THE NDSC TRANSCRIPTION SYSTEM FOR THE 2018 CHIME5 CHALLENGE

Dan Qu[*], *Cheng-Ran Liu*, *Xu-Kui Yang*, *Wen-lin Zhang*

National Digital Switching System Engineering and Technological R&D Center, Zhengzhou, China
qudanqudan@sina.com, chengranlearning@qq.com, gzyangxk@gmail.com

ABSTRACT

The National Digital Switching System Engineering and Technological R&D Center (NDSC) speech-to-text transcription system for the 2018 CHiME-5 is described. The time delay neural network (TDNN) and TDNN-long short term memory recurrent neural network (TDNN-LSTM) systems are trained using deep bottleneck features (BNF). Since the audio recordings from parallel worn microphone are available, the third system is trained, in which the alignments of audio recordings from Kinect device are generated from worn microphone audio recordings. At last, the minimum Bayes risk (MBR) combination was utilized to combine different systems and reduce WER further. The WER of our system on develop dataset is 74.61%, leading to a 6% absolute reduction comparing with the base-line system.

Index Terms— speech recognition, CHiME challenge, bottleneck feature, parallel data, Kaldi

1. INTRODUCTION

This paper describes the development of NDSC speech-to-text transcription systems for the 2018 CHiME challenge[1] [1]. CHiME-5 challenge considers the problems about distant microphone conversational speech recognition in home environments. Speech data are composed of twenty home parties in real home environments. Each party conversation has two participants acting as hosts and two acting as guests. Each session which lasted at least two hours was recorded by 6 distant microphone arrays with 4 microphones each and 4 binaural microphones worn by each participant.

The challenge includes single-array and mutiple-array. For each track, it has two separate rankings. Ranking A is focused on conventional acoustic model and official languages model, which means that the lexicon and language model can not be changed compared to the baseline system. while B is include all other systems and the lexicon and language model can be changed. The system addressed in this paper is single-array and corresponds to ranking A.

Different architectures of deep learning based acoustic model are trained. The time delay neural network (TDNN) [3] and TDNN-long short term memory recurrent neural network (TDNN-LSTM) [4][5] are trained using bottleneck features (BNF) [6]. Because of the audio recordings from parallel worn microphone are available, another TDNN is trained, in which the alignments are generated from worn microphone audio recordings[7][8]. All systems were trained by utilizing the Kaldi speech recognition toolkit [2]. Lattice-based minimum Bayes risk (MBR) combination method was utilized to combine different systems and reduce WER further [9].

The outline of this paper is as follows. Section 2 describes the related techniques used in our systems. A detailed description of the NDSC transcription system is given in Section 3. Section 4 presents the performance of our systems on the evaluation dataset, and conclusions are drawn in Section 5.

2. RELATED WORK

In this section, we present a brief introduction to acoustic models, language models, system combination used in our transcription.

2.1. Acoustic Models

With the development of speech recognition, deep neural networks (DNN) [10][11] with outstanding modeling capabilities and superior feature representations are widely applied. Among all kinds of DNNs, TDNN is particularly suitable for speech recognition as speech materials are difficult to segment precisely [12]. TDNN is known as a precursor to convolutional neural networks (CNN) [13][14], which utilized in speech recognition can date back to 1987. Despite lack of affine transform in the initial layer, the TDNN can model long-term temporal dependencies from short-term input speech features for the temporal resolutions that TDNN operates at increases from layer to layer by performing temporal convolution. To speed up the TDNN training and reduce the model size the sub-sampling processing can be used under the assumption that

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

* corresponding author: Dan Qu (Email: qudanqudan@sina.com)

10.21437/CHiME.2018-18

neighboring activations are correlated [15]. Moreover, performance gain will be get by setting higher frame rate at the lower layers while the computational efficiency can be still preserved.

Though recurrent neural network (RNN) [16] have a powerful advantage for long contextual information representations, the traditional RNNs encounter the gradient exploding or vanishing problems when being trained by the stochastic gradient descent (SGD) algorithm [17]. While the LSTM RNN is proposed to alleviate this problem. Compared with the traditional RNN, in LSTM linear recurrent connections are used instead of non-linear ones in the conventional RNN, which lead to more smooth back propagation of gradients. Moreover, Combining TDNN layers and LSTM layers reduces WER further, comparing with the LSTM model.

2.2. Bottleneck Features

Because neural nets can compress the input features and classify the features, BNF are typically extracted by training a DNN with a middle bottleneck layer, which concludes fewer nodes comparing to layers below or above it. BNF can be considered as a low dimensional vector compressed by the information about phonetic class and phonetic context [18]. The recognition performance is confirmed to be improved significantly by utilizing BNF.

2.3. Parallel Alignment Training

For the speech data from distant microphones are known as degraded, far-field speech recognition [19] is a tough task. To improve the far-field speech recognition performance, the DNN based acoustic models can be trained using the time-synchronize parallel data, e.g., the simultaneously collected worn microphone and distant microphone speech data, which can reduce the mismatch between systems trained on clean speech from worn microphones and noisy speech from distant microphone[20]. Most of the methods can be grouped into two categories. One of the categories, using all the data from different environment to train the models is named multi-condition training. Another category is environment-aware training which uses the environment features as auxiliary information features. The environment features are extracted from worn microphone and distant microphone data.

2.4. System Combination

To obtain better recognition results, system combination which utilizes the complementarities of different systems has been used in various speech processing tasks. The main combination methods include one-best-based combinations such as recognizer out voting error reduction (ROVER) and lattice-based combinations such as MBR combination. MBR combination is a lattice-based system combination under the MBR decoding framework. In this method, a further improve can be get by merging the lattices from different systems into one topology and then decoding the merged lattice.

3. THE DEVELOPED SYSTEM

3.1. Data Used

The 5th CHiME challenge used audios from 20 parties in different home with a total duration of 50 hours, which have been split into training, development set and evaluation test sets. Each session is composed of the recordings made by the binaural microphone worn by participants and by 6 microphone arrays with 4 microphones each.

The training data are composed of left and right channel of binaural microphone data and a subset of all distant microphone data. We use 100 thousands and 1 million distant utterances training the systems, respectively. The system K1 and K2 will performance better when the subset number of utterances is 100 thousands according to the results, and the system K3 is 1 million.

The approach of muti-channel speech enhancement, similar to CHiME 4 recipe [21], is using a weighted delay-and-sum beamformer (BeamformIt) [22], which is the same with the baseline system.

3.2. ASR Systems

3.2.1. BNF-TDNN

To extract 80-dimensional BNF and 320-dimensional BNF, two TDNNs were trained respectively. 40-dimensional Mel-frequency cepstral coefficients (MFCCs), appended with a 100-dimensional i-Vectors, are used as input. The TDNN has 8 hidden layers, and the 8th hidden layer is set as bottleneck layer with 80 or 320 nodes. The model configuration of TDNNs is shown in Table 1.

Layer	Extracting BNF TDNNs		
	Context	Layer-type	dimension
1	[-2,-1,0,1,2]	LDA	
2	-	TDNN	512
3	[-1,0,1]	TDNN	512
4	-	TDNN	512
5	[-1,0,1]	TDNN	512
6	-	TDNN	512
7	[-3,0,3]	TDNN	512
8	[-3,0,3]	TDNN	512
9	[-6,-3,0]	TDNN	80/320

Table 1.the model configuration of TDNN used for extracting BNF

The TDNN for recognition which has 8 hidden layers with 512 units in each layer uses 80-dimensional BNF and 320-dimensional BNF respectively. The WER of system with different BNF is shown in Table 2. The system with 80-dimensional BNF has a lower WER comparing to the system with 320-dimensional BNF. The reason for this result probably is that the 80 dimensional BNF are more abstract and have more powerful ability of classification.

The training data are made up of all worn microphones data and a subset of all distant microphone data. We use 100 thousands and 1 million distant utterances training the systems, respectively. As the result shown in Table 2, the system with 80-dimensional BNF utilizing 100k distant utterances performs better. Because of the distant data with strong noise, the system using overmuch distant utterances may fit the noise in the data. That is probably why using less distant utterances performs better.

Track	BNF dim	utterances	NN Type	%WER
Single	Baseline	100k	TDNN	81.28
	80	100k	BNF-TDNN	78.91
	320	100k	BNF-TDNN	79.01
	80	1m	BNF-TDNN	80.59

Table 2.the %WER of systems on different training data

3.2.2. BNF-TDNN-LSTM

The method of extracting BNF is the same with 3.2.1. The number of the subset of distant data used for training is 100k. The input of neutral network is 80-dimensional BNF without cepstral truncation, spliced across $\pm n$ (n may be 1 or 2) frames of context, and appended with a 100-dimensional i-vectors. The model configuration of TDNN-LSTM is shown in Table 3. The context in this table is in terms of the splicing indices, e.g. '[-1, 0]' means the input to the current layer at a given time step t is a spliced version of previous layer outputs at times t-1 and t. LSTMP means projected LSTMs. This model has TDNN layers with an output dimension of 1280 and LSTM layers with cell dimensions of 1024. The WER of the system on dev dataset is 77.11%, which is shown in Table 5 and leads to a total reduction in WER of 1.8% absolute over the BNF-TDNN model.

Layer	TDNN-LSTM	
	Context	Layer-type
1	[-1,0,1]	LDA
2	[-1,0]	TDNN
3	[0,1]	TDNN
4	[-1,0]	TDNN
5	[0,1]	TDNN
6	[-3,0]	TDNN
7	[-3,0]	TDNN
8	[0,3]	LSTMP
9	[-3,0]	TDNN
10	[-3,0]	TDNN
11	[0,3]	LSTMP
12	[-3,0]	TDNN
13	[-3,0]	TDNN
14	[0]	LSTMP

Table 3.the model configuration of TDNN-LSTM

3.2.3. Parallel alignment training

Both TDNN and TDNN-LSTM are trained. The alignments of audio are generated from the worn microphone audio recordings. 40-dimensional MFCCs, appended with a 100-dimensional i-Vectors, are used as input. The number of the subset of distant data used for training is 1 million. The TDNN has 8 hidden layers with 512 relu units in each layer. The model configuration of TDNN is the same with the baseline system. The TDNN-LSTM model is the same with the model in 3.2.2. Table 4 compares the WER between TDNN and TDNN-LSTM system. Generally, comparing to TDNN, TDNN-LSTM system performs better with enough data. But in our experiment the TDNN system performs better than TDNN-LSTM. We speculate that it may be caused by insufficient data.

Track	utterances	NN Type	%WER
Single	1 million	TDNN	76.00
	1 million	TDNN-LSTM	78.61

Table 4. The %WER of TDNN and TDNN-LSTM system

3.2.4. MBR combination

The MBR combination is utilized in decoding. The system B1 is a combination of K1 and K2 with the weight of K1 is 0.4. The system B2 is combined by B1 and K3, and the weight of B1 is 0.4. The results of systems are shown in Table 5. Compared with the system K1 and K2, the combined system B1 has a 3% absolute reduction in WER. The system B2 which is combined with B1 and K3 has a 2.5% absolute reduction in WER.

Track	System	NN Type	%WER
Single	Baseline	TDNN	81.28
	K1	BNF-TDNN	78.91
	K2	BNF-TDNN-LSTM	77.11
	K3	Parallel alignment	76.00
	B1	K1 + K2	76.43
	B2	B1 + K3	74.61

Table 5. The %WER of systems on development dataset

4. PERFORMANCE ON EVALUATION DATASETS

Track	Session		%WER			
			Kitchen	Dining	Living	Overall
Single	Dev	S02	83.78	73.68	71.83	74.61
		S09	73.06	72.77	69.07	
	Eval	S01	77.75	60.18	76.77	67.31
		S21	72.89	59.50	63.34	

Table 6. per session and location %WER together with the overall %WER

The system performance on development and evaluation dataset in different condition is shown in Table 6. The system used for recognition is system B2. We observe that the performance is the best in the dining room and the worst in the kitchen, which is probably caused by the noise interference and the speakers movement. The WER of system B2 on evaluation dataset is 67.31%.

5. CONCLUSION

This paper describes the structure and development of NDSC speech-to-text transcription system for the CHiME-5

challenge. Different architectures of deep learning based acoustic model as well as parallel alignment training and MBR combination have been evaluated. The WER of the combined system is 74.61%, which leads a 6% absolute reduction comparing with the base-line system.

6. ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China (No. 61673395, No. 61403415, and No. 61302107).

7. REFERENCES

[1] J. Barker, S.Watanabe, E. Vincent, and J. Trmal, "The fifth CHiME Speech Separation and Recognition Challenge: Dataset, task and baselines," in Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018)

[2] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek,N. Goel, M.Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J.Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit", *In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Waikoloa, Hawaii, USA, 2011.

[3] V. Peddinti, G. Chen, D. Povey, S. Khudanpur, "Reverberation robust acoustic modeling using i-vectors with time delay neural networks,"*In Proceedings of Inter speech*, Dresden, Germany, 2015, pp. 2440 – 2444

[4] A. Graves, A. Mohamed, G.E. Hinton,"Speech recognition with deep recurrent neural networks," *In Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver,Canada, 2013, pp.6645–6649.

[5] A. Graves, N. Jaitly, A. Mohamed,"Hybrid speech recognition with deep bidirectional LSTM," *In Proceedings of Automatic Speech Recognition and Understanding (ASRU)*, Olomouc,Czech Republic, 2013, pp.273–278.

[6] .Veselý, Karel, et al. "The language-independent bottleneck features." *Spoken Language Technology Workshop* IEEE, 2013:336-341.

[7] Qian, Yanmin, T. Tan, and D. Yu. "An investigation into using parallel data for far-field speech recognition." *IEEE International Conference on Acoustics, Speech and Signal Processing* IEEE, 2016:5725-5729.

[8] Himawan, Ivan, et al. "Learning feature mapping using deep neural network bottleneck features for distant large vocabulary speech recognition." *IEEE International Conference on Acoustics, Speech and Signal Processing* IEEE, 2015:4540-4544.

[9] H. Xu, D. Povey, L. Mangu, et al., "Minimum bayes risk decoding and system combination based on a recursion for edit distance," *Computer Speech and Language*, 2011, 25(4): 802-828.

[10] G.E. Dahl, D. Yu, L. Deng, et al., "Context-Dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, 2012, 20(1):30-42.

[11] L. Deng , J.Y. Li, J.T. Huang, et al., "Recent advances in deep learning for speech research at Microsoft," *In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[12] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J Lang, "Phoneme recognition using time-delayneural networks," *IEEETransactions on Acoustics, Speech and Signal* Processing, vol. 37, no. 3, pp. 328–339, 1989.

[13] O. Abdel-Hamid, A. Mohamed, H. Jiang, G. Penn,"Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition, "*In Proceedings of International Conference on Acoustics,Speech and Signal Processing (ICASSP)*.Kyoto,Japan, 2012, pp.4277–4280.

[14] X. Glorot, Y. Bengio,"Understanding the difficulty of training deep feed-forward neural networks,"*J.Mach.Learn.Res.*, vol. 9, 2010, pp. 249–256.

[15] V. Peddinti, D. Povey, S Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts,"*In Proceedings of Interspeech*, 2015.

[16] Graves, Alex, A. R. Mohamed, and G. Hinton. "Speech recognition with deep recurrent neural networks." 38.2003(2013):6645-6649

[17] X. Glorot, Y. Bengio,"Understanding the difficulty of training deep feed-forward neural networks,"*J.Mach.Learn.Res.*, vol. 9, 2010, pp. 249–256.

[18] Gehring, Jonas, et al. "Extracting deep bottleneck features using stacked auto-encoders." *IEEE International Conference on Acoustics, Speech and Signal Processing* IEEE, 2013:3377-3381.

[19] Yoshioka, Takuya, S. Karita, and T. Nakatani. "Far-field speech recognition using CNN-DNN-HMM with convolution in time." *IEEE International Conference on Acoustics, Speech and Signal Processing* IEEE, 2015:4360-4364.

[20] R. Giri, M. L Seltzer, J. Droppo, and D. Yu, "Improving speech recognition in reverberation using a room aware deep neural network and multi-task learning," in Proceedings of ICASSP, 2015, pp. 5014–5018

[21] E. Vincent, S.Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," Computer Speech and Language, vol. 46, pp. 535–557, 2017

[22] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," IEEE Trans. on Audio,Speech, and Lang. Proc., vol. 15, no. 7, pp. 2011–2023, 2007.