Robust Network Structures for Acoustic Model on CHiME5 Challenge Dataset

Alim Misbullah[1]

[1]da Vinci Innovation Lab, ASUSTek Computer Inc., Taiwan

misbullah@gmail.com

Abstract

Our work is focus on robust acoustic model for the 5th CHiME5 challenge. To boost the performance to be better than baseline, we designed different network structure that can be suitable for CHiME5 challenge dataset. Our final result can achieve 13% relative improvement compare to the baseline.

1. Background

In recent years, the deep neural network has been widely used for various machine learning tasks including speech recognition [1] [2] [3] and image recognition [4] [5]. In speech recognition task, the deep neural network is used to train both acoustic and language modelling.
Recently, most papers reported that the acoustic model performed perfectly on clean condition but not for noisy condition like happened in daily environment. The 5th CHiME challenge is held to encourage people who are interested to provide best solution for distant multi-microphone issue in everyday home environment [6]. The challenge provide two tracks, those are single and multiple array based on the information from the 5th CHiME challenge official website.
To involved in the challenge, we conduct experiments focus on acoustic model robustness by designing different network structures. The contribution will be briefly described in Section 2. The experimental setup and result will be presented in Section 3. Finally, the conclusion appear in Section 4.

2. Contributions

In this work, we contribute to design different network structures for CHiME5 challenge dataset. We only involved to boost performance for acoustic model robustness on single array dataset. The CNN-TDNN-LSTM and CNIN-TDNN-BLSTM network are designed to compete baseline TDNN network performance.

3. Experiments

3.1. CHiME5 Challenge Dataset

In our experiment, we use the full CHiME5 challenge dataset from official CHiME5 challenge website. We do not use any additional speech data in our experiment. The detail dataset description can be found in website[1] and recent CHiME5 paper [6]. The corpus consist of training, development and testing which recorded in different session e.g. kitchen, dining and living. In addition, the data is fully transcribed and segmented to provide correct utterances information.
To train the acoustic model, we only use subset of CHiME5 challenge dataset which is 249425 utterances or equal to 137 hours with single array. However, the total utterances that be used to train the DNN model is smaller after removing useless utterances by performing cleanup script [2].

3.2. Experimental Setup

In this experiment, the MFCC feature with 40 dimension and adding pitch with 3 dimension are extracted from speech datasets . We use Kaldi [3] toolkit and follow the avalaible script for CHiME5 challenge[3] for training and evaluation our acoustic model structure.
Our experiment is only focus on extended network structure to find robust acoustic model network for CHiME5 dataset. The experiment is started by training HMM-GMM model to obtain alignments. The HMM-GMM triphone model is performed to filter bad utterances before continuing DNN model training.

3.2.1. Baseline TDNN networks

The baseline system is trained based on chain model from NNET3 Kaldi with LF-MMI objective function [2]. The network structure use 8 hidden layers with 512 nodes in each layer. The total parameters are about 7.3 millions [1].
The training use 10 epochs with initial and final learning rate are 0.001 and 0.0001 respectively. In the example script, we found that the system use i-vector with 100 dimensions in top of network structure which is concatenated with input features.

3.2.2. CNN-TDNN-LSTMP networks

To boost the baseline system performance, we firstly try to use CNN-TDNN-LSTM network. The example script from Kaldi toolkit is used to train CHiME5 Challenge corpus. The networks structure is shown in Figure 1. In Figure 1, we denote input and output using white blocks, the green blocks are used to denote CNN layers, the blue blocks are used to denote TDNN layer with ReLU activation function and gold blocks are used to denote LSTM Projected [7].
In green blocks, the CNN use 256 number filter for each layers. The first CNN layer take input from *idct* layer then the pooling is performed after the first of CNN layer by 2 subsampling. In blue blocks, the TDNN layers use 1024 hidden node for each layer with different left and right context input. The LSTM Projected layers in gold blocks use 1024 cell dimension, 256 nodes recurrent projected and 256 nodes for non recurrent projected. The total parameter of Figure 1 structure is approximately 48.8 millions.

[1] http://spandh.dcs.shef.ac.uk/chime_challenge/data.html

[2] https://github.com/kaldi-asr/kaldi/blob/master/egs/wsj/s5/steps/cleanup/clean_and_segment_data.sh

[3] https://github.com/kaldi-asr/kaldi/blob/master/egs/chime5/s5/run.sh

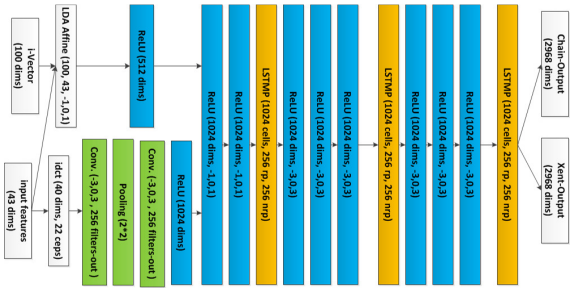

Figure 1: *CNN-TDNN-LSTMP networks structure*

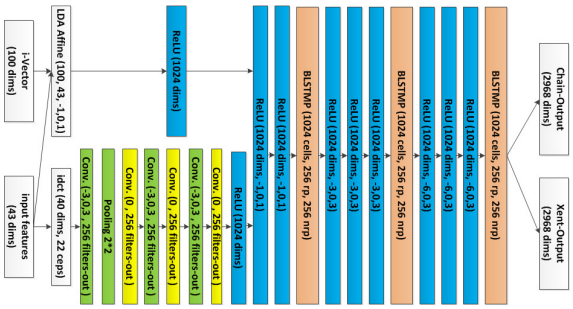

Figure 2: *CNIN-TDNN-BLSTMP networks structure*

3.2.3. CNIN-TDNN-BLSTMP networks

We extend the CNN-TDNN-LSTMP network from previous section by using Bidirectional LSTMP. In addition, we use zero time and height offsets between two CNN layer to reduce the network parameters. The zero time and height offsets for CNN layers is adopted from *Network in Network* (NIN) paper [4]. We named the network structure as CNIN-TDNN-BLSTMP and shown in Figure 2. The total network parameters are 69.6 millions approximately.

3.3. Experimental Result

In our experiment, we first evaluated the acoustic model using 3-gram language model. The language model is trained by using SRILM toolkit [4] from training corpus transcription. In example script [5], they use several approach to train language model and choose the best language which obtain lower perplexity.

In Table 1, we can achieve WER 71.78% using CNIN-TDNN-BLSTMP which has 11% relative improvement compare to the baseline system. We also evaluate the best model by combining posterior output of acoustic models before generate the lattices. The performance of combining posterior can achieve 13.85% relative improvement compare to the baseline.

In addition, we also perform RNNLM rescoring model [8] using available script from Kaldi [6] to train RNNLM model. The training transcription is used to train RNNLM model. The lattice output of decode acoustic model combine posterior is used as input to perform RNNLM rescoring. Finally, we can obtain slightly improvement compare by using RNNLM rescoring as shown in Table 1.

In Table 2, the result is shown for each session from different environments e.g. kitchen, dining, and living. The result show that S09 session can obtain better performance compare to S02 session. It is may caused by speaker distance and microphone quality. In addition, the evaluation result is obtained without perform RNNLM rescoring as shown in Table 2.

Table 1: *Overall WER (%) for the systems tested on the development test set.*

Track	System	WER
Single	TDNN Baseline	81.28
	CNN-TDNN-LSTMP	75.00
	CNIN-TDNN-BLSTMP	72.15
	CNIN-TDNN-BLSTMP + sMBR	71.78
Decode acoustic model combine posterior		70.02
+ RNNLM rescoring		69.80

Table 2: *Results for the best system. WER (%) per session and location together with the overall WER.*

Track	Session		Kitchen	Dining	Living	Overall
Single	Dev	S02	78.72	69.44	67.16	70.02
		S09	69.16	68.55	64.16	
	Eval	S01	68.41	53.74	72.62	61.31
		S21	67.02	52.52	58.47	

4. Conclusion

In this experiment, we tried to find appropriate network structure for CHiME5 challenge dataset. We did several experiments by adjusting network structures without adding more training corpus. By our limited time, we only can confirm that the best model for CHiME5 challenge dataset is CNIN-TDNN-BLSTMP based on our experiment. However, we think that the performance can be better if there has more training data from different sources for acoustic model training.

5. References

[1] V. Peddinti, D. Povey, and S. Khudanpur, "A Time Delay Neural Network Architecture for Efficient Modeling of Long Temporal Contexts," in *INTERSPEECH 2015*, 2015.

[2] V. Peddinti, D. Galvez, P. Ghahrmani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely Sequence-trained Neural Networks for ASR Based on Lattice-free MMI," in *INTERSPEECH 2016*, 2016.

[3] D. Povey, A. Ghoshal, and et. al, "The Kaldi Speech Recognition Toolkit," in *ASRU 2011*, 2011.

[4] M. Lin, Q. Chen, and S. Yan, "Network In Network," *arXiv:1312.4400v3 [cs.NE] 4 Mar 2014*, 2014.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv:1512.03385v1 [cs.CV] 10 Dec 2015*, 2015.

[4] http://www.speech.sri.com/projects/srilm/download.html

[5] https://github.com/kaldi-asr/kaldi/blob/master/egs/chime5/s5/local/train_lms_srilm.sh

[6] https://github.com/kaldi-asr/kaldi/blob/master/scripts/rnnlm/train_rnnlm.sh

[6] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth CHiME Speech Separation and Recognition Challenge: Dataset, task and baselines," in *INTERSPEECH 2018*, Hyderabad, India, Sep. 2018.

[7] H. Sak, A. W. Senior, and F. Beaufays, "Long Short-term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling," in *INTERSPEECH 2014*, 2014.

[8] H. Xu, T. Chen, D. Gao, Y. Wang, K. Li, N. Goel, Y. Carmiel, D. Povey, and S. Khudanpur, "A Pruned RNNLM Lattice-rescoring Algorithm for Automatic Speech Recognition," in *ICASSP 2018*, 2018.