The IOA Systems for CHiME-6 Challenge

Hangting Chen[1,2], *Pengyuan Zhang*[*1,2], *Qian Shi*[1,2], *Zuozhen Liu*[1,2]

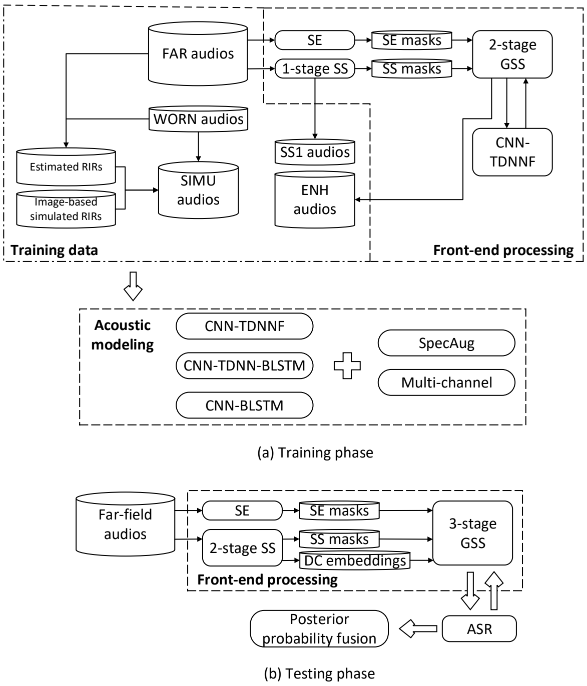
[1]Key Laboratory of Speech Acoustics & Content Understanding, Institute of Acoustics, CAS, China
[2]University of Chinese Academy of Sciences, Beijing, China

{chenhangting, zhangpengyuan, shiqian, liuzuozhen}@hccl.ioa.ac.cn

Abstract

The paper presents IOA's submission to the 6th CHiME Challenge. Our systems include the front-end enhancement combining deep learning-based and probabilistic model-based source separation, training data augmentation, acoustic modeling with multi-channel branches and system fusion. Tested on the evaluation sets, our best system for Track 1 Category A/B has yielded $35.11\%/34.53\%$ word error rate (WER) respectively, with an absolute reduction of $16.18\%/16.76\%$ compared with the baseline model.

1. System overview



(a) Training phase

(b) Testing phase

Figure 1: *The (a) training and (b) testing phase of our systems.*

This report describes our contribution to the 6th CHiME challenge (CHiME-6), which provides speech data recorded in the real party scenario via microphone arrays and presents extreme speech overlap and unrestrained speaking styles [1][2]. Our systems are designed for Track 1 Category A/B. Figure 1 shows the framework of the training and testing procedures of our systems. It consists 5 parts, including deep learning-based single-channel speech separation (SS), multi-channel speech enhancement with guided source separation (GSS), training data augmentation, acoustic modeling and system fusion.

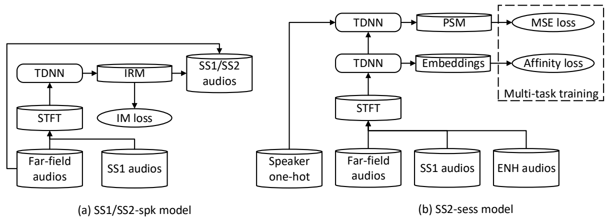
In the training phase, we first train 1-stage SS models for each speaker in each session (SS1), a universal speech enhancement model (SE). The separated audios serve as a part of training database. Then a 2-stage GSS is initialized with the speaker and noise masks, further refined by ASR alignments. 3 types of acoustic models with multi-channel branches are trained with the dataset augmented with additional data.

In the testing phase, we train 2-stage speech separation models (SS2). A 3-stage GSS is deployed to perform multi-channel speech separation. The final results are obtained with posterior probability fusion.

The detailed descriptions of the systems and the word error rate (WER) results on the development (Dev.) and evaluation (Eval.) sets can be found in the following sections.

2. Front-end processing

2.1. Deep learning-based single channel source separation



(a) SS1/SS2-spk model

(b) SS2-sess model

Figure 2: *The (a) SS1/2-spk and (b) SS2-sess model for single channel speaker separation.*

The deep learning-based single-channel source separation is to generate source masks and embeddings, which are provided to GSS module. The SS1-spk and SS2-spk models are trained for each speaker in each session. The SE models are training with progressive learning, similar with [3].

The SS2-sess models serve as unified models to separate speakers as well as to extract source embeddings for each session. The shallow layers receive the short-time Fourier transform (STFT) and output the unit-norm speaker embeddings. The high layers additionally utilize the speaker condition to output masks. The model is trained to optimize the multi-task loss of affinity matrix [4] and phase-sensitive masks [5] (Figure 2),

$$L_m = \frac{1}{T \times F} \sum_{t,f} (\hat{m}_{t,f,s} - m_{t,f,s})^2, \qquad (1)$$

$$L_e = \frac{1}{(T \times F)^2} \sum_{t,f,t',f'} (\mathbf{e}_{t,f}^T \mathbf{e}_{t',f'} - \mathbf{b}_{t,f}^T \mathbf{b}_{t',f'})^2, \qquad (2)$$

[*] Pengyuan Zhang is the corresponding author.

$$L = L_m + \alpha L_e, \tag{3}$$

where s is the target speaker index, $\hat{m}_{t,f,s}$, $m_{t,f,s}$, $\mathbf{e}_{t,f} \in \mathbb{R}^{D\times 1}$ and $\mathbf{b}_{t,f} \in \{0,1\}^{S\times 1}$ represent the estimated PSM, the oracle PSM, the estimated speaker embedding and the speaker membership indicator for each time-frequency (T-F) bin, T and F are the total number of frames and frequency bins in the sample, D and S are the embedding dimension and the number of speakers, α is the loss balance factor.

In our experiments, the SS1-spk models utilize non-overlapping utterances. The SS2-spk and SS2-sess models additionally use audios separated by SS1-spk and enhanced by 1-stage GSS. We set $D = 20$ and $\alpha = 1.0$.

2.2. Multi-channel guided source separation

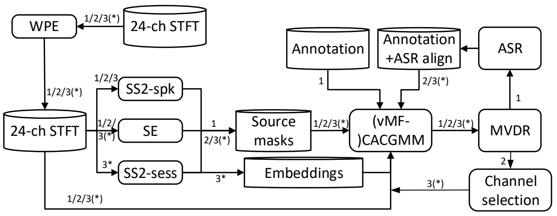


Figure 3: *The frame work of* 3*-stage GSS. The number represents the data flow in the* 1,2,3$(^*)$ *stages individually.*

We have developed the multi-channel separation based on the GSS [6]. The overall framework of systems is given in Figure 3. All the 24 channels audios are dereverbed with the weighted prediction error (WPE) [7]. SS and SE masks combined with annotations and alignments are served as an initialization of the complex angular central Gaussian model (CACGMM),

$$a(t,s) = (1-\beta)a_{annot}(t,s) + \beta a_{align}(t,s), \tag{4}$$

$$p_{init}(t,f,s) = \hat{m}(t,f,s)a(t,s), \tag{5}$$

where $\beta = 0.6$ is the confidence factor of the alignments, a_{annot} and a_{align} are frame-level source presence probability from the annotation and ASR alignment, $p_{init}(t,f,s)$ is the initialization mask. The interpolation is aimed to alleviate the inaccuracy of the alignment caused by the ASR transcription. After iterations, the masks representing the target speaker and the inference are used to do beamforming.

We briefly describe the 3 stages:

- The $1st$ stage uses $\hat{m}(t,f,s)$ to initialize CACGMM, generates ASR alignment by decoding the enhanced audios.
- The $2nd$ stage uses $p_{init}(t,f,s)$ to initialize CACGMM, generates signal-to-noise (SNR) information of each channel.
- The $3rd$ stage selects channels, uses $p_{init}(t,f,s)$ to initialize CACGMM, output enhanced audios.
- The 3^*rd stage differs by using von Mises-Fisher (vMF)-CACGMM model [8] with embeddings from SS2-sess and selecting channels with the fusion of SNR- and coherency-based [9] methods.

Each stage's performance of the front-end processing is presented in Table 1.

Table 1: *The Front-end results on the Dev. with CNN-TDNNF trained with WORN and ENH data [10].*

Stage	Baseline	1	2	3	3^*
WER(%)	45.42	43.02	42.62	42.14	**41.75**

The 2-stage GSS, which is adopted in the training phase to generate enough data, consists of the $1st$ and $3rd$ stages. The channel selection is in random to output 7-fold data, named ENH in Figure 1.

3. Acoustic models

3.1. Training data and settings

The whole training set contains worn headset data (WORN), far-field microphone array data (FAR), simulated data (SIMU), multi-channel enhanced data (ENH), totally 4 parts. The FAR data is made up of the original far-filed audios and single-channel audios enhanced by SS1 models. The SIMU data is generated by convolving the WORN data with image-based simulated room impulse responses (RIRs) and estimated RIRs calculated by the far and worn audio pairs. Moreover, it is observed that the short utterance combination can benefit the performance of the acoustic models. We have created 2 training sets, a small one with only WORN and ENH data, a large one with all mentioned data. The details of the training set are listed in Table 2.

Table 2: *The details of the training set, where* 1*-fold data is around* 40 *hours and* 120 *hours after* 3*-fold speed and volume perturbation.*

Data	Description	Fold
WORN	Headset microphone audios	$2/3$
FAR	Original far-field and SS1 separated audios	6
SIMU	Simulated far-field audios	6
ENH	Augmented enhanced audios	7
Small set	WORN+ENH	9
Large set	WORN+FAR+SIMU+ENH	22

3.2. Networks

Totally 15 acoustic models are trained for the final fusion. They are derived from CNN-TDNNF trained on the small set, CNN-TDNN-BLSTM trained on the large set and CNN-BLSTM trained on the large set (Table 3). CNN-BLSTM differs from CNN-TDNN-BLSTM by utilizing deeper CNN layers without interleaving TDNN and BLSTM layers. A multi-channel branch is introduced with CNN architectures, whose input is log power spectral (LPS) and magnitude squared coherence (MSC) [11]. The branch is trained in 2 ways, partial update and full update [12]. It was noticed that the multi-channel networks usually outperformed the single channel ones, audios from 3^*-stage GSS usually outperformed those from 3-stage.

The results of the acoustic model ensemble are plotted in Table 4. The fusion adopts the weighted average of log posterior probability inside each type of acoustic models. The final fusion was conducted across different types of acoustic models, and 3-stage and 3^*-stage front-end.

Table 3: *The architectures and performance of our acoustic models.*

Architecture	Dataset	Training settings	3-stage GSS Dev./Eval. WER(%)	3*-stage GSS Dev./Eval. WER(%)
CNN-TDNNF	Small	SpecAug	38.73/40.83	38.45/40.85
+Multichannel-CNN	Small	Partial update	38.10/39.16	37.95/38.98
+Multichannel-CNN-BLSTM	Small	Partial update	38.93/39.73	38.70/39.77
CNN-TDNNF-BLSTM	Small	SpecAug	38.41/40.04	37.90/39.95
+Multichannel-CNN	Small	Partial update	37.98/38.42	**37.76**/**38.27**
CNN-TDNNF-attention	Small	SpecAug	39.72/42.09	39.33/41.83
CNN-TDNN-BLSTM	Large	-	38.15/40.10	37.95/39.81
CNN-TDNN-BRLSTM	Large	-	38.19/40.26	37.89/40.16
+Multichannel-CNN	Large	Full update	38.27/40.42	38.02/40.16
+Multichannel-CNN	Small	Partial update	37.57/**38.60**	**37.29**/38.72
CNN-TDNN-BRLSTM-2	Large	SpecAug	42.15/44.03	41.92/43.70
CNN-BLSTM	Large	SpecAug	36.60/38.63	35.92/38.30
+Multichannel-CNN	Small	Partial update	37.47/38.45	37.17/38.30
CNN-BLSTM-deltaLayer	Large	SpecAug	37.69/39.41	37.30/39.27
CNN-BLSTM-resnet	Large	SpecAug	35.86/37.97	**35.54**/**37.95**

Table 4: *The ensemble results of different types of acoustic models and front-end processing.*

Acoustic model type (#)	3-stage GSS Dev./Eval. WER(%)	3^*-stage GSS Dev./Eval. WER(%)
CNN-TDNNF (3)	36.71/38.79	36.25/38.46
CNN-TDNNF + Multi-channel (3)	36.23/37.13	**36.07**/**37.05**
CNN-TDNN-BLSTM (3)	36.63/38.86	**36.38**/38.52
CNN-TDNN-BLSTM + Multi-channel (2)	37.02/38.32	36.67/**38.28**
CNN-BLSTM (3) CNN-BLSTM + Multi-channel (1)	34.88/36.37	**34.48**/**36.36**
Fusion with weight 0.05 : 0.15 : 0.1 : 0.1 : 0.6	34.18/35.67	**33.76**/**35.56**
Fusion with weight 0.4 : 0.6	**33.55**/**35.11**	
RNN rescore	**32.92**/**34.53**	

4. Conclusion

We present the performance details in Table 5, which are tuned on the Dev. set and tested on the Eval. set. For Category B, a language model based on the recurrent neural network (RNN) is trained for rescore. It yields around 0.6% improvement for both Dev. and Eval. sets.

Table 5: *The WERs (%) of our best systems for Category A and B.*

Category	Session	Dining	Kitchen	Living	Ave
A	S02	38.30	38.50	31.59	33.55
	S09	32.25	30.07	29.23	
	S01	29.58	48.49	42.72	35.11
	S21	29.76	39.66	28.60	
B	S02	37.51	38.02	31.06	32.92
	S09	31.54	29.69	28.11	
	S01	28.83	48.61	41.64	34.53
	S21	29.14	39.39	28.03	

5. References

[1] 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.

[2] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj, D. Snyder, A. S. Subramanian, J. Trmal, B. B. Yair, C. Boeddeker, Z. Ni, Y. Fujita, S. Horiguchi, N. Kanda, T. Yoshioka, and N. Ryant, "Chime-6 challenge:tackling multispeaker speech recognition for unsegmented recordings," 2020.

[3] L. Sun, J. Du, T. Gao, Y. Fang, F. Ma, and C.-H. Lee, "A speaker-dependent approach to separation of far-field multi-talker microphone array speech for front-end processing in the chime-5 challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 827–840, 2019.

[4] J. R. Hershey, C. Zhuo, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[5] H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *ICASSP 2015 - 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[6] N. Kanda, C. Böddeker, J. Heitkaemper, Y. Fujita, S. Horiguchi, K. Nagamatsu, and R. Haeb-Umbach, "Guided source separation meets a strong ASR backend: Hitachi/paderborn university joint investigation for dinner party ASR," *CoRR*, vol. abs/1905.12230, 2019. [Online]. Available: http://arxiv.org/abs/1905.12230

[7] L. Drude, J. Heymann, C. Boeddeker, and R. Haeb-Umbach, "NARA-WPE: A python package for weighted prediction error dereverberation in Numpy and Tensorflow for online and offline processing," in *13. ITG Fachtagung Sprachkommunikation (ITG 2018)*, Oct 2018.

[8] Z. Wang and D. Wang, "Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 457–468, 2019.

[9] V. M. Tavakoli, J. R. Jensen, M. G. Christensen, and J. Benesty, "A framework for speech enhancement with ad hoc microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 6, pp. 1038–1051, 2016.

[10] C. Zorila, C. Boeddeker, R. Doddipatla, and R. Haeb-Umbach, "An investigation into the effectiveness of enhancement in asr training and test for chime-5 dinner party transcription," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019.

[11] Z.-Q. Wang and D. Wang, "On spatial features for supervised speech separation and its application to beamforming and robust asr," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5709–5713.

[12] N. Kanda, Y. Fujita, S. Horiguchi, R. Ikeshita, K. Nagamatsu, and S. Watanabe, "Acoustic modeling for distant multi-talker speech recognition with single- and multi-channel branches," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6630–6634.