The USTC-NELSLIP Systems for CHiME-6 Challenge

Jun Du[1], *Yan-Hui Tu*[1], *Lei Sun*[1], *Li Chai*[1], *Xin Tang*[1], *Mao-Kui He*[1], *Feng Ma*[1], *Jia Pan*[1], *Jian-Qing Gao*[1], *Dan Liu*[1], *Chin-Hui Lee*[2], *Jing-Dong Chen*[3]

[1]University of Science and Technology of China, Hefei, Anhui, P. R. China
[2]Georgia Institute of Technology, Atlanta, Georgia, USA
[3]Northwestern Polytechnical University, Shanxi, P. R. China

{jundu, tuyanhui}@ustc.edu.cn

Abstract

This technical report describes our submission to the 6th CHiME Challenge. The submitted systems for CHiME-6 cover both the multiple-array speech recognition track and multiple-array diarization and recognition track. For each track, the results corresponded to Category A and Category B are reported. The main technique points of our submission include the deep learning based iterative speech separation, training data augmentation via different versions of the official training data, SNR-based array selection, front-end model fusion, acoustic model fusion.

Index Terms: speech recognition, human-computer interaction, computational paralinguistics

1. System Overview

For CHiME-6, we participate both the multiple-array speech recognition track and multiple-array diarization and recognition track. The overall system flowchart is given in Figure 1. In CHiME-4 and CHiME-5, we proposed the iterative mask estimation and two-stage speech separation methods, and both methods combined the neural network based mask estimation and the CGMM-based method in [1, 2]. In CHiME-6, the iterative beamforming speech separation model (IBF-SS model) is proposed for the multiple-array speech recognition track and multiple-array diarization and recognition track. In CHiME-4, an information fusion framework with multi-channel feature concatenation was proposed in [3]. In CHiME-6, the technique of the multi-feature concatenation is also utilized, but the features containing multi-channel spatial characteristics are firstly applied in our acoustic model. For Rank B, the 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 simple RNN-based language model (LM). In the following sections, we will give the detailed description on the multiple-array speech recognition track and multiple-array diarization and recognition track, respectively.

2. Multiple-Array Speech Recognition Track

First of all, due to rules defined by official, systems are allowed to exploit knowledge of the utterance start and end time, the utterance speaker label and the speaker location label. It's allowed to use binaural data and far-field data in the training set.

2.1. Speech separation model training

In order the make our speech separation model can utilize the spatial information of multi-channel data, the four beamformed outputs of CGMM is utilized for our iterative beamforming speech separation model (IBF-SS model). And the 257-dimensional log-power spectrograms (LPS) feature for four speakers and the corresponding phase features are concatenation as the input feature. For model architecture in both stages, we utilize a two layer Bi-directional long short-term memory (BLSTM) as the speech separation model, each direction with 512 cells. The Pytorch is used for training. After separation stage, the resulting waveforms can be directly sent to back-end acoustic models, or provide only masks to the beamforming.

2.2. Data simulation

For the acoustic model training data, a certain amount of far-field data are simulated by using the worn data and real far-filed data. The impulse responses are estimated through pairs of worn data and far-field data. With estimated impulse responses and worn data, simulated far-field data can be generated.

2.3. Acoustic model (AM)

All the acoustic models are trained using lattice-free maximum mutual information (LF-MMI) training with kaldi tool. For the single-input acoustic model, the input feature is the 40-dimensional MFCC. And for the multi-feature acoustic models, the various features containing multi-channel spatial characteristics are concatenated into 840-dimensional features as the input features.

2.3.1. AM with single feature

* Resnet-TDNNF-BLSTM

2.3.2. AM with multi-feature concatenation

* Resnet-CNN-TDNNF
* Hybrid CNN network
* CLDNN (CNN-BLSTM-DNN)
* Att.-CNN-TDNNF

3. Multiple-Array diarization and Recognition Track

First of all, due to rules defined by official, systems are not allowed to exploit knowledge of the utterance start and end time, the utterance speaker label and the speaker location label. And the acoustic models utilized in multiple-array speech recognition track are directly used in here. Because the utterance start and end time can not be used for the CGMM-based model as the initial information, so the IBF-SS is designed to directly learn the masks estimated by CGMM. So we extend

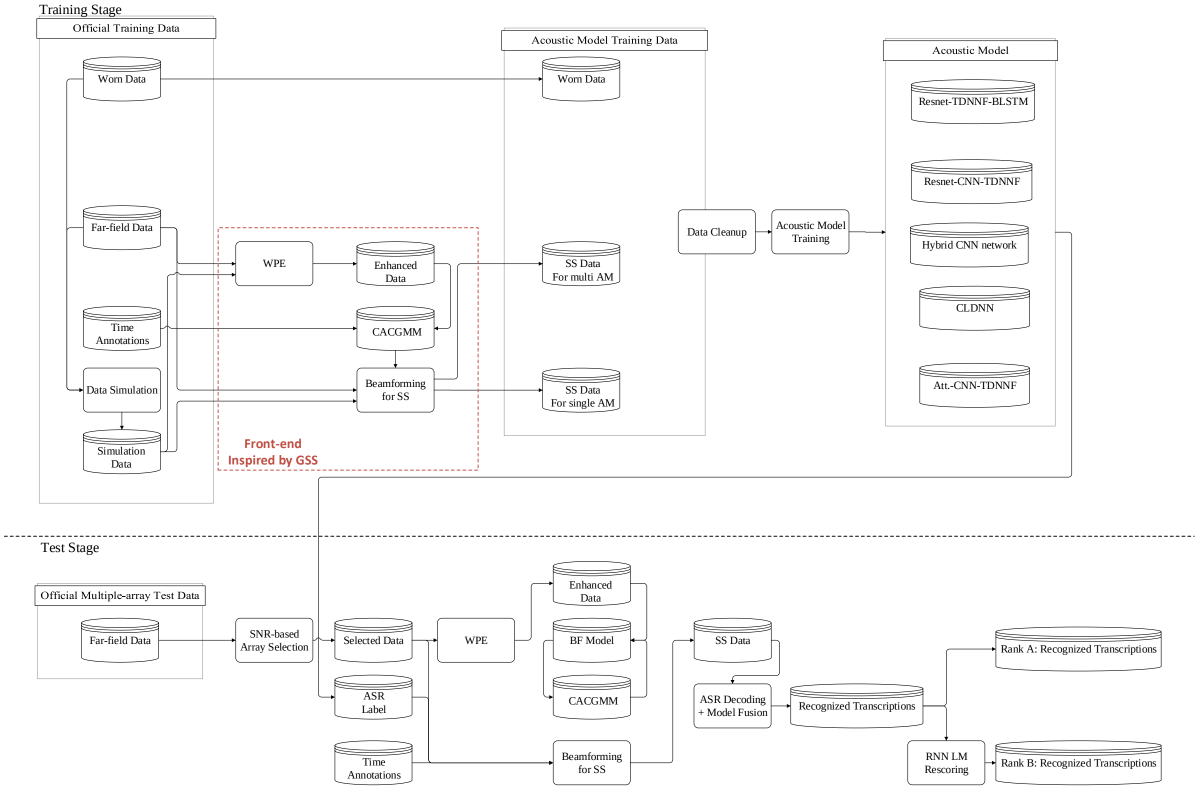

Figure 1: *An illustration of unified training stage, including front-end processing, data augmentation and acoustic modeling.*

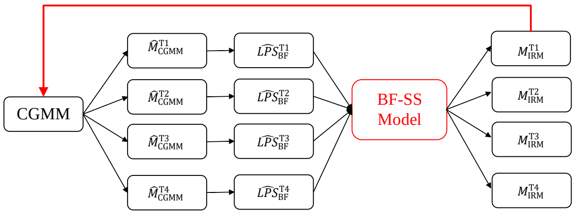
CGMM
$\widehat{M}^{T1}_{CGMM}$
$\widehat{M}^{T2}_{CGMM}$
$\widehat{M}^{T3}_{CGMM}$
$\widehat{M}^{T4}_{CGMM}$
$\widehat{LPS}^{T1}_{BF}$
$\widehat{LPS}^{T2}_{BF}$
$\widehat{LPS}^{T3}_{BF}$
$\widehat{LPS}^{T4}_{BF}$
BF-SS Model
M^{T1}_{IRM}
M^{T2}_{IRM}
M^{T3}_{IRM}
M^{T4}_{IRM}

Figure 2: *An illustration of the proposed iterative beamforming speech separation model (IBF-SS model) for multiple-array speech recognition track.*

our single-channel teacher-student learning framework [4] into multi-channel.And the CGMM can be seen as the teacher model to guide the IBF-SS model to learn the spatial information from the multi-channel training data showed as in Figure 3.

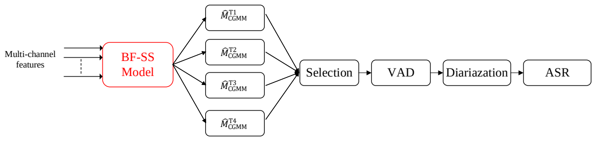
Multi-channel features
BF-SS Model
$\widehat{M}^{T1}_{CGMM}$
$\widehat{M}^{T2}_{CGMM}$
$\widehat{M}^{T3}_{CGMM}$
$\widehat{M}^{T4}_{CGMM}$
Selection
VAD
Diariazation
ASR

Figure 3: *An illustration of the whole framework for multiple-array diarization and recognition track.*

Table 1: *Results of the best system tested on the development and eval test set for multiple-array speech recognition. WER (%) per session and location together with the overall WER.*

Category	Session		Dining	Kitchen	Living	Overall
A	Dev	S02	34.95	35.13	29.77	31.11
		S09	29.60	28.10	27.65	
	Eval	S01	25.55	42.75	38.13	30.96
		S21	25.49	34.97	26.05	
B	Dev	S02	34.66	34.86	29.50	30.77
		S09	29.10	27.74	27.22	
	Eval	S01	25.01	42.66	37.44	30.50
		S21	25.14	34.84	25.34	

4. Results

The best system is taken to be the one that performs best on the development set. For multiple-array speech recognition, in addition to the overall WER, we report separate WERs for every location (kitchen, dining, living) in every development session (see Table 1). For multiple-array diarization and recognition, we report diarization error rates (DERs), Jaccard error rates (JERs), and WERs (%) for development and eval sets (see Table 2).

Table 2: *Results of the best system tested on the development test set for multiple-array diarization and recognition. DERs, JERs and WER (%).*

Category	Development set			Evaluation set		
	DERs	JERs	WER	DERs	JERs	WER
A	56.69	58.49	68.22	65.37	64.15	68.48
B	56.69	58.49	68.15	65.37	64.15	68.42

5. References

[1] Y. Tu, J. Du, L. Sun, F. Ma, H. Wang, J. Chen, and C. Lee, "An iterative mask estimation approach to deep learning based multi-channel speech recognition," *Speech Commun.*, vol. 106, pp. 31–43, 2019. [Online]. Available: https://doi.org/10.1016/j.specom.2018.11.005

[2] L. Sun, J. Du, T. Gao, Y. Fang, F. Ma, and C. Lee, "A speaker-dependent approach to separation of far-field multi-talker microphone array speech for front-end processing in the chime-5 challenge," *J. Sel. Topics Signal Processing*, vol. 13, no. 4, pp. 827–840, 2019. [Online]. Available: https://doi.org/10.1109/JSTSP.2019.2920764

[3] Y. Tu, J. Du, Q. Wang, X. Bao, L. Dai, and C. Lee, "An information fusion framework with multi-channel feature concatenation and multi-perspective system combination for the deep-learning-based robust recognition of microphone array speech," *Comput. Speech Lang.*, vol. 46, pp. 517–534, 2017. [Online]. Available: https://doi.org/10.1016/j.csl.2016.12.004

[4] Y. Tu, J. Du, and C. Lee, "Speech enhancement based on teacher-student deep learning using improved speech presence probability for noise-robust speech recognition," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 12, pp. 2080–2091, 2019. [Online]. Available: https://doi.org/10.1109/TASLP.2019.2940662