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 beyond baseline, we design different network structures for CHiME5 dataset. Our final result can achieve 17% relative improvement compared 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 having 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.

Contribution
In this work, we contribute to design different network structures that fit for CHiME5 Challenge dataset. We only involved in boost performance for acoustic model robustness on single array dataset. The CNNTDNN-LSTM and CNNTDNN-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 without training data and extra data. The detail dataset description can be found in website and in recent CHiME5 paper [6]. The corpus consists of training, development, and test which recorded in different session, e.g. kitchen, dining and living. In addition, the data is fully transcribed and segmented to provide correct utterance information.

To train the acoustic model, we only use subset of CHiME5 dataset which is 29,662 utterances or equal to 137 hours with single array. However, the final utterances that we use to train the TDNN 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 [7] toolkit and follow the available script for CHiME5. The LSTM-based model is only focused on extending different network structure to fit 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 uses 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 experiment, we try to find appropriate network structure for CHiME5 challenge dataset.

CNN-TDNN-LSTM 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 network 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 Layer [7].

In green blocks, the CNN uses 256 number filter for each layer, which is concatenated with input features.

CNN-TDNN-BLSTM networks
We extend the CNN-TDNN-LSTM network from previous section by using Bidirectional LSTM. In addition, we use zero time and height offsets for CNN layer to reduce the network parameters. The zero time and height offsets for CNN layers is adopted from https://github.com/kaldi-asr/kaldi/blob/master/egs/chime5/s5/local/train_lms_srilm.sh

In this experiment, we try 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 CNNTDNN-BLSTM which has 11% relative improvement compared to the baseline.

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 may be caused by speaker distance and microphone quality.

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

<table>
<thead>
<tr>
<th>Track</th>
<th>System</th>
<th>WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single</td>
<td>TDNN Baseline</td>
<td>81.28</td>
</tr>
<tr>
<td>Single</td>
<td>CNN-TDNN-LSTM</td>
<td>73.00</td>
</tr>
<tr>
<td>Single</td>
<td>CNNTDNN-BLSTM</td>
<td>72.15</td>
</tr>
<tr>
<td>Single</td>
<td>CNNTDNN-BLSTM + sMBR</td>
<td>71.78</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Track</th>
<th>Session</th>
<th>Kitchen</th>
<th>Dining</th>
<th>Living</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single</td>
<td>s02</td>
<td>78.72</td>
<td>69.44</td>
<td>65.78</td>
<td>68.52</td>
</tr>
<tr>
<td>Single</td>
<td>s09</td>
<td>59.09</td>
<td>69.16</td>
<td>68.53</td>
<td>64.16</td>
</tr>
<tr>
<td>Overall</td>
<td>70.02</td>
<td>63.61</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Conclusions

In this experiment, we try to find appropriate network structure for CHiME5 challenge dataset. We did several experiments by adjusting network structures and adding more training corpus. Due to our limited time, we only can confirm that the best model for CHiME5 challenge dataset is CNNTDNN-BLSTM based on our experiment. However, we think that the performance can be better if there is more training data from different sources for acoustic model training.

References