The NWPU System for CHiME-5 Challenge

Zhiwei Zhao[1], *Jian Wu*[1], *Lei Xie*[1]

[1] School of Computer Science, Northwestern Polytechnical University

zwzhao@mail.nwpu.edu.cn, cswujian@mail.nwpu.edu.cn, lxie@nwpu.edu.cn

Abstract

The 5th CHiME Speech Separation and Recognition Challenge (CHiME-5) considers the problem of distant multi-microphone conversational speech recognition in everyday home environments. In this challenge, we take advantage of several beamforming techniques, powerful TDNN-F acoustic models, pruned lattice-rescoring algorithm as well as ROVER based system fusion method. Compared to original released system, our best result on development set achieves around 20% absolute improvement on WER.

1. Background

For CHiME5, we participate in the single-array task, which uses only reference array to recognise given utterance. Our proposed system focus on the following three parts: 1) a multi-beamformer based front end system which could produce several kinds of enhanced speech for back end fusion; 2) the acoustic modeling method with semi-orthogonal low-rank matrix factorization ; 3) language rescoring technique with LSTM-TDNN structure and the ROVER based system fusion method. With proposed system, we finally get 63.54% best WER on the official development set.

2. Contributions

The overall framework of our system is as given in Fig.1.

2.1. Front-end

Our front-end is mainly based on two beamforming methods: a group of fixed beamformers with sampled DoA (Direction of Arrival) and WNG (White Noise Gain) constraint, and the MVDR beamformer with estimated speaker independent/dependent masks. We use IRM (Ideal Ratio Mask) in the following experiments.

For fixed beamformer, we set WNG constraint as 0dB and sample DoA every 30 degrees. After that, as several candidate results could be produced, we only keep several best results and take them into joint decoding or system fusion stage.

For MVDR beamformer, we take neural networks as speaker dependent/independent (SI/SD) mask estimator, which is used for covariance matrices estimation. We use CNN-TDNN structures to model SI-mask estimator and 3-layer BLSTMs for SD-mask estimator. Both of them use filter-bank as input features. For SI-mask modeling, we choose utterances from close-talk training data with lower WER across sessions and compute SI-masks via complex Gaussian mixture model (CGMM) . We take quantified CGMM masks as targets and corresponding far-field utterances as network input. For SD-mask models, we prepare speaker and noise dataset first and then simulate mixture data based on them. The speaker dataset comes from non-overlap segments in development/test set, which could be automatically extracted according to the annonations and the noise dataset comes from no-speaker segments on the training set. We use a simple VAD to filter out silence in the noise dataset. Finally, we train SI-mask estimator on simulated mixture data for each speaker independently.

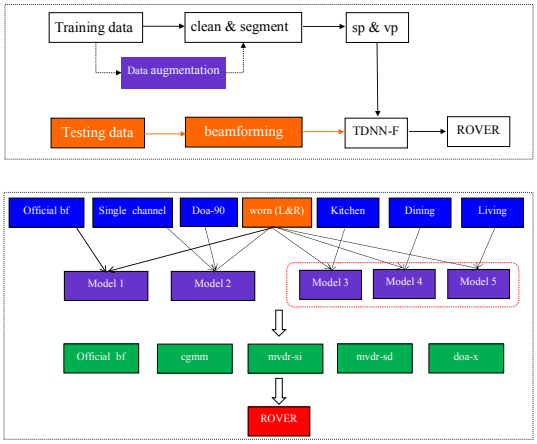

Figure 1: *System overview: upper) the overall pipeline; bottom) the details of each module.*

2.2. Acoustic model

For the acoustic model, we use TDNN-F that contains a convolutional layer, 19-TDNN layers and two fully connected layers. The model size is 175 MB. As to the acoustic feature, we use 40-dimensional log mel-filterbank (LMFB) features instead of the official MFCC features. The LMFB is fed into the convolutional layer and then the output of the convolutional layer is concatenated with a 100-dimension online ivector before it fed into the followed TDNN layer. We have trained various acoustic models using different types of training sets, denoted as model $\{1\sim 5\}$, as shown in Fig. 1. All the acoustic models are optimized using the lattice-free MMI [1].

2.3. Language model

In order to further improve the recognition performance, we use Kaldi-RNNLM tookit[2] to rescore the lattices from each dependent system. According to our experiments, an LSTM-TDNN structure with 3-order pruned lattice-rescoring algorithm yield the best results that achieves about 2% absolute WER reduction.

3. Experimental evaluation

3.1. Different architecture of acoustic models

Firstly, we have evaluated the performance of different acoustic models trained with the official baseline training set

(worn/82.70h + single channel/54.64h). The experimental results are shown in Tab. 1. By using the 1CNN + TDNN-F (19 layers) we can achieve about 10% absolute improvement compared to the official TDNN.

Table 1: *Comparison of various acoustic models trained with the baseline training data.*

Acoustic model	Feature	Dev (WER%)
Official GMM-HMM	MFCC	91.83
Official TDNN	MFCC	80.11
Official TDNN	LMFB	79.91
TDNN-F (11 layers)	LMFB	75.34
1CNN+TDNN-F (15 layers)	LMFB	73.16
1CNN + TDNN-F (19 layers)	LMFB	72.61

3.2. Different training set

We investigate the impact of training data on our model structure 1CNN + TDNN-F (19 layers) . As shown in Tab. 2 , training data of the baseline system is composed of about 64.39h all cleaned worn data and 36.26h single channel data. In the sub-scenario model training experiment, we choose about 36h kitchen/dining/living data within all single channel data respectively and then combine with all worn data. Finally , in the data augmentation experiment, we choose baseline data to combine with 15h cleaned DoA-90 data and about 147h cleaned official beamformit data generated by 878h all single channel data respectively. Tab. 3 and Tab. 6 shows the results on official develop set.

Table 2: *The details of the training data we used.*

ID	Data set	Original (hr)	Cleaned (hr)
1	Worn (L&R)	82.70	64.39
2	Single Channel All	878.41	-
3	Single Channel	54.64	39.26
4	DoA-90	25.89	15.68
5	Beamformit	219.60	147.78
6	Kitchen	55.04	35.78
7	Dining	55.04	37.25
8	Living	54.97	36.72

Table 3: *Comparison of acoustic models trained with different training data.*

ID	Training Data	Dev(WER%)
Baseline	1+3	72.61
Model 1	1+3+5	70.49
Model 2	1+3+4	71.32

3.3. Different front-end methods

Based on description of section 2.1, we apply different beamformer on develop data and results are shown in Tab. 4. The acoustic model is kept the same with the configuration of the baseline, including the input feature and the model structure. Compared with the official beamformit method, our best single beamforming approach yields 2.3% WER reduction.

Table 4: *Result of different beamformer based on baseline AM*

Methods	WER %
Beamformit	82.40
CGMM-MVDR	81.08
SD-MVDR	80.82
SI-MVDR	80.02
DoA-X joint decoding	80.69
SD-MVDR + DoA-X joint decoding	80.05

3.4. Final system

Finally, we use ROVER technique to vote all the recognized texts produced by multiple acoustic models on different beamformed speech and achieve the lowest WER of 63.54% on the development set.

Table 5: *Performance of the final system. (LM: N-gram)*

Beamformer	Model1	Model 2	Model{3∼5}
CGMM-MVDR	70.05%	70.80%	71.22%
DoA-105	69.67%	70.40%	70.72%
DoA-90	69.87%	70.62%	71.15%
DoA-60	69.66%	70.63%	70.94%
Beamformit	70.21%	71.07%	71.55%
SI-MVDR	68.83%	69.53%	70.07%
SD-MVDR	68.66 %	69.06%	69.57%

Table 6: *Performance of the final system. (LM: RNNLM)*

Beamformer	Model1	Model 2	Model{3∼5}
CGMM-MVDR	68.25%	68.68%	69.08%
DoA-105	68.13%	68.43%	68.65%
DoA-90	68.13%	68.65%	69.25%
DoA-60	68.09%	68.64%	69.09%
Beamformit	68.50%	69.28%	69.48%
SI-MVDR	66.98%	67.39%	68.07%
SD-MVDR	66.91%	67.00%	67.63%
ROVER	63.54%		

4. References

[1] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi." in *Interspeech*, 2016, pp. 2751–2755.

[2] 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," 2018.

[3] 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)*, Hyderabad, India, Sep. 2018.

[4] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*. IEEE, 1997, pp. 347–354.

[5] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohamadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018)*, Hyderabad, India, Sep. 2018.