Multiple Beamformers with ROVER for the CHiME-5 Challenge

Sining Sun[1], *Yangyang Shi*[2], *Ching-Feng Yeh*[2], *Suliang Bu*[3],
Mei-Yuh Hwang[2], *Lei Xie*[1]

[1]School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2]Mobvoi AI Lab, Seattle, USA
[3]Dept. of Electrical Engineering and Computer Science, University of Missouri-Columbia, USA
{snsun,lxie}@nwpu-aslp.org, {yyshi,cfyeh,mhwang}@mobvoi.com, sbkc6@mail.missouri.edu

Abstract

In this paper, we describe our systems and report our results for the CHiME-5 single-array track. We focus on front-end multi-channel speech processing, including beamforming and dereverberation. To address the complexity of the data and recording scenario, we use multiple beamformers where each beamformer targets at a predefined direction. N-Best lists are obtained from decoding each beamformed signal. These multiple N-best lists are further processed by ROVER to get the final result. Before beamforming, a multi-channel generalized weighted prediction error method is adopted to do the dereverberation. Comparing with the official baseline system, CNN-TDNN-F shows significant improvement. In language modeling, LSTM-based language model re-scoring generates additional improvement. Without system fusion, our single system can get 14.4% relative word error rate reduction on development set over the baseline system.

1. Background

For CHiME-5, we participate in the single-channel track. Our system focuses exclusively on multi-channel signal processing techniques, including dereverberation and beamforming. Therefore not much effort is spent on system fusion with various advanced acoustic models. Data augmentation and LSTM based language model (LM) re-scoring are used to improve the system performance.

2. System Description

2.1. Multi-channel dereverberation

The performance of speech recognition systems and the effectiveness of beamforming methods are degraded in reverberant environments. Many beamforming techniques rely on accurate estimation on the direction of arrival (DOA). However, reverberation makes it difficult to estimate DOA accurately. In our system, the generalized weighted prediction error (GWPE) [1] algorithm is used to do dereverberation before beamforming. As a multiple-input multiple-output method, GWPE is able to preserve the DOA information. Hence it can be applied to multi-channel signals before beamforming.

2.2. Multiple beamformers with ROVER

In our system, we propose to use multiple beamformers and each of them focuses on one specific direction, where fixed beamforming (FB) with a constraint for white noise gain (WNG) [2] is applied. We design 5 FBs focusing on 60, 90, 120, 150 and 180 degree respectively. Apart from 5 FBs, complex Gaussian mixture model based MVDR (CGMM-MVDR) beamforming, which has been proved to be effective in previous work [3, 4], is also applied to provide the 6th speech stream.

As mentioned earlier, it is difficult to estimate accurate DOA, and hence it is hard to choose the best beamformed signal from the 5 FBs and the CGMM-MVDR signal. Instead of choosing one best from the 6 signals, we decode the 6 beamformed signals in parallel, each generating an N-Best list. Then, SRILM [5] ROVER [6] toolkit is applied to get the final result out of the 6 N-best lists. Figure 1 shows the flowchart of our multiple beamforming system.

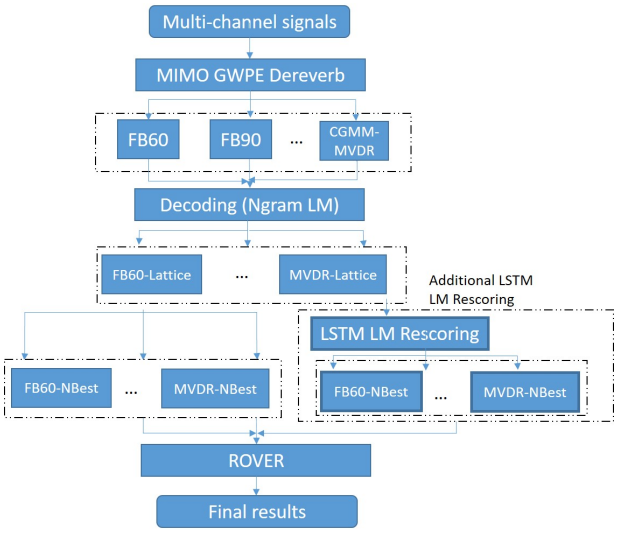

Figure 1: *Flowchart of ROVER over multiple beamformed outputs.*

2.3. Data augmentation

Data augmentation is a very directive and effective approach to improve the robustness of the acoustic model. On one hand, data augmentation can cover more data types, on the other hand, it can also alleviate the mismatch between training and test data. In our system, in order to further eliminate the distortion introduced by our front-end processing techniques, we also do a simple data augmentation during training. First, we randomly choose 22000 (22k) utterances from the training set and apply dereverberation using GWPE. Then, we generate 44k new enhanced training data by enhancing the randomly selected 22K utterances using FB90 and CGMM-MVDR beamforming techniques respectively. Finally, we augment the official training data (worn data + 100k far-field data) with the 44k enhanced

data.

2.4. Acoustic model

A factored form of time delay neural network (TDNN-F) introduced by [7] is used as our acoustic model (AM). Considering the noisy data, we add an extra CNN layer before TDNN-F. Compared with the official TDNN model, this AM structure can get better performance. For TDNN-F, we follow the configuration in the Kaldi [8] recipe.

2.5. LSTM language model

The language model is built based on a high-rank LSTM language model [9] with several optimization and regularization methods [10]. The high-rank LSTM language model uses a mixture of softmaxes (MoS) to make the softmax layer in LSTM language model more expressive. Similar to conventional LSTM language model [11, 12, 10], a sequence of hidden states is obtained after processing the input sequence over a stack of recurrent layers. On top of the hidden states, the MoS represents the conditional distribution of current word as weighted sum of different softmax layers.

3. Experimental evaluation

We only participate in the single array track. In this section, our single array results are reported.

3.1. Different acoustic models

In our work, we compared two different acoustic model structures, the official TDNN acoustic model and our CNN-TDNN-F model. Table 1 compares two different AM architectures, both using the official beamforming method (Beamformit) and the official training data, where the TDNN result is the official baseline result given by the challenge official website. CNN-TDNN-F gives significant improvement in this task. Therefore we will report results based on this CNN-TDNN-F acoustic model for the rest of the paper.

Table 1: *WER (%) on the dev set, using the official training data and Beamformit*

Track	Acoustic model	WER
Single	TDNN	81.30
	CNN-TDNN-F	75.91

3.2. Deverberation and multiple beamformers

Next we apply our multiple beamformers to the test speech, before sending it to be decoded by the CNN-TDNN-F model. Table 2 shows that without deverberation the proposed multiple beamformers followed by ROVER can reduce WER to 72.54%. When GWPE dereverberation is applied before beamforming, we get another 1% WER reduction. In all ROVER cases in Table 2, the official 3-gram LM is used to generate 10-best per speech stream. That is, 60 hypotheses per utterance are combined by ROVER in this table.

3.3. Data augmentation

Next, we add 44k utterances of enhanced speech into the official training data, as described in section 2.3. Although we add only 44k enhanced utterances, we are able to obtain almost 1% WER reduction, as shown in Table 3.

Table 2: *WER (%) on the dev set using different front-ends on the CNN-TDNN-F model*

Dereverb	Beamformimg	WER
No	Beamformit	75.91
No	Multi beamformers	72.54
Yes	Multi beamformers	71.56

Table 3: *WER (%) of data augmentation on the dev set*

Dereverb	Beamforming	Augmentation	WER
Yes	Multi beamformers	No	71.56
Yes	Multi beamformers	Yes	70.68

3.4. ROVER with LSTM LM N-Best lists

Finally we use our MoS LSTM LM to rescore the lattices generated by the official 3gram LM, and then dump another 10-best for each speech stream. Hence we obtain a new set of 60 hypotheses for each test utterance, from the LSTM nbest lists. Together with the original 60 hypotheses, the 120 hypotheses are then combined by ROVER to output the final hypothesis. Table 4 shows the WER drops to 69.57% when LSTM 10-best lists are added into ROVER.

Table 4: *WER on the dev set via ROVER before vs. after adding the LSTM nbest lists*

N-Best	WER
from 3-gram LM only	70.68
+LSTM LM	69.57

3.5. Results for the best system

Table 5 shows the breakdown WERs of our best system. Living room environment seems to be a bit easier than kitchen and dinning room. On development set, from 81.30% to 69.57%, we get 14.4% relative improvement. On evaluation set, we reduce the WER to 68.71%.

Table 5: *Results from the best system. WER (%) per session and location together with the overall WER.*

Track	Session		Kitchen	Dinning	Living	Overall
Single	Dev	S02	79.72	69.83	64.58	69.57
		S09	68.97	68.43	64.14	
	Eval	S01	80.98	61.75	78.50	68.71
		S21	73.73	58.88	65.15	

4. Conclusions

In this paper, we described our systems for CHiME-5 challenge.

- On the front-end signal processing, combining GWPE dereverberaton and multiple beamformers with N-Best ROVER gave significant improvement.
- On acoustic models, CNN-TDNN-F significantly improved over the standard TDNN backend.

- On language models, LSTM language model rescoring was used to further reduce the WER.

Our best system achieved 69.57% and 68.71% WER on development and evaluation sets.

5. References

[1] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening," *IEEE Transactions on Audio Speech and Language Processing*, vol. 20, no. 10, pp. 2707–2720, 2012.

[2] E. T. Mabande, A. Schad, and W. Kellerman, "Design of robust superdirective beamformers as a convex optimization problem," *IEEE International Conference on Acoustics, Speech, and Signal Processing*.

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

[4] S. Bu, Y. Zhao, M.-Y. Hwang, and S. Sun, "A probability weighted beamformer for noise robust ASR," *to appear in Interspeech*, 2018.

[5] A. Stolcke, "Srilm-an extensible language modeling toolkit," in *International Conference on Spoken Language Processing, Denver, Colorado, Usa, Sept*, 2002, pp. 901–904.

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

[7] 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."

[8] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[9] Z. Yang, Z. Dai, R. Salakhutdinov, and W. W. Cohen, "Breaking the Softmax Bottleneck: A High-Rank RNN Language Model," *CoRR*, vol. arXiv:1711.03953, 2017. [Online]. Available: http://arxiv.org/abs/1711.03953

[10] S. Merity, N. S. Keskar, and R. Socher, "Regularizing and Optimizing LSTM Language Models," *CoRR*, vol. arXiv:1708.02182, 2017. [Online]. Available: http://arxiv.org/abs/1708.02182

[11] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent Neural Network Regularization," *CoRR*, vol. arXiv:1409.2329, no. 2013, 2014. [Online]. Available: http://arxiv.org/abs/1409.2329

[12] X. Chen, A. Ragni, X. Liu, and M. J. F. Gales, "Investigating Bidirectional Recurrent Neural Network Language Models for Speech Recognition," in *The Proceedings of Interspeech*, 2017.

[13] 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.