The NPU System for DASR Task of CHiME-7 Challenge

Bingshen Mu[1], *Pengcheng Guo*[1], *He Wang*[1], *Yangze Li*[1], *Yang Li*[2], *Pan Zhou*[2], *Wei Chen*[2], *Lei Xie*[1*]

[1]Audio, Speech and Language Processing Group (ASLP@NPU), School of Computer Science,
Northwestern Polytechnical University, Xi'an, China
[2]Space AI, Li Auto

bsmu@mail.nwpu.edu.cn, lxie@nwpu.edu.cn

Abstract

This study describes the NPU system for the Distant Automatic Speech Recognition (DASR) task of the CHiME-7 Challenge. Specifically, two attention-based channel selection modules are introduced to automatically select the most advantageous channel subset from multiple signal channels. Furthermore, we incorporate additional spatial features during the cross-channel attention, which guides the model to capture the desired signals while suppressing the interference sources. It is noteworthy that these enhancements solely pertain to the ASR model, with no modifications made to the speaker diarization (SD). Our approach achieves a Macro diarization attributed word error rate (DA-WER) of 22.28% on CHiME-7 dev sets with oracle diarization and 41.04% on CHiME-7 dev sets with baseline SD results.

Index Terms: Distant automatic speech recognition, channel selection, multi-channel, spatial information

1. Introduction

With the advancements in deep learning, automatic speech recognition (ASR) has made significant progress, leading to noticeable improvements across various speech applications. However, ASR systems continue to encounter challenges in real-world distant scenarios characterized by factors like background noise, reverberation, speaker overlap, and diverse array topologies. To tackle these challenges, the CHiME Challenge series [1, 2, 3] has been established to boost the development of robust ASR systems by promoting research and innovation in multi-microphone signal processing algorithms.

The CHiME-7 Distant Automatic Speech Recognition (DASR) task this year focuses on designing a system that can generalize across various array geometries and provide reliable recognition performance in a wide range of real-world settings, even under adverse acoustic conditions [3]. In this task, multiple recording devices are used to capture audio from different spatial locations simultaneously, enabling a better coverage of the sound source. However, effectively fusing the information from different channels remains a challenge. Besides, some of the far-field microphone arrays or channels may be contaminated by background noise, resulting in significant degradation of ASR performance.

Automatic channel selection proves to be a potent strategy for selecting the most promising subset of microphones for each utterance. An inherent advantage of this method lies in its capacity to transcend various array configurations and application contexts. Historically, channel selection methodologies can be categorized into two primary groups: signal-based and decoder-based measures. It's important to note that prior channel selection techniques necessitated either audio preprocessing or post-processing of ASR outcomes, culminating in protracted and unwieldy processing pipelines. To mitigate the impact posed by noisy channels and enhance the utilization of multi-channel audio, we first propose two attention-based channel selection modules: coarse-grained channel selection (CGCS) and fine-grained channel selection (FGCS). These modules learn to assign higher weights to channels that are beneficial for ASR while assigning lower weights to channels that are detrimental to ASR.

Recently, the concept of cross-channel attention has emerged as a novel method for directly harnessing multi-channel signals within neural speech recognition systems. This approach circumvents the intricacies associated with front-end processing and integrates beamforming and acoustic modeling into a comprehensive end-to-end neural solution. In this cross-channel attention framework, frame-wise multi-channel signals serve as input, enabling the learning of global correlations between sequences originating from distinct channels. The multi-frame cross-channel attention (MFCCA) mechanism, as detailed in the reference [4], adeptly captures both channel-specific and frame-specific information concurrently. It places heightened emphasis on channel context between contiguous frames, thereby effectively modeling dependencies that pertain to both individual frames and entire channels. Considering that MFCCA inherently captures spatial information between channels via the attention mechanism, we have introduced an augmentation to MFCCA by integrating supplementary inter-channel spatial attributes, including Inter-Channel Phase Difference (IPD) [5, 6, 7, 8]. These spatial characteristics serve as guiding cues for the model, facilitating the discernment of target signals while concurrently mitigating interference from extraneous sources.

After combining the results of various systems by the Recognizer Output Voting Error Reduction (ROVER) [9] technique, we achieve a final Macro DA-WER of 22.28% on the dev sets with oracle diarization and 41.04% on the dev sets with baseline speaker diarization (SD) results.

2. Proposed system

2.1. Data processing

In Figure 1, we illustrate the progression of our data processing workflow. This involves the utilization of three distinct datasets: CHiME-6 [2], characterized by a dinner party scenario, encompassing distant speech captured by 6 Kinect array devices, each equipped with 4 microphones, amounting to a total of 24 microphones; DiPCo [10], similarly featuring a dinner party setting, with distant speech recorded via 5 far-field devices, each boasting a 7-microphone circular array, summing to a total of 35 microphones; and Mixer 6 Speech [11], which portrays a meet-

[*] : Corresponding author.

ing scenario, with recordings gathered from 14 microphones of varied styles. Due to the substantial array and channel configurations within each dataset, it becomes imperative to undertake preprocessing of the data. The three datasets undergo initial preprocessing employing the weighted prediction error (WPE) [12] and guided source separation (GSS) [13] algorithms. This preprocessing step yields enhanced clean signals for each individual utterance. Subsequent to the WPE processing, the multichannel audio derived from multiple arrays undergo a transformation, facilitated by the array-based BeamformIt [14] algorithm. This transformation results in converting the multi-array multi-channel audio into a unified single multi-channel audio format, where the number of channels corresponds to the number of arrays.

2.2. Attention-based CGCS

The CGCS process is executed by evaluating the abundance of semantic information encapsulated within each channel of the multi-channel audio. We utilize the Gated Recurrent Unit (GRU) [15] network as the audio feature extractor (AFE). The final hidden state of this GRU network serves as the feature representation for the entire audio content of each channel. To extract semantically relevant audio features, we employ the CTC loss function to guide the audio feature extractor. Fig. 2 illustrates the input for CGCS. The query $\mathbf{A}_{\text{GSS}}$ and key $\mathbf{A}_{\text{WPE+BF}}$ in the attention mechanism are features extracted from the GSS and WPE+BF audio by the WavLM [16] and AFE, respectively. For each encoder layer, the value $V(l)$ is the output of the previous encoder layer, while the value $V(0)$ is $\mathbf{X}_{\text{WPE+BF}}$ extracted from the WPE+BF audio by the WavLM when l equals 0.

2.3. MFCCA with cosIPD features

Spatial information assumes paramount significance in the context of multi-channel scenarios. However, it's worth noting that the MFCCA framework, while potent in its own right, does not possess explicit spatial features as inputs. It relies exclusively on multi-channel audio features for input, thereby enabling MFCCA to implicitly acquire spatial information. To address this limitation, we incorporate the cosIPD features into our model to leverage spatial information. Specifically, we concatenate the cosIPD features with the multi-channel audio features and utilize the MFCCA to better learn spatial information. We concatenate the output of CGCS with the cosIPD features $\mathbf{C}_{\text{WPE+BF}}$, extracted by the cosIPD feature extractor (CFE) from the WPE+BF audio. This concatenated representation is then used as the key and value in the FGCS attention mechanism.

2.4. Attention-based FGCS

CGCS undertakes the filtration of multi-channel audio through the exploitation of semantic information encompassing the entirety of each channel's audio content. It discerns and selects channels characterized by substantial semantic content. In contrast, FGCS centers its attention on computing frame-level resemblances between the GSS audio features and the multichannel audio features. It identifies and selects multi-channel audio frames that exhibit similarity to each frame of the GSS audio. Fig. 2 illustrates the input for FGCS. The query $\mathbf{X}_{\text{GSS}}$ in the attention mechanism is the feature extracted from the GSS audio by the WavLM, while the key and value are obtained from the results of CGCS, which include the cosIPD features.

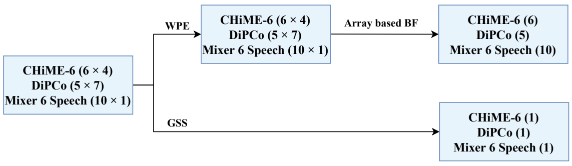

Figure 1: *The flow chart of data processing. (N × M) refers to N array devices each with M microphones before data processing. (K) refers to K channels after data processing.*

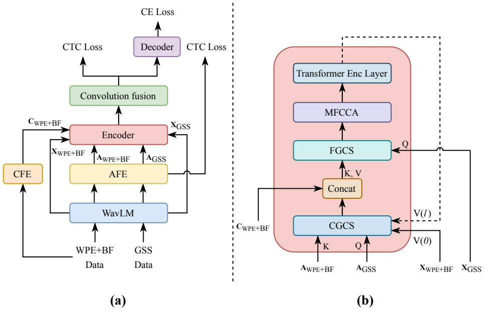

Figure 2: *(a) An overview of the proposed ASR model, where "AFE" denotes the audio feature extractor and "CFE" denotes the cosIPD feature extractor. (b) A detailed description of each Encoder layer, where the subscript "l" refers to layer index.*

2.5. Inference procedure

During the inference, the dev and eval sets are first segmented by the baseline SD model results for the main track and oracle diarization for the sub-track, enhanced by WPE, BF, and GSS, transcribed by our ASR models, and rescored by a Transformer based language model (LM) trained on a combination of the CHiME-7 and LibriSpeech [17] corpora. We also tune the decoding parameters, including beam size, CTC weight, and LM weight, during the decoding process. The results from different models are fused by the ROVER technique finally.

3. Experiments

3.1. Setup

All of the models are implemented with the ESPnet [18] toolkit. We follow the setup of the CHiME-7 baseline ASR model to build our model, which consists of a WavLM frontend, a 12-layer Transformer encoder, and a 6-layer transformer decoder. The dimensions of MHSA and FFN layers are set to 256 and 2048, respectively. All of our models are trained on the full CHiME-7 train sets which are segmented by oracle segmentation and processed following the procedures outlined in 1. Besides, all of our systems are rescored by a transformer-based language model (LM) trained on a combination of the CHiME-7 and LibriSpeech corpora. During the training, we freeze the parameters of the WavLM. AFE is initialized with a well-trained ASR model that utilizes a bi-directional GRU encoder trained solely with the CTC loss. The spatial feature cosIPD is extracted with window length, frameshift, and STFT length are 32ms, 16ms, and 512, respectively.

Table 1: *The DER(%) and JER(%) results of baseline SD system on the dev sets.*

Scenario	DER	JER
CHiME-6	39.97	51.19
DiPCo	29.85	41.41
Mixer 6	16.56	12.78
Macro	28.79	38.46

Table 2: *The main track DA-WER(%) results of ASR models on the dev sets segmented by baseline SD results.*

ASR model	Scenario			Macro
	CHiME-6	DiPCo	Mixer 6	
Baseline	62.40	56.64	22.58	47.21
MFCCA	59.51	58.04	24.23	47.26
+CGCS	57.71	51.82	19.23	42.92
+FGCS	56.75	52.39	18.42	42.52
+cosIPD	57.91	52.98	18.51	43.13
+ALL	56.14	52.10	18.33	42.19
ROVER	**55.17**	**50.46**	**17.49**	**41.04**

3.2. Results

Table 1 presents the results achieved by the baseline SD system in terms of diarization error rate (DER) and Jaccard error rate (JER) with 250 ms collar. The DER and JER results are in line with the findings reported in [3]. Table 2 presents the DA-WER results of various ASR models on the main track. Table 3 presents the DA-WER results of various ASR models on the sub-track. From Tables 2 and 3, it can be seen that all of our models give better results over the official baseline and achieve up to 22.66% relative Macro DA-WER improvement on the sub-track and 13.05% relative Macro DA-WER improvement on the main track due to the effective channel selection and spatial information utilization strategies. Among the array of proposed methods, FGCS stands out as the most impactful. Its effectiveness stems from its dual capability: not only does FGCS identify the highly correlated features within the multichannel audio data compared to the GSS audio features, but it also actively harnesses the inter-channel spatial information. Furthermore, CGCS, which selects channels based on the richness of semantic information across different channels, emerges as the second most effective approach after FGCS. Additionally, the inclusion of cosIPD features significantly bolsters the ASR model's capacity to discern spatial information. We achieve the best performance by incorporating all the proposed methods into the ASR system.

Table 3: *The sub-track DA-WER(%) results of ASR models on the dev sets segmented by oracle diarization.*

ASR model	Scenario			Macro
	CHiME-6	DiPCo	Mixer 6	
Baseline	32.64	33.54	20.25	28.81
MFCCA	31.13	34.37	21.73	29.07
+CGCS	28.86	30.57	18.60	26.01
+FGCS	28.24	30.41	16.19	24.95
+cosIPD	29.15	30.79	15.92	25.29
+ALL	27.66	29.14	14.75	23.85
ROVER	**25.58**	**27.34**	**13.92**	**22.28**

4. Conclusions

In this study, we describe our system for DASR Task of CHiME-7 Challenge. Our efforts include data processing, channel selection, and spatial information fusion strategies. By combining various systems, we get a Macro DA-WER of 41.04% on the main track and 22.28% on the sub-track.

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 *Proc. Interspeech 2018*, 2018, pp. 1561–1565.

[2] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj *et al.*, "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," in *Proc. CHiME Workshop*, 2020.

[3] S. Cornell, M. Wiesner, S. Watanabe, D. Raj, X. Chang, P. Garcia, Y. Masuyama, Z.-Q. Wang, S. Squartini, and S. Khudanpur, "The CHiME-7 DASR Challenge: Distant Meeting Transcription with Multiple Devices in Diverse Scenarios," *arXiv preprint arXiv:2306.13734*, 2023.

[4] F. Yu, S. Zhang, P. Guo, Y. Liang, Z. Du, Y. Lin, and L. Xie, "MFCCA: Multi-Frame Cross-Channel attention for multi-speaker ASR in Multi-party meeting scenario," in *Proc. SLT. IEEE*, 2023, pp. 144–151.

[5] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li, and Y. Gong, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *Proc. SLT. IEEE*, 2018, pp. 558–565.

[6] T. Yoshioka, H. Erdogan, Z. Chen, and F. Alleva, "Multi-microphone neural speech separation for far-field multi-talker speech recognition," in *Proc. ICASSP. IEEE*, 2018, pp. 5739–5743.

[7] R. Gu, J. Wu, S.-X. Zhang, L. Chen, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "End-to-end multi-channel speech separation," *arXiv preprint arXiv:1905.06286*, 2019.

[8] L. Shubo, Y. Fu, J. Yukai, L. Xie, W. Zhu, W. Rao, and Y. Wang, "Spatial-DCCRN: DCCRN Equipped with Frame-Level Angle Feature and Hybrid Filtering for Multi-Channel Speech Enhancement," in *Proc. SLT. IEEE*, 2023, pp. 436–443.

[9] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proc. ASRU. IEEE*, 1997, pp. 347–354.

[10] M. Van Segbroeck, A. Zaid, K. Kutsenko, C. Huerta, T. Nguyen, X. Luo, B. Hoffmeister, J. Trmal, M. Omologo, and R. Maas, "DiPCo–Dinner Party Corpus," *arXiv preprint arXiv:1909.13447*, 2019.

[11] L. Brandschain, D. Graff, C. Cieri, K. Walker, C. Caruso, and A. Neely, "The mixer 6 corpus: Resources for cross-channel and text independent speaker recognition," in *Proc. LREC*, 2010.

[12] L. Drude, J. Heymann, C. Böddeker, and R. Haeb-Umbach, "NARA-WPE: A python package for weighted prediction error dereverberation in numpy and tensorflow for online and offline processing," in *Proc. ITG Symposium on Speech Communication*. VDE / IEEE, 2018, pp. 1–5.

[13] C. Boeddeker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, and R. Haeb-Umbach, "Front-end processing for the CHiME-5 dinner party scenario," in *Proc. CHiME5 Workshop*, 2018.

[14] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE/ACM TASLP*, vol. 15, no. 7, pp. 2011–2022, 2007.

[15] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. EMNLP*, 2014.

[16] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE/JSTSP*, vol. 16, no. 6, pp. 1505–1518, 2022.

[17] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proc. ICASSP. IEEE*, 2015, pp. 5206–5210.

[18] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "Espnet: End-to-end speech processing toolkit," in *Proc. Interspeech. ISCA*, 2018, pp. 2207–2211.