Robust Network Structures for Acoustic Model on CHiME5 Challenge Dataset

Alim Misbullah
da Vinci Innovation Lab, ASUSTek Computer Inc., Taiwan
misbullah@gmail.com

Abstract
Our work is focused on robust acoustic model for the 5th CHiME Challenge. To boost the performance better than baseline, we designed different network structures for CHiME5 dataset. Our final result can achieve 13% relative improvement compare to the baseline TDNN network.

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 will conduct experiments focus on acoustic model robustness by designing different network structures. The detail experiment is described in Section 3.

Contributions

In this work, we contribute to design different network structures that fit 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.

CHiME5 Challenge Dataset

In our experiment, we use the full CHiME5 Challenge dataset that provided by team without adding extra data. The detail dataset description can be found in website[1] and recent CHiME5 paper [6]. The corpus consists 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 dataset which is 249425 utterances or equal to 137 hours with single array. However, the final utterances that we use to train the DNN model is smaller due to the cleanup process.

Experimental Setup

In this experiment, the speech datasets are extracted using 40 dimension MFCC features and pitch. To conduct the experiment, we use Kaldi [3] toolkit and follow the avalaible script for CHiME5 [2].
Our experiment is only focus on extended network structure to find robust acoustic model network for CHiME5 dataset. We first start experiment by training GMM-HMM model to obtain alignments. We then continue to train the baseline TDNN model based on GMM-HMM alignment.

Baseline TDNN networks

The baseline system is trained based on chain model from NNET3 Kaldi [6]. 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.

CNN-TDNN-LSTMP networks

To boost the performance from the baseline system, we first try to use CNN-TDNN-LSTM network. We use the example script from Kaldi toolkit to train CHiME5 corpus. The networks structure is shown in Figure 2. In Figure 2, 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 2 structure is approximately 48.8 millions.

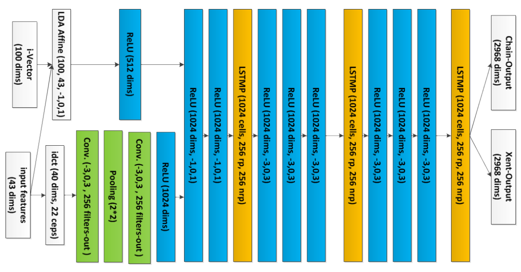

Figure 1: CNN-TDNN-LSTMP networks structure

[1]http://spandh.dcs.shef.ac.uk/chime_challenge/data.html
[2]https://github.com/kaldi-asr/kaldi/blob/master/egs/chime5/s5/run.sh
[3]http://www.speech.sri.com/projects/srilm/download.html
[4]https://github.com/kaldi-asr/kaldi/blob/master/egs/chime5/s5/local/train_lms_srilm.sh
[5]https://github.com/kaldi-asr/kaldi/blob/master/scripts/rnnlm/train_rnnlm.sh

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 **??**. The total network parameters are 69.6 millions approximately.

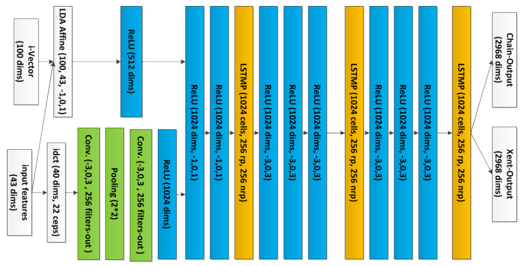

Figure 2: CNIN-TDNN-BLSTMP networks structure

Experimental Results

In our experiment, we first evaluated the acoustic model using 3-gram language model. The language model is trained by using SRILM toolkit [3] from training corpus transcription. In example script [4], 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 tried to train RNNLM rescoring model [8] using available script from Kaldi [5]. We use training transcription as corpus to train RNNLM model. We use lattice output from decode acoustic model combine posterior to perform RNNLM rescoring. Finally, we can obtain slightly improvement compare after rescoring using RNNLM 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.

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	

Conclusions

In this experiment, we tried to find appropriate network structure for CHiME5 challenge dataset. We did several experiments by adjusting network structures and 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.

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.