



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}

¹Key Laboratory of Modern Acoustics, Nanjing University, Nanjing 210093, China

²NJU-Horizon Intelligent Audio Lab, Horizon Robotics, Beijing 100094, China

{qinwen.hu, tianchi.sun, xinan.chen, xiaobin.rong}@smail.nju.edu.cn, lujing@nju.edu.cn

Abstract

The paper describes the NJU-AALab team’s entry to the Natural Office Talkers in Settings of Far-field Audio Recordings (NOTSOFAR-1) task, part of the CHiME-8 Challenge. The approach uses a pipeline consisting of a continuous speech separation (CSS) module based on TF-GridNet, the state-of-the-art speech separation model, a speech recognition module utilizing Whisper ”large-v2”, and a speaker diarization module based on the multi-level normalized maximum eigengap-based spectral clustering (NME-SC) method. Our proposed system achieves a time-constrained minimum permutation word error rate (tcpWER) of 33.5% on the evaluation set and 36.4% on the development set of the NOTSOFAR-1 real recordings, which outperforms the baseline by a large margin and ranks 3rd in the single track of the challenge.

Index Terms: Continuous speech separation, speaker diarization, ASR, CHiME-8 NOTSOFAR-1

1. Introduction

The NOTSOFAR-1 task [1] of CHiME-8 Challenge focuses on speaker diarization and automatic speech recognition (ASR) in real-world scenarios. The task comprises two tracks: (1) the single-channel device track and (2) the known-geometry multi-channel device track. Participants are required to submit speaker-diarized transcriptions inferred using recordings from a single device. The evaluation set of the NOTSOFAR-1 task presents many challenges for existing speaker diarization and recognition systems due to its realistic features, such as varying and complicated acoustic environments and diverse speaking dynamics. Moreover, the single-channel track emphasizes the practical scenario in that the outcome is already the processing result from the microphone array and may include speech distortion, audio clipping and notch distortion.

Diarization-and-Recognition systems designed to handle multi-talker conversations usually utilize modularized methods composed of CSS, speaker diarization, and ASR. The pipeline containing these modules can be

flexible. One promising approach is to perform CSS first to handle overlapping speech, followed by ASR based on the CSS output. Finally, speaker diarization is conducted on the CSS output using boundary information from ASR outputs, allowing the diarization to benefit from syntactic information. Our system follows this CSS-ASR-Diarization pipeline.

Our submitted system for the single-channel track achieves a tcpWER of 33.5% on the evaluation set and 36.4% on the development set of NOTSOFAR-1. Training the speech separation model takes approximately 3 days using six RTX 3090 GPUs, and the entire inference process for the evaluation set takes 2 hours with four RTX 3090 GPUs.

2. System Description

2.1. System Overview

Figure 1 shows the diagram of the pipeline of our system. The overall structure of our system follows the baseline of the NOTSOFAR-1, consisting of a CSS module, an ASR

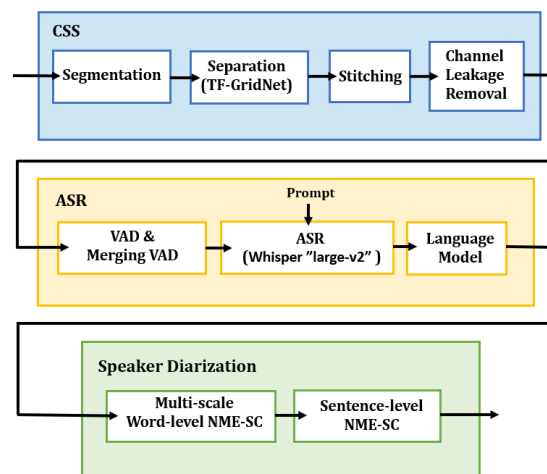


Figure 1: The schematic diagram of the proposed system.

* The authors contribute equally to this work.

module, and a speaker diarization module sequentially. The CSS module processes mixed signals in a streaming fashion, implicitly detecting overlaps and separating overlapped speech into different streams. The outputs of the CSS module are then regularized and fed into the ASR module to acquire transcriptions with time boundaries for each stream. Finally, we apply speaker diarization based on the CSS results and ASR transcriptions.

2.2. CSS Module

The CSS module is structured within a segmentation-separation-stitching processing scheme, where the separation step is conducted using a conventional speech separation model trained on audio clips with a fixed input length. Following the previous work [2] on multi-talker meeting transcription, we adopt TF-GridNet [3] as our separation model, which has demonstrated state-of-the-art performance in separating fully overlapped speech through a complex spectrogram mapping-based architecture.

The separator is trained with the signal-to-noise-ratio (SNR) loss in the segment-level permutation-invariant-training (PIT) manner. Speech separation models often encounter the issue of channel leakage, which can be categorized into under-separation and over-separation [4], occurring in overlapping and single-speaker regions, respectively. For under-separation, TF-GridNet demonstrates superior performance in overlapping scenarios, resulting in less channel leakage at overlapping frames compared to the baseline. This type of channel leakage is challenging to eliminate using post-processing methods. To address channel leakage in single-speaker regions caused by over-separation, we implement a two-step removal process using energy-based speech presence probability (SPP), as detailed in Section 3.1.

In line with the baseline, TF-GridNet outputs four streams, three for speech and one for noise. During the inference stage, the input mixed signal is segmented by a sliding-window [5] into overlapping segments consisting of a fixed number of frames. The frames within the current segment undergo processing by TF-GridNet to generate output streams, with the sliding-window moving forward each time. Then we stitch the estimated segments based on the alignment of the overlapped region.

2.3. ASR Module

For ASR, we employ Whisper "large-v2" [6], which supports word-level timestamps. ASR is applied independently to each audio stream produced by CSS. However, Whisper suffers from hallucination, meaning it will create transcripts even during silence frames. Therefore, it is necessary to perform voice activity detection (VAD) on the CSS output first to isolate speech frames. Consider-

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 and Whisper "large-v2" as the ASR model.*

Version	VAD	Merging	Prompt	LM	tcorcWER
v1	×	×	×	×	37.6%
v2	✓	×	×	×	36.1%
v3	✓	✓	×	×	33.0%
v4	✓	✓	✓	×	31.0%
v5	✓	✓	✓	✓	31.1%

ing that too short segments are prone to errors due to the lack of contextual information, we merge short speech segments into longer ones as the input for Whisper.

Additionally, in real-world meeting scenarios, there are many hesitations, filler words and word repetitions. An appropriate prompt, which includes manually added redundant words and filler phrases, has proven effective in improving the situation. Furthermore, the transcripts generated by Whisper undergo an additional refinement process through rescoreing with a pre-trained language model (LM). This rescoreing step leverages the LM's contextual understanding to correct potential errors in the initial transcription, improving overall accuracy and fluency. By incorporating the LM, the system can better handle ambiguous or unclear speech segments.

Table 1 shows the effectiveness of every technique we apply to the ASR processing. Time-constrained optimal reference combination word error rate (tcorcWER) metric is used here to validate effectiveness of the techniques on speech recognition alone, without considering their impact on subsequent speaker diarization. Although LM rescoreing worsens the tcorcWER metric of v5 compared to v4, our subsequent experiments show that it can improve the tcpWER. Therefore, we ultimately submit the version that incorporates all ASR techniques.

2.4. Speaker Diarization Module

As per the post-ASR diarization method presented in the baseline system, multi-scale speaker embedding vectors are extracted for each word using the word boundaries obtained from the ASR results, and the affinity matrix is formed by calculating the cosine similarity between the words. Each scale corresponds to different window lengths and the final affinity matrix is the average of the affinity matrices from all scales. It should be noted that when the window length surpasses the duration of a word, the timestamp of the word extends to both sides. Then, offline clustering is performed using the NME-SC [7] algorithm, which employs eigengap analysis to determine a threshold for affinity binarization and to estimate the number of speakers in the meeting recording. K-means

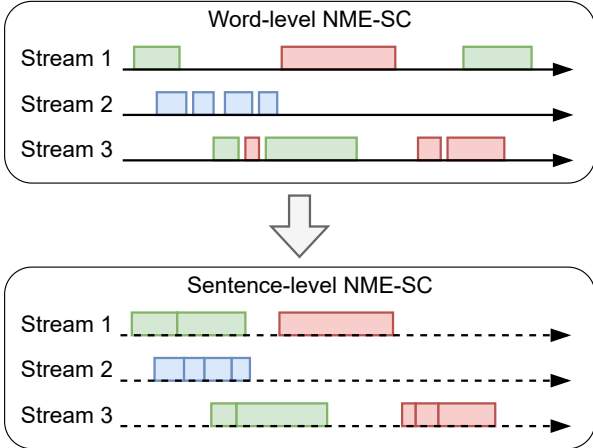


Figure 2: *The diagram of multi-level NME-SC speaker diarization. Different colors represent different speakers.*

clustering is then used to assign a speaker label to each word.

Leveraging the word-level timestamps provided by the ASR model, speaker diarization for an unknown number of speakers is simplified compared to the end-to-end approach. However, in overlapping speech segments, the speaker embedding extracted for a word can become confusing if a window spans a speaker turn. Besides, speaker embeddings are considered to become uninformative if computed on segments shorter than 2 seconds [8]. Although the baseline system includes deduplication to mitigate channel leakage by deleting recognized words that appear in multiple output streams simultaneously, we have made further improvements. As illustrated in Fig. 2, the results from the word-level NME-SC algorithm are considered preliminary diarization results. In each audio stream, we concatenate the consecutive words belonging to the same speaker to form sentences. Then, the speaker embedding vectors extracted for each sentence are clustered to obtain the final diarization results, which can be considered as sentence-level NME-SC. Table 2 presents the tcpWER and diarization error rate (DER) results from the ablation study on diarization approaches.

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 and Whisper "large-v2" as the ASR model.*

Sentence-level diarization	tcpWER	DER
×	35.1%	17.0%
✓	34.0%	17.0%

3. Experimental Configurations

3.1. Training Datasets and Configurations for CSS

We use the 200-hour version of the simulated dataset from the NOTSOFAR-1 task to train the CSS model. In the data cleaning stage, audio samples where more than one speaker is present in a single target stream are filtered out. The TF-GridNet model implemented in our experiments consists of 5 TF-Grid blocks, each with a hidden size of 112 across the 3 modules in the block. Further details about the network can be found in [3]. The model is trained on data with a sampling rate of 16 kHz, and the short-time Fourier transform is performed with a 32ms frame length and a 16ms frame shift. We train the model with an initial learning rate of 0.001, which is halved if no improvement is observed on the validation set after every 10 validation steps.

During the CSS model’s inference stage, the input stream is segmented into 4-second segments with a hop length of 2 seconds. We align adjacent segments in the stream dimension by minimizing the mean square error (MSE) in the overlapping regions.

In most cases, the separation model outputs exhibit channel leakage of the over-separation type. To address this issue, we employ a two-step masking method. First, we calculate the relative energy ratios of each stream to the total energy of the three streams, frame by frame, on the output spectrograms to estimate an energy-based SPP. Next, we apply smoothing to the SPP results, setting the energy of frames without prominent speech to zero. However, some low-level residual noise may still remain in frames where no speech is present in the original mixture. Additionally, we recalculate the smoothed SPP results based on the absolute energy and repeat the previous operations. This two-step, SPP-based binary masking can effectively remove most of the channel leakage and streamline the process of performing further VAD to prepare the speech segments for the ASR module.

3.2. Configurations for ASR

It should be noted that the CSS output has already undergone binary masking, any simple VAD method can be used to obtain the start and end times of speech. In our system, we implement the MarbleNet [9] from the NeMo toolkit as the VAD network, to ignore silence segments and combine the speech segments separated by a short pause. During Whisper’s inference process, short speech segments from each output stream are merged into longer segments, not exceeding 26 seconds, to be transcribed and rescored by a pre-trained BERT [10] model. To avoid word omissions in segments with repetitions, we provide

Table 3: The *tcpWER* and *tcorcWER* scores on the development (*Dev*) and evaluation (*Eval*) set of NOTSOFAR-1 calculated on all sessions of single-channel devices.

System	<i>tcpWER</i>		<i>tcorcWER</i>	
	Dev	Eval	Dev	Eval
Baseline	45.8%	41.4%	38.6%	35.5%
Submission	36.4%	33.5%	33.2%	30.4%

Whisper with the prompt suggested in this page¹. The ASR model and LM are applied to all test datasets without any fine-tuning.

3.3. Configurations for Speaker Diarization

During the diarization process, we utilize five different window lengths—2.5, 2.0, 1.5, 1.0, and 0.5 seconds—to extract speaker embedding vectors for each word based on the word timestamps. The pre-trained TitaNet [11] from the NeMo toolkit is implemented as the speaker embedding model, consistent with the baseline. After the preliminary word-level NME-SC, each word is assigned a speaker label. 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 0.5 seconds are concatenated into sentences. Finally, the speaker embedding vectors extracted for each sentence are processed using NME-SC, assigning a speaker label to each sentence. This sentence-level NME-SC approach yields fine-tuned diarization results.

4. Results and Analysis

Table 3 presents the *tcpWER* and *tcorcWER* scores on the development and evaluation sets of the single-channel track. The results obtained demonstrate that our proposed modifications to the meeting transcription pipeline lead to significant improvements compared to the baseline. Enhancements in the CSS, ASR, and speaker diarization modules all contribute to a reduction in the *tcpWER* metric. Ultimately, our system achieves a *tcpWER* of 33.5% on the evaluation set, reflecting a 19.2% reduction from the baseline (41.4%).

Nevertheless, there remains substantial room for further improvement. For instance, the CSS module has difficulty generalizing to single-channel data corrupted by front-end signal processing. Enhancing the system’s performance through fine-tuning with real-world data will be a focus of our future work.

¹<https://github.com/openai/whisper/discussions/2003>

5. Conclusions

In this paper, we summarize the approach of NJU-AALab’s entry to the CHiME-8 NOTSOFAR-1 Task. Our system is composed of a TF-GridNet based CSS module, an ASR module utilizing the Whisper “large-v2” model, and a multi-level NME-SC speaker diarization module. The proposed system demonstrates robust performance on realistic conversational speech data, achieving a *tcpWER* of 33.5% on the evaluation set of the NOTSOFAR-1 single-channel track.

6. References

- [1] A. Vinnikov, A. Ivry, A. Hurvitz, I. Abramovski, S. Koubi, I. Gurvich, S. Peer, X. Xiao, B. M. Elizalde, N. Kanda, X. Wang, S. Shaer, S. Yagev, Y. Asher, S. Sivasankaran, Y. Gong, M. Tang, H. Wang, and E. Krupka, “Notsofar-1 challenge: New datasets, baseline, and tasks for distant meeting transcription,” in *Interspeech 2024*, 2024, pp. 5003–5007.
- [2] T. Von Neumann, C. Boeddeker, T. Cord-Landwehr, M. Delcroix, and R. Haeb-Umbach, “Meeting recognition with continuous speech separation and transcription-supported diarization,” in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE, 2024, pp. 775–779.
- [3] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, “Tf-gridnet: Integrating full-and sub-band modeling for speech separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.
- [4] Y. Bando, T. Nakamura, and S. Watanabe, “Neural blind source separation and diarization for distant speech recognition,” in *Interspeech 2024*, 2024, pp. 722–726.
- [5] S. Chen, Y. Wu, Z. Chen, J. Wu, J. Li, T. Yoshioka, C. Wang, S. Liu, and M. Zhou, “Continuous speech separation with conformer,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5749–5753.
- [6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.
- [7] T. J. Park, K. J. Han, M. Kumar, and S. Narayanan, “Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap,” *IEEE Signal Processing Letters*, vol. 27, pp. 381–385, 2019.
- [8] T. Zhou, Y. Zhao, and J. Wu, “Resnext and res2net structures for speaker verification,” in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 301–307.
- [9] F. Jia, S. Majumdar, and B. 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.

- [10] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: <http://arxiv.org/abs/1810.04805>
- [11] N. R. Koluguri, T. Park, and B. 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.