The USTC-iFlytek System for CHiME-4 Challenge

Jun Du[1], Yan-Hui Tu[1], Lei Sun[1], Feng Ma[2],
Hai-Kun Wang[2], Jia Pan[2], Cong Liu[2], Jing-Dong Chen[3], Chin-Hui Lee[4]

[1]University of Science and Technology of China, Hefei, Anhui, P. R. China
jundu@ustc.edu.cn, {tuyanhui,sunlei17}@mail.ustc.edu.cn
[2]iFlytek Research, iFlytek Co., Ltd., Hefei, Anhui, P. R. China
{fengma,hkwang,jiapan,congliu2}@iflytek.com
[3]Northwestern Polytechnical University, Xian, Shaanxi, P. R. China
jingdongchen@ieee.org
[4]Georgia Institute of Technology, Atlanta, Georgia, USA
chl@ece.gatech.edu

Abstract

The submitted system for CHiME-4 this year includes significant improvements over the previous one for CHiME-3, including the front-end design, training data augmentation via different versions of the official training data, acoustic model fusion, and language model fusion. The final average WERs of the real test set are 2.24%, 3.91%, 9.15% for 6-channel, 2-channel, and 1-channel, respectively.

1. Background

For CHiME-4 [1], we participate all the tracks including 1 ch, 2 ch, and 6 ch tasks. In comparison to CHiME-3 challenge [2, 3], our new progress mainly includes: 1) a closed-loop optimization for beamforming by leveraging the information of deep neural network (DNN) based single-channel speech enhancement and the recognition results; 2) diversified training data using the noisy data of each channel, the multiple beamformers' outputs data of 6 channels and 2 channels; 3) the acoustic model upgrade via the deep convolutional neural networks (DCNNs) [4, 5]; 4) the long short-term memory (LSTM) based language modeling [6, 7]. In the next section, we will elaborate these contributions.

2. Contributions

The overall system flowchart is given in Fig. 1, where a unified framework for all three tasks, namely 1/2/6-channel cases, is designed. In the training stage, both the acoustic models with multiple front-ends and language models are built. In the recognition, multiple acoustic models are fused at the state-level first and then first-pass decoding is performed with the HMM and 3-gram to generate the lattice as the hypotheses , which are served for the second-pass decoding with a LSTM-based LM. The details can refer to the following subsections.

2.1. Beamforming

The beamforming approach showed in Fig. 2 is similar to the work in [8], namely the generalized sidelobe canceller (GSC) with a post-filtering. First, the time-frequency (T-F) masking is calculated via the complex Gaussian mixture model (CGMM) [9] to estimate the covariance matrices of noise and noisy speech. The relative transfer function is implemented by the eigenvector-based estimation in [10]. To further improve the estimation of the time-frequency masking, both the VAD information from the segmentation results of recognizer based on the beamformed speech and the ideal ratio mask (IRM) estimated using a DNN are used for a second-pass beamforming. The input of IRM-DNN is the log-power spectra (LPS) of the beamformed speech while the output is the masking values of T-F units calculated between the noisy speech of the channel 5 and the underlying clean speech. Obviously, the VAD and IRM information are based upon the beamforming results, which forms a feedback loop optimization [11] among them with multiple iterations. Experiments show that this new framework could significantly improve the recognition accuracy, yielding a remarkable gain over the best beamforming approach of CHiME-3 [2].

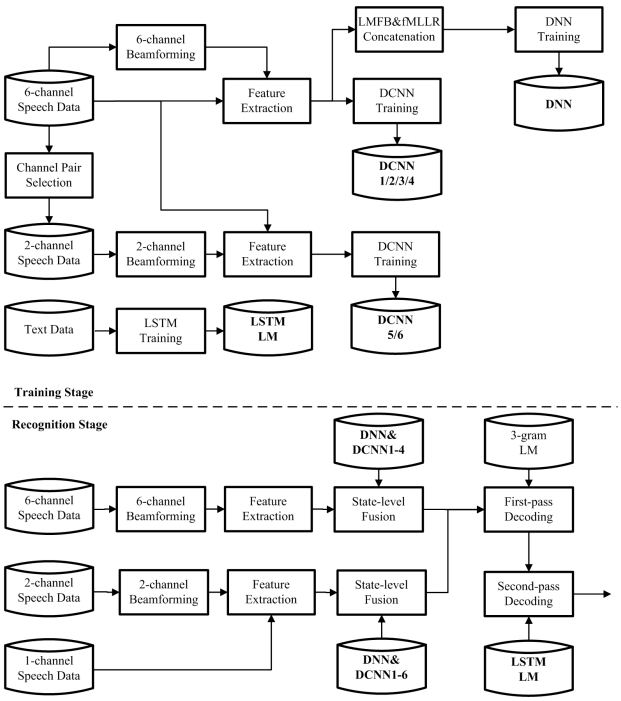

Figure 1: System overview.

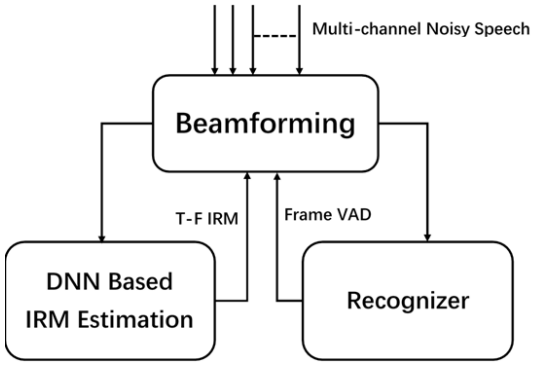

Figure 2: Beamforming.

2.2. Training data augmentation

The training data augmentation is a straightforward way to enlarge the data coverage, especially for 3 tasks with different settings of channel number in CHiME-4. Three data types are employed. First, the noisy speech of 5 channels (excluding the channel 2 with the most degraded speech) are used to simulate the 1-channel testing case. Then, the enhanced version using the beamforming approach applied to all 6 channels matches the 6-channel testing cases. Finally, we randomly select some channel pairs from 5 channels and the beamformed results of the corresponding channel pairs can correspond to the 2-channel testing cases. As illustrated in Fig. 1, both the noisy speech and 6-channel beamformed data are adopted to train the models (DNN and DCNN1/2/3/4) for all testing. Meanwhile, the noisy speech plus 2-channel beamformed data are combined to learn two other models (DCNN5/6) for 2-channel and 1-channel testing.

2.3. Acoustic models

We train mainly 2 types of neural networks. One is the conventional DNN and the other is DCNN. For 6-channel system, 5 models are built and fused via the state-level posterior average [3], including one DNN and 4 DCNNs (DCNN1/2/3/4). The DNN system concatenates the log mel-filterbank (LMFB) and fMLLR features. 4 DCNNs consist of LMFB-based one, fMLLR-based one and two others with different parameter settings. For 2-channel and 1-channel systems, two additional DCNNs (DCNN5/6) are used, namely 7 models in total. The DCNN system shows the strong complementarity when fused with the DNN system.

2.4. Language models

Besides the 5-gram and RNNLM provided officially, we also train an LSTM-based LM to further improve the recognition accuracy. According to our experiments, the LSTM-based LM alone could yield a relative WER reduction of more than 30% over the 5-gram+RNNLM based system.

3. Experimental evaluation

3.1. Beamforming

The Word Error Rates (WERs) on the evaluation data of the official and our proposed beamformers for the 2 ch and 6 ch track have been showed in Table 1. We adopted the DNN based official baseline system, 11 frames of 40-dimension fMLLR features. The DNN architecture is 440-2048*7-1987, namely 40*11 dimension for fMLLR input features, 7 hidden layers with 2048 nodes for each, and 1968 nodes for the output layers as our ASR model. The IRM-DNN is trained using 7 frames of 257-dimension LPS features of CH5. The IRM-DNN architecture is 1799-2048*3-257, namely 257*7 dimension for LPS input features, 3 hidden layers with 2048 nodes for each, and 257 nodes for the output T-F IRM. The significant reduction of WERs on the evaluation data for both the development and test sets can be found in Table 1, and our beamformer is more effective for more adverse environments and more microphones than official beamformer.

Table 1: WERs obtained with the proposed and official beamformers on the evaluation data for 2 ch and 6 ch tracks using the official DNN acoustic model.

Track	System	Dev		Test	
		real	simu	real	simu
2ch	Official	8.50	9.92	17.07	15.98
	Proposed	6.20	8.12	10.86	11.69
6ch	Official	6.25	7.15	11.82	11.43
	Proposed	4.18	4.17	6.13	5.23

Table 2: WERs between the baseline and data augmentation based systems on the evaluation data for 2 ch and 6 ch tracks.

Track	System	Dev		Test	
		real	simu	real	simu
2ch	Official	6.20	8.12	10.86	11.69
	Retrained	4.68	6.26	7.14	9.39
6ch	Official	4.18	4.17	6.13	5.23
	Retrained	3.24	3.33	4.33	4.21

3.2. Training data augmentation

The Word Error Rates (WERs) on the evaluation data of the official baseline and retrained by data augmentation DNN systems for the 1 ch, 2 ch and 6 ch tracks have been showed in Table 2. As for the retrained DNN system, 42-dimensional LMFB features and 40-dimensional fMLLR features with their first-order and second-order derivatives are used. The 20-dimensional i-vector features [3] are concatenated. The DNN architecture is 2234-2048*7-1965, namely (42+40)*3*9+20 dimension for LMFB+fMLLR+ivector combined input features, 7 hidden layers with 2048 nodes for each, and 1965 nodes for the output layer. The training data contains 1,3,4,5,6 channels data and 4 kinds of beamformered data, totally 78642 utterances(8738*9), and the beamformered data by our proposed method is used as our test set. Approximately 20% WERs reduction can be found between the official and our proposed systems in the all test sets.

3.3. Acoustic models

The Word Error Rates (WERs) on the real evaluation data of the different acoustic models for the 1 ch, 2 ch and 6 ch tracks have been showed in Table 3. The main difference of our DCNNs and conventional CNNs is the number and the size of the filters. The multi-layer small convolution kernels (3x3 and 3x5) are used, and the total number of convolutional layers is 25. And the learning rate is set to 0.002, and the batch size is 2048. Batch normalization is also used to speed up the training. In the Table 3, we can find that the performance of DCNNs is sig-

Table 3: WERs with the different acoustic models on the real evaluation data for 1 ch, 2 ch and 6 ch tracks.

Track	Set	System					
		DNN	DCNN1	DCNN2	DCNN3	DCNN4	Ensemble
1ch	Dev	8.29	7.70	7.71	9.87	9.86	6.10
	Test	14.58	15.47	14.72	17.05	17.45	11.12
2ch	Dev	4.68	4.05	4.13	5.24	5.43	3.55
	Test	7.14	6.87	6.94	8.34	8.36	5.40
6ch	Dev	3.24	2.88	2.99	3.37	3.50	2.61
	Test	4.33	3.87	4.09	4.67	4.90	3.22

Table 4: Average WER (%) for the tested systems.

Track	System	Dev		Test	
		real	simu	real	simu
1ch	Official LM	6.10	8.24	11.15	13.62
	LSTM LM	4.55	6.61	9.15	11.81
2ch	Official LM	3.56	4.89	5.41	7.30
	LSTM LM	2.33	3.46	3.91	5.74
6ch	Official LM	2.55	2.61	3.24	3.06
	LSTM LM	1.69	1.78	2.24	2.12

Table 5: WER (%) per environment for the best system.

Track	Envir.	Dev		Test	
		real	simu	real	simu
1ch	BUS	5.84	4.90	14.10	7.58
	CAF	5.09	9.84	9.64	14.98
	PED	2.66	4.84	6.89	11.58
	STR	4.63	6.86	5.98	13.09
2ch	BUS	2.74	2.83	5.16	3.83
	CAF	2.18	4.29	3.83	5.66
	PED	1.73	2.94	3.18	6.14
	STR	2.65	3.79	3.49	7.32
6ch	BUS	2.05	1.64	2.65	1.36
	CAF	1.50	1.99	2.09	1.87
	PED	1.50	1.55	1.74	2.35
	STR	1.71	1.93	2.48	2.91

nificantly better than DNN, and it can bring approximately 20% WERs reduction comparing to DNN on the real test set. Finally, the model ensemble is used by the state posterior average of single system output, it also can bring about 20% WERs reduction.

3.4. Language models

The Word Error Rates (WERs) on the evaluation data of the official and our language models for the 1 ch, 2 ch and 6 ch tracks have been showed in Table 4. The forward and backward LSTM models are trained for the combination of language models. We can find that the performance of LSTM-LM is more effective when the front-end and acoustic models are better in Table 4. Finally, Table 5 presents the results per environment for our best system, and we can find the improvement is significantly comparing to baseline system.

4. Acknowledgments

The team would like to thank other colleagues, Zi-Rui Wang, Tian Gao, Xiao Bao, Ye-Bo Bao, Di-Yuan Liu, and Shi-Liang Zhang for their contributions of building the final systems.

5. References

[1] 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*, 2016.

[2] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime' speech separation and recognition challenge: Dataset, task and baselines," in *IEEE ASRU*, 2015.

[3] J. Du, Q. Wang, Y.-H. Tu, X. Bao, L.-R. Dai, and C.-H. Lee, "An information fusion approach to recognizing microphone array speech in the chime-3 challenge based on a deep learning framework," in *IEEE ASRU*, 2015, pp. 430–435.

[4] D. Yu, W. Xiong, J. Droppo, A. Stolcke, G. Ye, J. Li, and G. Zweig, "Deep convolutional neural networks with layer-wise context expansion and attention," in *INTERSPEECH*, 2016.

[5] T. Sercu, C. Puhrsch, B. Kingsbury, and Y. LeCun, "Very deep multilingual convolutional neural networks for lvcsr," in *ICASSP*, 2016.

[6] M. Sundermeyer, R. Schluter, and H. Ney, "Lstm neural networks for language modeling," in *INTERSPEECH*, 2012.

[7] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu, "Exploring the limits of language modeling," in *arXiv:1602.02410v2*, 2016.

[8] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function gsc and postfiltering," *IEEE Transactions on Speech Audio Processing*, vol. 12, no. 6, pp. 561–571, 2004.

[9] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust mvdr beamforming using time-frequency masks for online/offline asr in noise," in *ICASSP*, 2016.

[10] A. Krueger, E. Warsitz, and R. Haeb-Umbach, "Speech enhancement with a gsc-like structure employing eigenvector-based transfer function ratios estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 206–219, 2011.

[11] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE Transactions on Speech Audio Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.