



The Fano Labs System for the CHiME-8 NOTSOFAR-1 Challenge Single-channel Track

Samuel J. Broughton, Lahiru Samarakoon, Harrison Zhu

Fano Labs, Hong Kong SAR, China

samuel.broughton@fano.ai

Abstract

This technical report outlines our submissions to the CHiME-8 NOTSOFAR-1 Challenge single-channel track, which focuses on distant speaker diarization and automatic speech recognition. This track evaluates far-field meeting transcriptions on a single audio channel using a single device. Our submission features a highly streamlined and efficient system incorporating both a non-autoregressive speaker diarization and automatic speech recognition model. Each submitted system outperforms the baseline in terms of Time-Constrained minimum Permutation Word Error Rate (tcpWER) on the development set. We provide an analysis on models sizes and inference throughput under constrained computational resources with the most practical system using less than 100 million parameters. Additionally, we report new state-of-the-art results for the AMI Mix and AMI SDM datasets with DER values of 11.83 % and 17.55 %, respectively.

Index Terms: speaker diarization, DASR, ASR

1. Introduction

Distant conversational speech recognition presents a significant challenge, necessitating a robust speaker partition mechanism alongside advanced speech transcription capabilities.

The CHiME (Computational Hearing in Multisource Environments) series of challenges aims to enhance conversational speech recognition in real-world environments [1, 2, 3].

The CHiME-8 NOTSOFAR-1 (Natural Office Talkers in Settings of Far-field Audio Recordings) challenge focuses on single-device distant speaker diarization and automatic speech recognition (DASR) [4]. The challenge features both single-channel and known-geometry multi-channel tracks, using recording equipment that reflects typical commercial setups. The primary motivation is to improve the accuracy of ASR systems in every day situations.

However, focusing solely on accuracy can sometimes lead to systems with significant latency and computational overhead, making them impractical for real-world deployment and everyday usage. Applications such as live broadcasting and meeting transcription demand efficient and prompt responses. Additionally, end users often have limited processing power and may lack access to large-scale computing resources such as GPUs, particularly when systems are required to run on mobile or embedded devices.

To address these issues, CHiME-8 introduces a jury award for the most practical and efficient system [5]. This award encourages the development of systems that strike a balance between high performance and minimal latency and resource consumption, thereby facilitating more accessible and practical applications for everyday use.

The baseline system for NOTSOFAR-1 consists of three key steps: continuous speech separation (CSS), automatic speech recognition (ASR), and speaker diarization [4]. The CSS module generates up to three individual audio streams, each of which is then processed by the Whisper “large-v2” ASR model [6]. Offline speaker diarization is also applied to each of the separated audio streams produced by the CSS module. The final transcription is achieved by aligning each word with its corresponding speaker label, utilizing word boundary information from Whisper and speaker information from diarization.

Performing long-form inference with the Whisper “large-v2” ASR model is resource inefficient due to its approximately 1.5 billion parameters and the quadratic complexity of self-attention based on sequence length. These inefficiencies are particularly notable as CSS module audios often have long periods of silence.

In this technical report, we present the effectiveness of a simple and streamlined system design for single-device single-channel far-field meeting transcription. Our approach integrates a robust end-to-end diarization module that directs segments of active speech to an efficient ASR deployment. That submitted system is highly versatile, capable of operating in non-autoregressive (NAR) and autoregressive (AR) modes, with a total of only two models required to be trained.

2. System Description

2.1. Data Preparation

The training dataset comprises the complete CHiME-8 “Core Datasets” and a selected subset of “Allowed External” data [7].

Datasets employed for ASR model training include the official training splits of LibriSpeech [8], CHiME-6 [2], DiPCo [9], Mixer 6 [10], both NOTSOFAR1 SC (single-channel) and NOTSOFAR1 MC (multi-channel) [4], and both AMI SDM (single distant microphone) and AMI Mix (headset mixtures) [11]¹. All datasets are utilized to train the ASR model from scratch with only the AMI corpus being augmented with external noise data as permitted by the official rules.

The diarization model is first pretrained using the LibriSpeech corpus and then finetuned on the NOTSOFAR1 SC, NOTSOFAR1 MC, AMI-SDM and AMI-Mix datasets. Each dataset contains augmentations as permitted by the rules. Pre-training data is generated using a standard diarization mixture simulation algorithm [12, Alg. 1]. We simulated 400,000 mixtures with 1 to 8 speakers, resulting in 50,000 generated simulations for each possible number of active speakers in a mixture. Similarly to recent work, pretraining is conducted in a single

¹Full-corpus ASR partitions can be found at <https://github.com/BUTSpeechFIT/AMI-diarization-setup>

step on all 400,000 mixtures [13].

For diarization finetuning, datasets containing multi-channel audio recordings are downmixed to a single channel.

2.2. Speaker Diarization

The speaker diarization component of the overall system employs EEND-TA [14]. End-to-end diarization models are straightforward to train and efficient during inference, requiring only a single model for deployment. In contrast, clustering-based diarization methods often rely on multiple components, some of which are trained separately on non-diarization objectives, making them less practical for deployment [15, 16, 17]. Using a pipeline of models to achieve diarization puts a system at risk of increasing latency and computational overheads.

Similar to other other EEND-based models, EEND-TA approaches diarization as a multi-label classification problem. These models directly map an input audio feature sequence to the probability of speech activities for multiple speakers, inherently handling overlapping speech. The backbone of EEND-TA consists of a Conformer encoder [18], which transforms an input audio sequence to a latent low-resolution representation. The Transformer Attractor (TA) module then computes a set of attractors to identify valid speakers. The input queries to the TA module is a combination of conversational summary representations [19] and learnable global embeddings. Diarization results are obtained by computing the matrix product between the latent representation and valid speaker attractors.

The model is trained in 2 stages. The pretraining stage uses simulated mixtures from LibriSpeech, while the finetuning stage employs real data, as detailed in Section 2.1. Models are pretrained for three million steps with a batch size of 64×4 GPUs with an utterance length of 200s. Models are finetuned for 100 epochs with a batch size of 8 on a single GPU with an utterance length of 600 seconds. The Adam optimizer [20] and Noam scheduler [21] with 100,000 warm-up steps was used during pre-training. At fine-tuning the Adam optimizer was used with a fix learning rate of 1×10^{-5} . The maximum number of speakers that TA can predict is set to 8. The model parameters of the best 10 epochs in terms of validation accuracy are averaged for inference.

During inference, diarization results are simply obtained by inputting the entire audio sequence into the model. Speaker existence and diarization thresholds are set to 0.45 and 0.5, respectively.

2.3. Speech Recognition

For speech recognition, we made use of both Whisper ASR models and a model trained by us with only permitted datasets.

Prioritizing accuracy, we directly use Whisper “large-v3” to transcribe audio within the time boundaries specified by the diarization module without any further training. Whisper “large” models are significant in size, boasting 1.55 billion parameters [6]. Compared to Whisper “large-v2”, Whisper “large-v3” uses 128 Mel frequency bins instead of 80 and is trained on an extensive dataset comprising 1 million hours of weakly labeled audio and 4 million hours of pseudo-labeled audio collected using Whisper “large-v2” [22].

Given the Whisper “large-v3” model is parameter inefficient and costly to run inference, we also experimented with the smaller ASR model trained by us. This model uses the joint CTC-attention objective with a CTC weight of 0.3 [23]. It was trained using the ESPnet package [24] and comprises a Uconv-Conformer Encoder [25] along with a Transformer decoder,

consisting of 12 and 6 layers respectively. Uconv-Conformer includes 3 down-sampling blocks and 1 up-sampling block, with each block consisting of 4 layers, and there is also a skip connection between blocks 2 and 4. Each attention layer in the encoder-decoder architecture has 512 hidden units and 8 attention heads and each feedforward layer consists of 2048 units. The depth-wise convolutional layers of the conformer have a kernel size of 5.

Model is trained for 100 epochs on 80 dimensional Mel-scale filterbank features using the data outlined in Section 2.1. Network parameters are trained using the Adam optimizer [20] with a warm-up of 40,000 steps.

The final model is obtained by averaging the model parameters of the top 10 epochs with the highest validation accuracy. Inferences are performed using both AR and NAR mechanisms. AR decoding utilises beam search with the Transformer decoder, while NAR decoding involves removing the Transformer decoder layers and applying CTC-greedy decoding. NAR decoding is more efficient than AR decoding but shows slightly reduced performance [26]. Our experiments involve evaluating these two methods.

2.4. Overall System Design

The overall system design is streamlined into two steps. First speaker diarization is performed, followed by speech recognition on the diarized segments. Diarization is conducted once per audio file, after which each diarized audio segment is processed in parallel by multiple ASR deployments.

During inference no information about the session device is considered, this is to simulate a more realistic and general-purpose use-case.

Each of our submitted systems use EEND-TA with 6 conformer layers for speaker diarization (EEND-TA C6). For the asr module, system “sys_a” uses Whisper “large-v3”, “sys_b” uses an AR Uconv ASR model and “sys_c” uses a NAR Uconv ASR model.

3. Results and Discussion

3.1. Speaker Diarization

We selected the diarization model EEND-TA C6 for our submission systems by comparing the diarization error rates (DER) and mean speaker counting error rates (MSCE) across datasets outlined in Table 1. Metrics are calculated by including all speech overlaps, no oracle voice activity detection, no oracle speaker counting and without any forgiveness collar. Notably, we achieved new state-of-the-art DER results for the AMI Mix and AMI SDM evaluation sets [13].

Table 2 further breaks down the results for EEND-TA C6 on NOTSOFAR1 by providing an analysis of metrics based on the number of speakers in a recording. This includes DER and its components: missed speech (MS), false alarm (FA), and speaker confusion error (SE), as well as the MSCE. These results show the largest hindrance to diarization model performance are missed speech segments. Also, as the number of active speakers in a recording increases, so does the MSCE. This could be explained by the training dataset containing a greater number of recordings with 4 active speakers. This is because the AMI corpus is comprised mostly of 4 speaker recordings, which could contribute to the lower MSCE results for NOTSOFAR1 SC 4 speaker recordings.

It is important to note that the NOTSOFAR1 train set includes 4 to 8 speaker conversations, whereas the development

Table 1: *Diarization Error Rate (DER) and Mean Speaker Counting Error (MSCE) across several datasets. Lower is better. "NSF1" refers to NOTSOFARI and "All" refers to a combination of all datasets shown. Values marked with * depict new state-of-the-art results.*

Model	Set	Dataset							
		AMI Mix		AMI SDM		NSF1 SC		All	
		DER	MSCE	DER	MSCE	DER	MSCE	DER	MSCE
EEND-TA C4	DEV	12.07	0.00	16.25	0.00	25.35	0.33	18.79	0.26
	EVAL	12.71	0.06	20.71	0.00	33.54	0.41	30.64	0.40
EEND-TA C6	DEV	11.03	0.00	16.16	0.06	20.56	0.30	16.50	0.24
	EVAL	11.83*	0.00	17.55*	0.06	32.47	0.37	29.78	0.36

Table 2: *EEND-TA C6 results for each number of active speakers in the NOTSOFARI SC (NSF1 SC) and NOTSOFARI MC (NSF1 MC) datasets. Included are the Missed Speech (MS), False Alarm (FA), Speaker Error (SE), Diarization Error Rate (DER) and Mean Speaker Counting Error (MSCE) metrics. For all metrics, lower is better.*

Set	NS3					NS4					NS5					NS6					NS7										
	MS	FA	SE	DER	MSCE	MS	FA	SE	DER	MSCE	MS	FA	SE	DER	MSCE	MS	FA	SE	DER	MSCE	MS	FA	SE	DER	MSCE	MS	FA	SE	DER	MSCE	
DEV	-	-	-	-	-	10.94	7.67	4.19	22.80	0.09	8.96	4.49	3.93	17.38	0.38	11.37	6.63	6.41	24.41	0.81	-	-	-	-	-	-	-	-	-	-	-
EVAL	7.55	19.44	5.07	32.06	0.84	6.57	16.94	3.15	26.66	0.10	11.32	15.09	7.89	34.30	0.36	10.41	17.83	8.57	36.81	0.40	14.11	14.89	12.56	41.56	1.06	-	-	-	-	-	

Table 3: *Diarization model sizes and inference throughput with 2 Intel(R) Xeon(R) Gold 6430 vCPUs.*

Model	# Params (M)	RTF
EEND-TA C4	10.2	5.90×10^{-3}
EEND-TA C6	13.3	7.73×10^{-3}

set only includes 4 to 6 speaker conversations. In contrast, the evaluation set consists of 3 to 7 speakers. A noticeable degradation in DER is observed when comparing the evaluation set to the development set, likely due to selecting the top 10 best-performing models based on the development set. Differences in data characteristics are evident from the high false alarm rates reported for the evaluation set, as shown in Table 2.

Details on model sizes and real-time factor (RTF) results are provided in Table 3. The RTF metric was calculated by running diarization on each of the recordings in the NOTSOFARI SC dataset, one recording after another. During this process, we limited the resources available to 2 vCPUs of an Intel(R) Xeon(R) Gold 6430 processor. Under these conditions the EEND-TA C6 model was still approximately 130 times faster than real-time. Given less constraints on hardware resources this takes approximately 7 seconds when using an RTX 4090 GPU, making the model 6260 times faster than real-time.

We chose EEND-TA C6 over EEND-TA C4 due to its enhanced accuracy, even though it comes with a minor increase in latency. The slight latency degradation is insignificant considering the overall efficiency of the EEND-TA architecture. With just 13.3 million parameters, EEND-TA C6 is a lightweight model, making it well-suited for deployment in resource-constrained settings.

3.2. Speech Recognition

Due to the removal of the CSS module, our solution is inherently disadvantaged when transcribing overlapping speech segments. The NOTSOFARI SC development set alone contains an overlapped speech ratio of approximately 20 %. To understand this design drawback, Table 4 presents the tcpWER for each ASR system using oracle diarization segments in replace of our trained models. This illustrates the best result our system can achieve for each ASR model given perfect outputs from di-

Table 4: *Time-Constrained minimum Permutation Word Error Rate (tcpWER) using oracle diarization segments on the NOTSOFARI single-channel DEV set.*

Model	tcpWER
Whisper small	44.23
Whisper large-v3	35.18
AR Uconv	34.85
NAR Uconv	35.12

Table 5: *ASR Model sizes and inference throughput with 10 Intel(R) Xeon(R) Gold 6430 vCPUs.*

Model	# Params (M)	RTF
Whisper small	244	1.36
Whisper large-v3	1550	5.02
AR Uconv	115	0.11
NAR Uconv	83	0.04

arization. We include results using Whisper "small" to show the performance gain when using Uconv models.

With perfect segmentation, the NAR Uconv model displays competitive performance in terms of accuracy when compared to Whisper "large-v3" and shows an absolute improvement of 9.11 % over Whisper "small". The AR Uconv model outperforms Whisper "large-v3" in this experiment. A vastly large proportion of high tcpWER results for all models are attributed to the considerable number of insertion errors made during overlapped speech regions. For example, in an overlapped speech region of three speakers, each speaker's segment transcription will contain the recognition outputs for all active speakers at that time.

As shown in Table 5, the NAR Uconv is nearly 19 times smaller and approximately 125 times faster than Whisper "large-v3". RTF results are calculated by running each of the diarized segments for the recording "MTG_30500/sc_plaza_0/ch0.wav" through a single ASR deployment, one segment after another. This is to simulate how the end-to-end system will run. Here, diarized segments were pre-computed using EEND-TA C6, this resulted in 397.45 seconds of audio for all segments and an average segment length of 2.18 seconds. During this benchmark we constrained the re-

Table 6: Overall system performance on CHiME-8 NOTSO-FAR1 single-channel track.

System	Set	tcpWER	tcorcWER
Baseline	DEV	45.8	38.6
	EVAL	41.4	35.5
sys.a: EEND-TA C6 + Whisper large-v3	DEV	37.4	34.5
	EVAL	44.2	40.7
sys.b: EEND-TA C6 + AR Uconv	DEV	40.6	37.7
	EVAL	44.0	40.7
sys.c: EEND-TA C6 + NAR Uconv	DEV	40.7	37.9
	EVAL	43.1	40.0

sources available to 10 vCPUS of an Intel(R) Xeon(R) Gold 6430 processor.

3.3. Overall System Results

As shown in Table 6, for the development set, our lightweight and efficient “sys.c” achieves improved performance over the baseline in terms of tcpWER on the development set. Our results further outperform the baseline when the ASR module is swapped out for Whisper “large-v3”, shown by the results for “sys.a”.

The reduced gap between speaker-agnostic tcorcWER metric and tcpWER metric in each of our submitted systems for the development set suggested an improvement in terms of speaker labelling accuracy when compared to the baseline. Final results on the development set for the most accurate system, “sys.a”, showed an absolute degradation of 2.19 % tcpWER when compared to using oracle diarization segments.

However, due to the poor performance of the diarization model on the evaluation set, end-to-end tcpWER results for each of our submitted systems perform worse than the baseline. With a more generalized diarization model, less biased towards the development data, this outcome might have been avoided.

Under the constrained computational resources outlined in Tables 3 and 5, “sys.c” processed the entire evaluation set in 50 minutes and 58 seconds, an RTF of approximately 0.05.

Future work will look to address the poor performance in overlapping regions of speech, whilst still keeping the overall system practical and efficient.

4. Conclusion

In this technical report, we presented a simple, lightweight and streamlined approach for single-channel DASR. Reported systems are submitted to the single-channel track of the NOTSO-FAR-1 Challenge. The most practical system submitted required only 2 models to be trained and used a less than 100 million parameters during inference. The EEND-TA speaker diarization module of each system achieved new state-of-the-art results on the publicly available AMI corpus.

5. References

- [1] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, “The fifth chime speech separation and recognition challenge: dataset, task and baselines,” *arXiv preprint arXiv:1803.10609*, 2018.
- [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,” *arXiv preprint arXiv:2004.09249*, 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] A. Vinnikov, A. Ivry, A. Hurvitz, I. Abramovski, S. Koubi, I. Gurchich, 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.
- [5] “NOTSO-FAR-1,” <https://www.chimechallenge.org/current/task2/index>, [Accessed 13-07-2024].
- [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] “CHiME-8 Data,” <https://www.chimechallenge.org/current/task1/data>, [Accessed 13-07-2024].
- [8] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.
- [9] 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.
- [10] 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. of LREC*, 2010.
- [11] W. Kraaij, T. Hain, M. Lincoln, and W. Post, “The ami meeting corpus,” in *Proc. International Conference on Methods and Techniques in Behavioral Research*, 2005.
- [12] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, “End-to-End Neural Speaker Diarization with Permutation-Free Objectives,” in *Proc. Interspeech 2019*, 2019, pp. 4300–4304.
- [13] M. Härkönen, S. J. Broughton, and L. Samarakoon, “Eendm2f: Masked-attention mask transformers for speaker diarization,” *arXiv preprint arXiv:2401.12600*, 2024.
- [14] L. Samarakoon, S. J. Broughton, M. Härkönen, and I. Fung, “Transformer attractors for robust and efficient end-to-end neural diarization,” in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.
- [15] H. Bredin, “pyannote. audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe,” in *24th INTERSPEECH Conference (INTERSPEECH 2023)*. ISCA, 2023, pp. 1983–1987.
- [16] K. Kinoshita, M. Delcroix, and N. Tawara, “Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7198–7202.
- [17] —, “Advances in integration of end-to-end neural and clustering-based diarization for real conversational speech,” *arXiv preprint arXiv:2105.09040*, 2021.
- [18] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, “Conformer: Convolution-augmented transformer for speech recognition,” *arXiv preprint arXiv:2005.08100*, 2020.
- [19] S. J. Broughton and L. Samarakoon, “Improving End-to-End Neural Diarization Using Conversational Summary Representations,” in *Proc. INTERSPEECH 2023*, 2023, pp. 3157–3161.
- [20] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

- [21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [22] "Hugging Face openai/whisper-large-v3," <https://huggingface.co/openai/whisper-large-v3>, [Accessed 14-07-2024].
- [23] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.
- [24] 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," *arXiv preprint arXiv:1804.00015*, 2018.
- [25] A. Andrusenko, R. Nasretdinov, and A. Romanenko, "Uconv-conformer: High reduction of input sequence length for end-to-end speech recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [26] Y. Li, L. Samarakoon, and I. Fung, "Improving non-autoregressive speech recognition with autoregressive pretraining," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.