System Description of NJU-AALab's Submission for the CHiME-8 NOTSOFAR-1 Challenge

Qinwen Hu[1,2,], Tianchi Sun[1,2,*], Xin'an Chen[1,2], Xiaobin Rong[1,2], Jing Lu[1,2]*

[1]*Key Laboratory of Modern Acoustics, Nanjing University, Nanjing 210093, China*
[2]*NJU-Horizon Intelligent Audio Lab, Horizon Robotics, Beijing 100094, China*

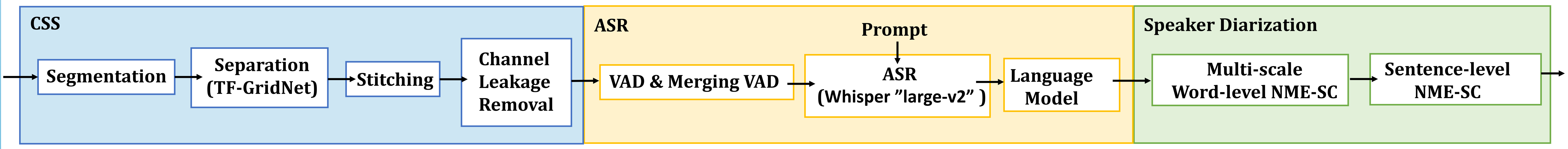
[*] *These authors contribute equally to this work.*

INTRODUCTION

Our submission to the NOTSOFAR-1 task [1] is a modularized system consisting of sequential continuous speech separation (CSS), automatic speech recognition (ASR) and speaker diarization modules. The CSS system processes mixed signals in a streaming fashion, implicitly detecting overlaps and separating overlapped speech into different streams. The outputs of the CSS system are then regularized and fed into the ASR system to acquire transcriptions with word-level time boundaries for each stream. Finally, we apply speaker diarization based on the CSS results and ASR transcriptions. Our entry achieves a tcpWER of 33.5% on the evaluation set and 36.4% on the development set.

System Description

System Overview



The CSS system

The CSS system is structured within a segmentation-separation-stitching processing scheme, where the separation step is conducted with a conventional speech separation model, a lite version of TF-GridNet [2]. It is trained with a fixed input length in the segment-level permutation-invariant-training (PIT) manner using signal-to-noise-ratio (SNR) loss. We use the 200-hour version of the simulated dataset from the NOTSOFAR-1 task to train TF-GridNet. We filter out samples where more than one speaker is present in a single target stream. Following the baseline, TF-GridNet outputs 3 speech streams and 1 noise stream. During inference, the input stream is separated into 4-second segments with a hop length of 2 seconds, processed by TF-GridNet. Then we stitch the estimated segments based on the alignment of the overlapped region and apply energy-based channel leakage removal afterwards.

The ASR system

For ASR, we apply transcription with Whisper "large-v2" [3] independently to each audio stream produced by CSS. We perform a series of pre-processing steps on the output of CSS system before sending the signal to the ASR module. We apply voice active detection (VAD) using MarbleNet [4] from the NeMo toolkit on the signal to get speech segments and remove non-speech frames to avoid hallucination of ASR. We concatenate short speech segments into a single segment to provide more context, ensuring that the total length of the concatenated segment does not exceed 26 seconds.

We further improve the accuracy of ASR results by providing a prompt to the ASR system to avoid word omissions in segments with repetitions, and apply a pretrained language model (LM), the BERT [5] model to rescore the transcribed results. The ASR model and LM are applied to all test datasets without any fine-tuning.

The diarization system

We perform an offline speaker diarization approach on the CSS output streams leveraging time boundaries of words obtained with Whisper "large-v2". We extract multiple-scale speaker embedding vectors for each word following the baseline. Each scale corresponds to different window lengths and the final affinity matrix is the average of the affinity matrices of all the scales. The pre-trained TitaNet [6] from the NeMo toolkit is implemented as the speaker embedding model.

Then, offline clustering is performed by using the normalized maximum eigengap based spectral clustering (NME-SC) [7] algorithm, to assign a speaker label to each ASR word. The results from the word-level NME-SC algorithm are considered preliminary diarization results. Then, all the words undergo deduplication to suppress duplicate context in different streams caused by channel leakage. In each stream, words belonging to the same speaker and with intervals less than 0.5s are concatenated into sentences. Finally, speaker embedding vectors extracted for each sentence are processed by sentence-level NME-SC to generate the fine-tuned diarization results.

RESULTS AND ANALYSIS

Table 1 illustrates the effectiveness of each technique applied to ASR processing. The implementation of VAD, merging VAD, and using prompts all contribute to improvements in ASR accuracy. Although the usage of a LM results in a slight decrease in accuracy for the "plaza_0" device data, it leads to an overall increase in accuracy across the entire single-channel test set.

VAD	Merging VAD	Prompt	LM	tcorcWER
×	×	×	×	37.4%
✓	×	×	×	36.3%
✓	✓	×	×	33.3%
✓	✓	✓	×	**31.0%**
✓	✓	✓	✓	31.1%

Table 1: The ablation study of ASR techniques on the development set using recordings from the "plaza_0" device. We use the TF-GridNet model as the CSS model, Whisper "large-v2" as the ASR model, and the "word-nmesc" approach for diarization.

Table 2 presents the results of the ablation study on diarization approaches. Sentence-level diarization further reduces the tcpWER.

Sentence-level Diarization	tcpWER	DER
×	35.1%	**17.0%**
✓	**34.0%**	17.0%

Table 2: The ablation study of diarization methods on the development set using recordings from the "plaza_0" device. We use the TF-GridNet model as the CSS model, Whisper "large-v2" as the ASR model.

System	tcpWER	tcorcWER
Baseline	45.8%	38.6%
Submitted System	**36.4%**	**33.2%**

Table 3: The recognition and diarization scores on the development set of NOTSOFAR-1 calculated on all sessions of single-channel devices.

System	tcpWER	tcorcWER
Baseline	41.4 %	35.5%
Submitted System	**33.5 %**	**30.4%**

Table 4: The recognition and diarization scores on the evaluation set of NOTSOFAR-1 calculated on all sessions of single-channel devices.

Tables 3 and 4 present the tcpWER and tcorcWER for the single-channel data from the development and evaluation set respectively. The results demonstrate that integrating all the proposed techniques into the meeting transcription pipeline yields significant improvements over the baseline. Enhancements to the CSS, ASR, and speaker diarization modules all contribute to reducing the tcpWER metric. However, there remains considerable room for improvement in this system. For instance, the CSS module struggles with generalization when handling single-channel data corrupted by frontend signal processing. Future work will focus on fine-tuning the system's performance using real-world data.

REFERENCES

[1] Alon Vinnikov, Amir Ivry, Aviv Hurvitz, Igor Abramovski, Sharon Koubi, Ilya Gurvich, Shai Peer, Xiong Xiao, Benjamin Martinez Elizalde, Naoyuki Kanda, et al., "Notsofar-1 challenge: New datasets baseline, and tasks for distant meeting transcription," *arXiv preprint arXiv:2401.08887*, 2024.

[2] Zhong-Qiu Wang, Samuele Cornell, Shukjae Choi, Youn glo Lee, Byeong-Yeol Kim, and Shinji Watanabe, "Tf-gridnet: Integrating full-and sub-band modeling for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[3] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28492–28518.

[4] Fei Jia, Somshubra Majumdar, and Boris Ginsburg, Marblenet: Deep 1d time-channel separable convolutional neural network for voice activity detection," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6818–6822.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: pre-training of deep bidirec tional ransformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.

[6] Nithin Rao Koluguri, Taejin Park, and Boris Ginsburg, "Titanet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8102–8106.

[7] Tae Jin Park, Kyu J Han, Manoj Kumar, and Shrikanth Narayanan, "Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap," *IEEE Signal Processing Letters*, vol. 27, pp. 381–385, 2019.