



THE FOSAFER SYSTEM FOR THE CHiME-8 MMCSG CHALLENGE

Shangkun Huang, Dejun Zhang, Yankai Wang, Jing Deng, Rong Zheng

Beijing Fosafer Information Technology Co., Ltd.

{huangshangkun, zhangdejun, wangyankai, dengjing, zhengrong}@fosafer.com

Abstract

This paper presents the system designed by FOSAFER for the CHiME-8 MMCSG challenge. Our system generates text transcriptions with speaker attributes from natural conversations between two participants in a streaming format. To meet the challenge requirements, we developed a directed automatic speech recognition (ASR) system based on a multi-channel microphone array. The system follows a two-stage training approach and incorporates the SpecAugment dynamic data augmentation technique to improve model performance. Its architecture includes a front-end for speaker label detection and crosstalk suppression using the Non-Linearly Constrained Minimum Variance (NLCMV) beamformer, and a back-end with a streaming hybrid Transducer ASR model that integrates CTC and RNNT decoders. Additionally, the system handles overlapping speech and speaker switching through Sequential Output Training (SOT). Experimental results demonstrate that our system significantly outperforms the official baseline across various delay conditions, underscoring its effectiveness in complex, real-world environments and its potential for practical applications.

Index Terms: multi-channel, multi-talker ASR, smart glasses, CHiME-8 MMCSG

1. Introduction

The CHiME challenge has advanced research and application of speech recognition in real-world environments by gradually introducing complex natural scenes and advanced technologies [1–6]. The CHiME-8 MMCSG task [7] involves obtaining speaker-attributed transcriptions of natural conversations between two participants, recorded with smart Aria glasses, in a streaming fashion using audio, video, and IMU inputs. The challenge includes transcribing both sides of the conversation with minimal latency, while addressing issues like noise, target speaker identification, speech enhancement, diarization, and the impact of a non-static microphone array affected by the wearer’s head movements. Additionally, the task explores how integrating signals from multiple modalities (e.g., cameras, accelerometers, and gyroscopes) can improve transcription performance compared to audio-only systems.

Smart glasses are gaining popularity for speech-related applications such as ASR and enhanced hearing [8]. Advances in audio sensing and augmented reality (AR) have enabled new use cases, but these devices often face challenges from background noise, reverberation, and overlapping speech, which can degrade speech intelligibility [9] [10] [11]. Multi-channel ASR systems, utilizing microphone arrays, offer a solution by improving the signal-to-noise ratio (SNR) and separating relevant speech from noise. This technology makes it possible to transcribe conversations more accurately, even in complex

environments, and can be especially beneficial for applications like real-time captioning for the hearing-impaired or hands-free voice interfaces in noisy conditions. As smart glasses continue to evolve, their ability to handle spatial audio and process speech in real-world scenarios will play a key role in expanding their use cases.

Microphone-array based ASR methods are broadly categorized into end-to-end and hybrid approaches [12–19]. End-to-end methods optimize the multi-channel ASR model using an ASR criterion, with or without explicit separation modules. For example, MIMO-speech [13] uses source-specific time-frequency masks as latent variables for transcription, while later improvements integrate localization sub-networks. Some studies directly incorporate spatial features into the ASR system without separation modules. In contrast, hybrid methods employ a pipeline where speech separation modules explicitly extract target speech or predict speaker-related masks before feeding the processed signal into the ASR system.

In this paper, we designed and implemented a multi-channel directional automatic speech recognition system for the CHiME-8 MMCSG Challenge. The system adopts a two-stage training method and introduces the SpecAugment dynamic data augmentation technique to enhance model performance. The system architecture includes a front-end using an NLCMV beamformer that not only handles speaker label detection but also suppresses crosstalk, as well as a streaming hybrid Transducer ASR model that combines CTC and RNNT decoders, further strengthened by SOT to better handle overlapping speech and speaker transitions. Our experimental results show that, compared to the official baseline, our system demonstrates superior performance across different latencies, achieving notable improvements. These enhancements underscore the practical potential of our proposed solution in complex real-world environments.

2. Multi-channel Directional ASR System Architecture

Fig. 1 illustrates the system architecture of our adopted directed speech recognition system, which integrates NLCMV beamformers, a feature frontend, and a streaming Hybrid Transducer-CTC ASR model trained with SOT.

2.1. Beamforming Front-end

Beamforming is crucial for our system, handling both speaker tag detection and cross-talk suppression. We process raw multi-channel audio using $K + 1$ fixed beamformers: K for horizontal directions around the smart-glasses and one towards the speaker’s mouth, with predetermined coefficients. This shifts the problem from comparing phase differences to evaluating magnitudes and features across multiple directions. we use a new NLCMV criterion that incorporates white noise gain and null direction control, improving performance [20–23]. Specif-

This work is supported by the National Key Research and Development Program of China (No.2022YFF0608504).

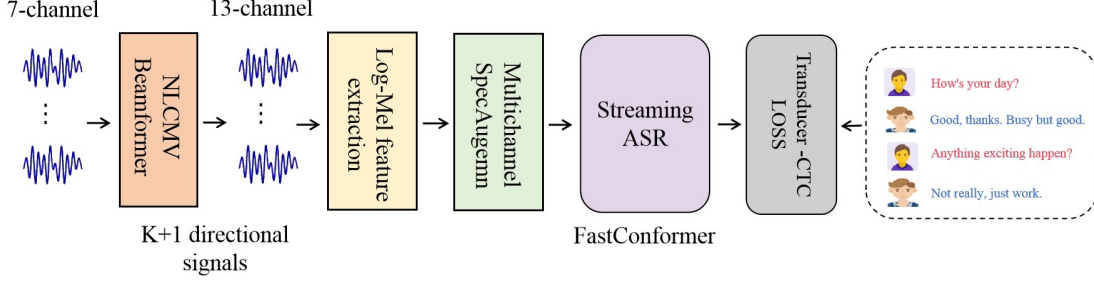


Figure 1: Overview of the Directed Speech Recognition Architecture Used.

ically, the objective function $\mathcal{L}[h(j\omega)]$ of the NLCMV beamformer is as follows:

$$\mathbf{h}^H(j\omega) \left[\Phi_{dd}(j\omega) + \underbrace{\phi_{pp}(w) \sum_{n=1}^N \alpha_{p,n} \cdot \mathbf{g}_n(j\omega) \mathbf{g}_n^H(j\omega)}_{\text{soft control of null directions.}} \right] \mathbf{h}(j\omega) \quad (1)$$

which is subject to the linear equality and nonlinear inequality constraints, which are simplified to the following form:

$$\begin{cases} \mathbf{h}^H(j\omega) \mathbf{g}(j\omega) = 1, \\ c(w) \triangleq \mathbf{h}^H(j\omega) \mathbf{\Psi}(j\omega) \mathbf{h}(j\omega) \leq 0, \end{cases} \quad (2)$$

constraint on white noise gain.

where $\Phi_{dd}(j\omega)$ is the covariance matrix of diffuse noise,

$$\mathbf{\Psi}(j\omega) \triangleq \mathbf{I} - \mathbf{g}(j\omega) \mathbf{g}^H(j\omega) \cdot M / \left[\sum_{m=1}^M |G_m(j\omega)|^2 \right] \quad (3)$$

The $G_m(j\omega)$ are measured channel responses from the target speech source to the m -th of M microphones (ATFs), N is the number of point noise sources, $\phi_{pp}(w)$ is the PSD of point noise, $\alpha_{p,n}$ is the n th point noise weight, and \mathbf{I} is the identity matrix. From the multiple channels produced by the beamformers, we extract per-channel log-Mel features, which are normalized with respect to the corpus mean and variance to enhance convergence. The log-Mel processing removes phase information, which in raw audio encodes directional cues. This is acceptable, as the beamformers have already exploited this directional information, which is now represented as amplitude variations.

We also added a SpecAugment [24] after the feature extraction. SpecAugment is a dynamic data augmentation technique that can be directly applied to spectral speech features for DNN training. The enhancement strategy aims to train more generalized ASR models by predicting data changes in the temporal direction, partial information loss in the frequency direction, and loss of small speech fragments. To achieve this, masks are constructed to dynamically block or modify information in the time and frequency directions. The width and position of the masks are randomly determined, ensuring that the DNN encounters different versions of the input speech in each training period. We enabled time warping with a window size of 5 using bicubic interpolation; frequency masking with a width range of 0 to

27, applied twice; and time masking with a width ratio range of 0% to 5%, applied ten times.

2.2. Cache-Aware Streaming FastConformer ASR with Serialized Output Training

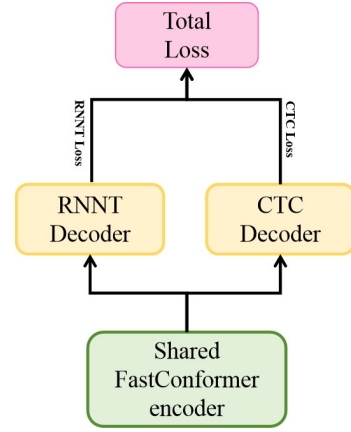


Figure 2: Architecture of the hybrid CTC/RNNT model.

The Cache-Aware Streaming FastConformer model [25] is an optimized architecture for streaming automatic speech recognition, designed to balance efficiency and accuracy. It builds upon the FastConformer [26] by adapting it for streaming applications through two key modifications: constraining the look-ahead and past contexts within the encoder, and introducing an activation caching mechanism. This allows the non-autoregressive encoder to function in an autoregressive manner during inference, thus maintaining consistent performance between training and real-time use.

The model's design includes limiting the context window of each audio frame, ensuring that only a fixed amount of future and past information is considered, which is critical for low-latency streaming. By avoiding normalization in the mel-spectrogram feature extraction and employing fully causal convolution layers, the FastConformer ensures that each input sequence step is processed independently. Layer normalization is used instead of batch normalization, enabling the model to handle streaming data without the need for full-context information.

This architecture supports various decoder configurations, including CTC (Connectionist Temporal Classification) and RNN-Transducer (RNNT) decoders, as well as a hybrid

CTC/RNNT setup, which enhances both the accuracy and computational efficiency. The Cache-Aware Streaming FastConformer has been evaluated on datasets such as LibriSpeech and a large multi-domain dataset, demonstrating superior accuracy, lower latency, and faster inference times compared to conventional buffered streaming models. We used a hybrid architecture which consists of two decoders, one CTC decoder and one RNNT decoder to train our models. Both decoders share a single encoder. The architecture of our hybrid model is shown in Fig. 2. During the training the losses of the CTC decoder (l_{ctc}) and RNNT decoder (l_{rnnt}) are mixed with a weighted summation as the following:

$$l_{total} = \alpha * l_{ctc} + l_{rnnt} \quad (4)$$

where l_{total} is the total loss to get optimized, and α is the hyperparameter to control the balance between these two losses.

Our model also integrates SOT [27] [28], which helps in detecting speaker transitions between the user and another speaker, as well as in recognizing overlapping speech. In our SOT approach, we sort and interleave transcriptions from multiple speakers based on the end times of their words, inserting special markers ($\gg 0$ or $\gg 1$) at each speaker change. This enables the model to learn to annotate ASR transcripts, indicating whether the speech is from the wearer or the other participant.

3. Experiments

3.1. Dataset

Our approach implements a two-phase training method. In the first phase, we generate simulated bifacial conversations using permitted single-channel data, producing approximately 1,000 hours of training data via the MCAC_simulator tool¹. This is combined with 8.5 hours from the MMCSG training dataset, resulting in a total of 1,008.5 hours of data for pre-training over 50 epochs. The model is initialized with streaming Hybrid FastConformer weights.

Given the limited availability of real multichannel data, we simulated 7-channel datasets based on the array configuration of Aria glasses. Project Aria’s microphone array consists of 7-channels as shown in Fig. 3. See [29] for more information about Aria glasses. This was done using the single-channel LibriSpeech [30] and TEDLIUM [31] datasets. Long utterances were segmented into shorter pieces (0.5 to 15 seconds) using single-channel forced alignment. We then used the MCAS dataset to simulate multi-speaker conversations between the wearer (SELF), a conversation partner (OTHER), and an interferer. The MCAS dataset contains information on microphone geometry, real ATFs from Aria glasses, and 10k multichannel room impulse responses (RIRs) from rooms ranging in size from [5, 5, 2] to [10, 10, 6] meters. To simulate conversations, we positioned SELF, OTHER, and the interferer in space, with conversation overlap between SELF and OTHER, and crosstalk from bystanders. The OTHER’s speaking angle ranged from -60° to $+60^\circ$, while bystanders were placed randomly outside this range. Noise from the DNS Challenge was added, with signal-to-noise ratios (SNR) between -5 dB and 30 dB, in 1 dB intervals. The bystanders’ speech did not overlap with the main speakers’. For specific configurations, refer to the MCAC_simulator’s transform.json².

¹https://github.com/facebookresearch/MMCSG/tree/main/tools/MCAC_simulator

²https://github.com/facebookresearch/MMCSG/blob/main/tools/MCAC_simulator/transform.json

In the second phase, the model is fine-tuned using the MMCSG dataset, which includes 8.5 hours of data—consistent with the baseline set. The MMCSG dataset contains 172 training recordings, 169 development recordings, and 189 evaluation recordings, each capturing conversations between two participants, both wearing Aria glasses. The recordings feature 7-channel audio with a 48 kHz sampling rate. SpecAugment is employed to enhance model robustness by masking time and frequency regions randomly during training. This improves generalization, reduces overfitting, and increases data diversity, especially for multi-channel input. SpecAugment is applied throughout both training phases.

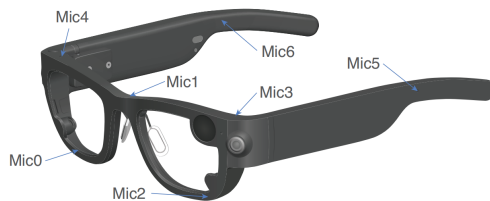


Figure 3: Microphone locations on Project Aria glasses [29].

3.2. Model Setup

The model utilizes the FastConformer-Hybrid-Transducer architecture, input features are extracted using an 80-dimensional log-Mel filterbank from beamformed multichannel audio data. For each audio channel, the encoder network’s input layer projects the concatenated feature vector into 128 dimensions, followed by stacking four consecutive frames, reducing the sequence length by 4x, resulting in a 512-dimensional vector. This vector is then passed through 17 layers of the Conformer encoder with a model dimension of 512. The encoder utilizes depth-wise striding for subsampling, self-attention with relative position encoding, and multi-head attention with 8 heads. Causal convolutions ensure that the model handles streaming input efficiently. Regularization techniques include a 0.1 dropout rate in most layers. The prediction network consists of a single LSTM layer with 640 hidden units, while the joint network projects both encoder and prediction outputs to 640 dimensions before passing them to the RNNT joint net.

For training, the AdamW optimizer is used with the NoamAnnealing scheduler, starting with a learning rate of 0.5, a 5000-step warmup, and a weight decay of 0.001. The model is trained for 50 epochs using a batch size of 8 and mixed precision. Gradient accumulation is set to 2, and the learning rate is reduced automatically based on validation performance. No external language model is used, and fused batch processing optimizes memory efficiency for loss and WER computation. The auxiliary CTC decoder is not employed in this configuration. Additionally, the training process incorporates FastEmit regularization to enhance accuracy and reduce latency during streaming inference.

The model’s parameter configuration largely follows a standard repository³. During the pre-training phase, AdamW is employed as the optimizer with a learning rate of 0.5, momentum parameters (betas) of [0.9, 0.98], and a weight decay of $1e-3$. The scheduler utilizes the NoamAnnealing strategy, with

³https://github.com/facebookresearch/MMCSG/blob/main/config/finetune_asr.yaml

Table 1: Comparison of SELF and OTHER speakers against the official baseline speaker-attributed word error rate (MTWER%) across various training stages and delays. Att. Context Size represents the size of the contextual window used in the attentional mechanism.

System	Att. Context Size	Latency (s)	SELF speakers					OTHER speakers				
			MTWER	INS	DEL	SUB	ATTR	MTWER	INS	DEL	SUB	ATTR
Official baseline	[70, 1]	0.15	17.9	1.7	4.2	10.5	1.6	24.4	2.6	7.3	12.3	2.2
	[70, 6]	0.34	15.0	1.4	3.9	8.4	1.4	21.4	2.2	7.2	10.1	1.8
	[70, 13]	0.62	14.3	1.3	3.8	7.9	1.3	20.3	2.1	7.1	9.6	1.6
	Average		15.7	1.5	4.0	8.9	1.4	22.0	2.3	7.2	10.7	1.9
Pre-training in the first phase	[70, 1]	0.22	10.5	1.2	2.6	6.3	0.5	18.9	2.3	6.7	8.5	1.3
	[70, 6]	0.43	9.9	1.2	2.5	5.7	0.5	18.0	2.4	6.6	7.9	1.2
	[70, 13]	0.71	9.6	1.2	2.4	5.5	0.4	17.6	2.4	6.5	7.6	1.2
	Average		10.0	1.2	2.5	5.8	0.5	18.2	2.4	6.6	8.0	1.2
Fine-tuning in the second phase	[70, 1]	0.19	10.7	1.3	2.5	6.2	0.6	18.2	2.6	5.8	8.6	1.3
	[70, 6]	0.40	9.7	1.2	2.3	5.6	0.5	17.4	2.6	5.6	8.0	1.2
	[70, 13]	0.67	9.6	1.3	2.3	5.6	0.5	17.0	2.6	5.5	7.8	1.1
	Average		10.0	1.3	2.4	5.8	0.5	17.5	2.6	5.6	8.1	1.2

20,000 warmup steps and a minimum learning rate of $1e-6$. For fine-tuning, the learning rate is lowered to 0.1, and the warmup steps are decreased to 5,000, ensuring the model adapts effectively to the fine-tuning phase.

3.3. Evaluation metrics

Evaluation of speech recognition performance by multitalker word error rate (MTWER) used by the system, MTWER evaluates the transcription of SELF and OTHER speakers jointly and it expects the words to be correctly attributed to these two speakers. It breaks down the error into substitutions, insertions, deletions and speaker-attribution errors. The final multitalker WER is then computed for SELF and OTHER as:

$$MTWER_{self} = \frac{INS_{self} + DEL_{self} + SUB_{self} + ATTR_{self}}{NREF_{self}} \quad (5)$$

$$MTWER_{other} = \frac{INS_{other} + DEL_{other} + SUB_{other} + ATTR_{other}}{NREF_{other}} \quad (6)$$

3.4. Result and Analysis

The performance of our system was evaluated through a series of experiments, with results summarized in Table 1. Our method involved two main phases: pre-training and fine-tuning. During pre-training, we observed an average reduction in MTWER of 5.7% for SELF speakers and 3.8% for OTHER speakers compared to the baseline. The most significant improvements were seen for SELF speakers, demonstrating the model’s effectiveness in transcribing the wearer’s speech. For instance, with a context size of [70, 1], the MTWER for SELF speakers dropped from 17.9% in the baseline to 10.5%, while for OTHER speakers, it decreased from 24.4% to 18.9%.

In the fine-tuning phase, using the MMCSG dataset, the model achieved further MTWER reductions, particularly for OTHER speakers, without diminishing the performance for SELF speakers. This indicates that fine-tuning helped the model better adapt to the characteristics of the other speaker’s voice, even though the overall MTWER for OTHER speakers remained higher. Specifically, for a context size of [70, 6],

the MTWER for OTHER speakers improved to 17.4%, down from 18.0% during pre-training and significantly lower than the 21.4% baseline.

We also explored different attentional context sizes—[70, 13], [70, 6], and [70, 1]—which corresponded to average latencies of 674 ms, 401 ms, and 196 ms, respectively, on the DEV dataset. This allowed us to examine the trade-off between latency and recognition accuracy. The results showed that larger context windows generally led to lower MTWER, indicating that a broader contextual view improves transcription quality but at the cost of increased latency.

Despite the improvements, the MTWER for OTHER speakers remained higher than for SELF speakers, highlighting ongoing challenges with far-field speech separation and recognition. The relatively low separation quality for OTHER speakers may have negatively impacted the system’s performance. As noted in previous research, neural beamforming techniques in multi-channel source separation could enhance ASR performance, and future work will focus on improving the system’s ability to handle the speech of OTHER speakers.

4. Conclusions

This paper presents the system developed by Fosafer for the CHiME-8 MMCSG Challenge, which employs a two-stage training approach and incorporates SpecAugment for dynamic data augmentation. The system’s architecture features a front-end that uses the NLCMV beamformer for speaker label detection and crosstalk suppression, and a back-end with a streaming hybrid Transducer ASR model that combines CTC and RNNT decoders. SOT is also integrated to improve the handling of overlapping speech and speaker transitions. Experimental results demonstrate that the system outperforms the official baseline across various latency conditions, particularly in low-latency scenarios. However, speech recognition accuracy remains lower for OTHER speakers compared to SELF speakers, highlighting ongoing challenges in far-field speech separation, which will be addressed in future work.

5. References

- [1] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The pascal chime speech separation and recognition challenge," *Computer Speech & Language*, vol. 27, no. 3, pp. 621–633, 2013.
- [2] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'chime' speech separation and recognition challenge: Datasets, tasks and baselines," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 126–130.
- [3] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime' speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 504–511.
- [4] 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.
- [5] 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.
- [6] S. Cornell, M. Wiesner, S. Watanabe, D. Raj, X. Chang, P. Garcia, M. Maciejewski, Y. Masuyama, Z.-Q. Wang, S. Squartini *et al.*, "The chime-7 dasr challenge: Distant meeting transcription with multiple devices in diverse scenarios," *arXiv preprint arXiv:2306.13734*, 2023.
- [7] "The chime-8 mmsg challenge: Multi-modal conversations in smart glasses," in *CHI ME Workshop on Speech Processing in Everyday Environments*, 2024.
- [8] J. Engel, K. Somasundaram, M. Goesele, A. Sun, A. Gamino, A. Turner, A. Talattof, A. Yuan, B. Souti, B. Meredith *et al.*, "Project aria: A new tool for egocentric multi-modal ai research," *arXiv preprint arXiv:2308.13561*, 2023.
- [9] N. Kern, B. Schiele, H. Junker, P. Lukowicz, and G. Troster, "Wearable sensing to annotate meeting recordings," in *Proceedings. Sixth International Symposium on Wearable Computers*. IEEE, 2002, pp. 186–193.
- [10] T. Feng, A. Nadarajan, C. Vaz, B. Booth, and S. Narayanan, "Tiles audio recorder: an unobtrusive wearable solution to track audio activity," in *Proceedings of the 4th ACM Workshop on Wearable Systems and Applications*, 2018, pp. 33–38.
- [11] A. Dey, M. Billinghurst, R. W. Lindeman, and J. E. Swan, "A systematic review of 10 years of augmented reality usability studies: 2005 to 2014," *Frontiers in Robotics and AI*, vol. 5, p. 37, 2018.
- [12] X. Chang, W. Zhang, Y. Qian, J. Le Roux, and S. Watanabe, "End-to-end multi-speaker speech recognition with transformer," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6134–6138.
- [13] X. Chang, W. Zhang, Y. Qian, J. L. Roux, and S. Watanabe, "Mimo-speech: End-to-end multi-channel multi-speaker speech recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 237–244.
- [14] A. S. Subramanian, C. Weng, S. Watanabe, M. Yu, Y. Xu, S.-X. Zhang, and D. Yu, "Directional asr: A new paradigm for e2e multi-speaker speech recognition with source localization," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 8433–8437.
- [15] J. Heymann, L. Drude, C. Boeddeker, P. Hanebrink, and R. Haeb-Umbach, "Beamnet: End-to-end training of a beamformer-supported multi-channel asr system," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5325–5329.
- [16] S. Huang, Y. Du, Y. Wang, J. Deng, and R. Zheng, "The fosafer system for the icassp2024 in-car multi-channel automatic speech recognition challenge," in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE, 2024, pp. 5–6.
- [17] Y. Shao, S.-X. Zhang, and D. Yu, "Multi-channel multi-speaker asr using 3d spatial feature," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6067–6071.
- [18] Z.-Q. Wang, P. Wang, and D. Wang, "Complex spectral mapping for single-and multi-channel speech enhancement and robust asr," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 28, pp. 1778–1787, 2020.
- [19] 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 *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 558–565.
- [20] T. Feng, J. Lin, Y. Huang, W. He, K. Kalgaonkar, N. Moritz, L. Wan, X. Lei, M. Sun, and F. Seide, "Directional source separation for robust speech recognition on smart glasses," *arXiv preprint arXiv:2309.10993*, 2023.
- [21] J. Lin, N. Moritz, Y. Huang, R. Xie, M. Sun, C. Fuegen, and F. Seide, "Agadir: Towards array-geometry agnostic directional speech recognition," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 951–11 955.
- [22] J. Lin, N. Moritz, R. Xie, K. Kalgaonkar, C. Fuegen, and F. Seide, "Directional speech recognition for speaker disambiguation and cross-talk suppression," in *Proc. INTERSPEECH 2023*, 2023, pp. 3522–3526.
- [23] Y. Yang, D. Raj, J. Lin, N. Moritz, J. Jia, G. Keren, E. Lakomkin, Y. Huang, J. Donley, J. Mahadeokar *et al.*, "M-best-rq: A multi-channel speech foundation model for smart glasses," *arXiv preprint arXiv:2409.11494*, 2024.
- [24] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.
- [25] V. Noroozi, S. Majumdar, A. Kumar, J. Balam, and B. Ginsburg, "Stateful conformer with cache-based inference for streaming automatic speech recognition," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 041–12 045.
- [26] D. Rekish, N. R. Koluguri, S. Kriman, S. Majumdar, V. Noroozi, H. Huang, O. Hrinchuk, K. Puvvada, A. Kumar, J. Balam *et al.*, "Fast conformer with linearly scalable attention for efficient speech recognition," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.
- [27] N. Kanda, J. Wu, Y. Wu, X. Xiao, Z. Meng, X. Wang, Y. Gaur, Z. Chen, J. Li, and T. Yoshioka, "Streaming Speaker-Attributed ASR with Token-Level Speaker Embeddings," in *Proc. Interspeech 2022*, 2022, pp. 521–525.
- [28] X. Chang, N. Moritz, T. Hori, S. Watanabe, and J. Le Roux, "Extended graph temporal classification for multi-speaker end-to-end asr," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7322–7326.
- [29] J. Engel, K. Somasundaram, M. Goesele, A. Sun, A. Gamino, A. Turner, A. Talattof, A. Yuan, B. Souti, B. Meredith *et al.*, "Project aria: A new tool for egocentric multi-modal ai research," *arXiv preprint arXiv:2308.13561*, 2023.
- [30] 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)*. IEEE, 2015, pp. 5206–5210.
- [31] A. Rousseau, P. Deléglise, and Y. Esteve, "Ted-lium: an automatic speech recognition dedicated corpus." in *LREC*, 2012, pp. 125–129.