The NWPU System for CHiME-5 Challenge

Zhiwei Zhao, Jian Wu, Lei Xie

Audio, Speech and Language Processing Group, School of Computer Science,
Northwestern Polytechnical University, Xi'an, China

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) [1] 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 [2] acoustic models, pruned lattice-rescoring algorithm as well as ROVER [3] based system fusion. Compared to the official baseline, our best system achieves around 18% and 17% absolute WER reduction on the development and test sets respectively.

1. Background

We participate in the CHiME5 single-array task, which uses only reference array to recognize given utterances. Our proposed system focuses on the following aspects:

- A multi-beamformer based front-end system which can produce several kinds of enhanced speech as candidate for back-end fusion;
- Acoustic modeling with semi-orthogonal low-rank matrix factorization;
- Language rescoring technique with LSTM-TDNN structure and the ROVER based system fusion.

With the proposed system, we finally get 63.54% and 56.10% WER on the official development and test sets respectively with RNNLM rescoring.

2. Contributions

The overall framework of our system, which focuses on front-end processing, data augmentation and acoustic modeling, is given in Fig.1.

2.1. Front-end

Our front-end is mainly based on two beamforming methods: a group of fixed beamformers [4] 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) and MVDR (Minimum Variance Distortionless Response) beamforming in all our experiments.

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

For MVDR beamformer, we train neural networks to predict speaker dependent/independent (SI/SD) masks, which are used for covariance matrices estimation. We use CNN-TDNN structure to model SI-mask estimator and 3-layer BLSTMs

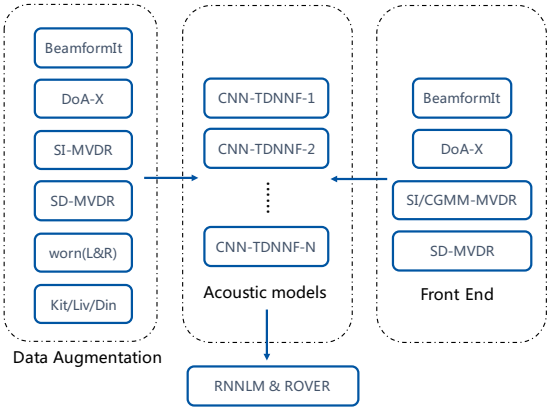

Figure 1: *System overview*

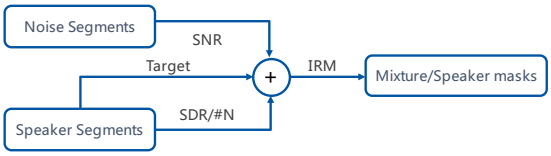

Figure 2: *Data simulation for SD mask estimator training*

for SD-mask estimator. Both of them use log mel-filterbank (LMFB) as input features. For SI-mask training, we choose utterances from close-talk dataset with lower WER across sessions and estimate SI-masks via complex Gaussian mixture model (CGMM) methods [5]. These CGMM masks are used as targets during the training stage and corresponding far-field utterance are used as the network input. For SD-mask models, we prepare speaker and noise dataset first and then simulate mixture data based on them (See Fig. 2). The speaker dataset comes from non-overlap segments in the development/test set, which could be automatically extracted according to the groundtruth. The noise dataset comes from non-speaker segments on the training set. We use a simple VAD to filter out silence segments. Finally, we train SI-mask estimator on simulated mixture data for each speaker independently with the same configurations.

2.2. Acoustic model

We use CNN-TDNN-F (19 TDNN-F layers following 1 CNN layer) structures for acoustic modeling with 40-dimensional log mel-filterbank (LMFB) features instead of the official MFCC. LMFB features are fed into the CNNs directly and the output of CNN is concatenated with 100-dimension online ivector before feeding into the following TDNN layers. We have trained various acoustic models using different types of training data, denoted as model $\{1\sim 5\}$, as shown in Fig. 1. All the acoustic

models are optimized using the lattice-free MMI criterion [6].

2.3. Language model

In order to further improve the recognition performance, we use Kaldi-RNNLM tookit [7] to rescore the lattices from each dependent system. According to our experiments, an LSTM-TDNN structure with 3-order pruned lattice-rescoring algorithm yields the best result with 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 training data (worn/82.70h + far field/54.64h). The experimental results on the development set are shown in Tab. 1. By using the CNN-TDNN structures 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 data

We have also investigated the impact of training data based on CNN-TDNN-F structures. Tab. 2 shows the amount of training data we used for acoustic modeling. The official training data set is composed of 64h worn data and 39h far field data and we denote it as the baseline training set. For Model 1, we add cleaned official beamforming data into the training set, and get 2% absolute WER reduction. For Model 2, we add 20k selected utterances which were processed with DoA-90 fixed beamformer and get slightly better results. And we also train 3 models for each scenario (Living/Kitchen/Dining), which were denoted as model 3$\sim$5. For the training data of each model, apart from far-field data recorded in the corresponding room, worn data is also included. Tab. 3 shows the results on official development data 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	Far Field(All)	878.41	-
3	Far Field(Baseline)	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 data augmentation.*

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 the description in Section 2.1, we apply different beamformers on the development data and results are shown in Tab. 4. Joint decoding here means posterior level average technique, and we can give different weights to each beamformed results. The acoustic model used has same the structures with the baseline system. Compared with the official beamformit method, our best single beamforming approach (SD-MVDR) yields 2.3% absolute WER reduction. Tab. 5 shows the absolute WER improvement for each speaker on the development set.

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

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

Table 5: *Absolute WER reduction for each speaker using SD-MVDR*

Speaker ID	Impr(WER%)
P05	2.91
P06	2.52
P07	3.00
P07	3.62
P25	1.86
P26	2.17
P27	0.39
P28	2.39

3.4. System ensemble

Finally, we use ROVER technique to vote all the recognized texts produced by multiple acoustic models on different beamformed speech. For Ranking A, we only use official N-gram LM during decode stage, and for Ranking B, we use RNNLM rescoring as in Section 2.3. At last, we achieve the lowest WER of 65.17% and 63.54% on the development set for Ranking A and B respectively. Results are summarized in Tab. 6 and Tab. 7.

3.5. Final results

Based on the description in Section 3.4, our final results on the official development and test sets are reported in Tab. 8.

Table 8: *Final results achieved by our best system*

Track	Ranking	Dev (WER %)	Test (WER %)
Single	A	65.17	57.90
	B	63.54	56.10

Table 6: *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%
ROVER	65.17%		

Table 7: *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] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth'chime'speech separation and recognition challenge: Dataset, task and baselines," *arXiv preprint arXiv:1803.10609*, 2018.

[2] 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.

[3] 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.

[4] E. Mabande, A. Schad, and W. Kellermann, "Design of robust superdirective beamformers as a convex optimization problem," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 77–80.

[5] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust mvdr beamforming using time-frequency masks for online/offline asr in noise," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5210–5214.

[6] 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.

[7] 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.

[8] 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.