



# The CHiME-8 MMCSG Challenge: Multi-modal conversations in smart glasses

*Katerina Zmolikova, Simone Merello, Kaustubh Kalgaonkar, Ju Lin, Niko Moritz, Pingchuan Ma, Ming Sun, Honglie Chen, Antoine Saliou, Stavros Petridis, Christian Fuegen, Michael Mandel*

Meta AI

kzmolikova@meta.com

## Abstract

The increasing adoption of smart glasses has opened up the way for innovative applications such as live speech captioning and translation. This presents new exciting research problems and opportunities. To increase the visibility of this research topic and support the researchers in this field, we are introducing the Multi-modal Conversations in Smart Glasses (MMCSG) dataset and challenge. The MMCSG dataset includes two-party conversations, recorded through smart Aria glasses worn by one of the participants, accompanied by manual annotations. Several modalities including multi-channel audio, video and inertial measurement unit (IMU) measurements are available. Additionally, we are releasing the Multi-channel Audio Conversation Simulator (MCAS) dataset and tools. The simulator is designed to generate extensive simulated training data, simplifying development of robust systems. In the challenge, we will evaluate speaker-attributed speech recognition systems on both multi-talker word error rate and algorithmic latency. To assist the challenge participants, we are providing two baseline models. These models serve as starting points for development and as benchmarks for comparison. We hope that these resources will lower the barriers to entry for researchers interested in the potential of smart glasses in enhancing communication.

**Index Terms:** CHiME challenge, speaker-attributed speech recognition, smart glasses, multi-modality.

## 1. Introduction

Smart glasses are growing in popularity, especially for speech and audio use cases like audio playback and communication. Equipped with multiple microphones, cameras, and other sensors, and positioned on the user's head, they offer various advantages over other devices such as phones or static smart speakers. One particularly interesting application is closed captioning of live conversations, which can be particularly beneficial for individuals with hearing impairments and which could also lead to applications such as real-time translation between languages. Integrating such system for smart glasses poses some unique research challenges. Besides speech recognition, the system must address speaker attribution in dynamic and often acoustically challenging environments. It must effectively utilise various information sources provided by the glasses, including multi-channel audio from the microphones, video signals, or other sensor data, such as from accelerometer or gyroscope. Given that all the sensors are head-worn, the system has to deal with rapid movements associated with natural head rotations. Importantly, all processing must be efficient, ideally operating in real-time and directly on the device.

While some research efforts have been already devoted to this problem [1, 2, 3, 4, 5, 6], it still remains relatively under-explored. In the past, research challenges like the CHiME series

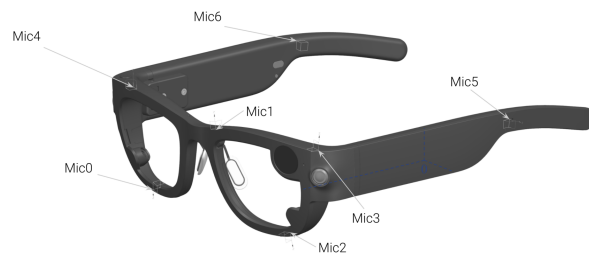


Figure 1: Aria smart glasses.

[7, 8, 9] have driven innovation and often shaped the direction of research. Recognising the need for further research in the area of speech recognition on smart glasses, we have organised a task in CHiME-8 Challenge, named Multi-modal conversations in smart glasses (MMCSG). This challenge aims to not only enhance the visibility of this topic but also to equip researchers with datasets and tools designed for this application, thereby making this research direction more accessible. Through this, we hope to encourage new developments and accelerate advancements in smart glasses technology.

With the launch of the challenge, we have published MMCSG dataset<sup>1</sup>, which includes two-party conversations accompanied by manual annotations. The MMCSG dataset was recorded using Aria smart glasses and features multiple modalities, including multi-channel audio, video and inertial measurement unit (IMU) measurements. In addition to the MMCSG dataset, we have also published room impulse responses and acoustic transfer functions in the MCAS dataset<sup>2</sup>. These can be used together with the MCAS tools that we have released to simulate conversations mimicking the characteristics of real recordings.

To support participants in developing systems for the challenge, we have published two baseline systems. The first baseline system leverages a publicly available pre-trained model, which is then fine-tuned on the MMCSG data. This approach yields competitive performance while not being very computationally demanding. The second baseline involves training a model from scratch on data simulated from Librispeech [10] and TEDLIUM [11]. This baseline serves as a good starting point for participants who prefer not to use a pre-trained model. It also demonstrates how to utilize the MCAS data and tools to generate training data effectively.



Figure 2: Sample snapshots from the MMCSG dataset videos.

## 2. Datasets

We have released two datasets for use in the challenge. The MMCSG dataset serves as the primary data source for the challenge, while the MCAS dataset contains room impulse responses and acoustic transfer functions that can be useful for system development. Both datasets were recorded with Aria smart glasses. In this section, we first provide an overview of the Aria glasses, followed by description of both datasets.

### 2.1. Aria glasses

Aria glasses, shown in Figure 1, are a research device developed as part of Project Aria<sup>3</sup> [12]. They are designed to facilitate research in egocentric machine perception and augmented reality. The glasses are equipped with a variety of sensors including: two monoscene cameras, an RGB camera, eye-tracking cameras, two inertial measurement units (IMUs), seven microphones, a magnetometer, a barometer, a thermometer, a GNSS receiver, and Wi-Fi and Bluetooth transceivers. The configuration of the sensors can be customized through recording profiles, which allow the users to enable or disable individual sensors and adjust frame rate and resolution to accommodate the power and bandwidth constraints of the glasses.

The sensors in the Aria glasses utilized for the MMCSG dataset and challenge are:

- An IMU sensor operating at 1000 Hz positioned at the right temple of the glasses.
- A 7-channel spatial microphone array with a sampling rate of 48 kHz.
- Video footage captured using the RGB camera.

### 2.2. MMCSG dataset

The MMCSG dataset is the cornerstone of the challenge, containing data for training, development and evaluation of the systems. It features recordings between two conversational partners with one participant wearing Aria glasses, which capture

<sup>1</sup><https://ai.meta.com/datasets/mmcs-g-dataset/>

<sup>2</sup><https://ai.meta.com/datasets/mcas-dataset/>

<sup>3</sup><https://www.projectaria.com/>

Table 1: Details of the MMCSG dataset.

subset	number of recordings	total duration	average recording duration	number of speakers
train	172	8.5 h	3 min	49
dev	169	8.4 h	3 min	45
eval	189	9.4 h	3 min	44

Table 2: Positions of microphones at Aria glasses. Directions are given from the point of view of the wearer of the glasses.

position	x [cm] back→front	y [cm] right→left	z [cm] bottom→up
lower-lens right	9.95	-4.76	0.68
nose bridge	10.59	0.74	5.07
lower-lens left	9.95	4.49	0.76
front left	9.28	6.41	5.12
front right	9.93	-5.66	5.22
rear right	-0.42	-8.45	3.35
rear left	-0.48	7.75	3.49

the scene. All data have been manually labeled with transcriptions and speaker activity labels. The detailed statistics about the datasets can be found in Table 1.

The recordings are captured from the ego-centric perspective of one of the participants. Figure 2 displays examples from the dataset. For privacy reasons, all faces in the videos are blurred using the EgoBlur algorithm [13]. While this prevents the application of many audio-visual methods, which leverage lip movement, there is information in the gestures of the speakers that could be used for improving speaker attribution. Note that there is a third person in the room, the moderator, who is present in some of the video recordings, but does not participate in the conversation. The recordings were made indoors, with 30% featuring re-played background noise such as music, traffic, or domestic sounds. On average, 11% of the recordings time includes overlapped speech. The participants in the recordings represent diverse range of ages, genders and ethnicities.

### 2.3. MCAS dataset

In addition to the MMCSG dataset, we are also releasing the MCAS dataset to assist the participants in developing their systems. The MCAS dataset includes:

- Real room impulse responses (RIRs) recorded with the Aria glasses.
- Simulated RIRs generated based on the microphone coordinates of the Aria glasses (Table 2).
- Acoustic transfer functions (ATFs) recorded with the Aria glasses in an anechoic room.

Both real RIRs and ATFs were recorded using a manikin. The real RIRs were captured in five different rooms, with recordings taken from various positions within each room, categorized into four types: far-field, distractor, noise, and mouth, based on the location of the sound source. For mouth RIRs, the source was positioned at the mouth location of the manikin. Far-field RIRs were recorded with the source within an azimuthal range of  $[-60, 60]$  degrees relative to the wearer’s gaze

direction. Distractor sources were positioned in the broader azimuthal range of  $[60, 300]$  degrees. Noise sources are placed randomly throughout the room.

The simulated RIRs follow the same placement of the sources used in the real RIRs. These RIRs are generated using the coordinates of the microphones on the glasses, employing the image-source method through the PyRoomAcoustics toolkit [14], across 10000 simulated rooms. Although these RIRs are less realistic—not considering the head shadows or the directivity of the microphones on glasses—they provide a broader variety compared to the real RIRs.

The ATFs include both far-field and near-field responses, captured in an anechoic room. The near-field ATF source is positioned at the mouth of the manikin, while the far-field ATF sources are distributed across 1674 different locations on a 1.5m sphere surrounding the glasses.

### 3. Challenge and rules

In this challenge, participants are tasked with developing streaming speaker-attributed speech recognition systems using audio, video and IMU modalities. This section outlines the evaluation criteria for the systems and details the rules governing the challenge.

#### 3.1. Evaluation

##### 3.1.1. Multi-talker word error rate

The submitted systems are required to perform speaker-attributed speech recognition, which entails that each transcribed word must be accompanied by a speaker label indicating whether the word was spoken by the wearer of the glasses (SELF) or by the conversation partner (OTHER). To assess both the transcription accuracy and the speaker attribution, we employ multi-talker Word Error Rate (WER).

The multi-talker WER evaluates the transcription of both SELF and OTHER speakers, requiring accurate attribution of words to each speaker. It breaks down errors into substitutions, insertions, deletions and speaker-attribution errors. The final multi-talker WER is calculated for SELF and OTHER as follows:

$$\text{multitalkerWER}_{\text{self}} = \frac{\text{ins}_{\text{self}} + \text{del}_{\text{self}} + \text{sub}_{\text{self}} + \text{attr}_{\text{self}}}{\text{nref}_{\text{self}}} \quad (1)$$

$$\text{multitalkerWER}_{\text{other}} = \frac{\text{ins}_{\text{other}} + \text{del}_{\text{other}} + \text{sub}_{\text{other}} + \text{attr}_{\text{other}}}{\text{nref}_{\text{other}}}, \quad (2)$$

where  $\text{ins}_x$ ,  $\text{del}_x$ ,  $\text{sub}_x$ ,  $\text{attr}_x$  represent the numbers of insertions, deletions, substitutions and attribution errors for speaker  $x$ , respectively.  $\text{nref}_x$  denotes the number of reference words for speaker  $x$ . These error counts are obtained by jointly searching for the best alignment of the reference words of SELF and OTHER with the words in the hypothesis. Note that some words can be aligned to a wrong word of a wrong speaker, thus being both substituted and mis-attributed. We choose to include these types of errors into attributions  $\text{attr}_x$ .

Before calculating the multi-talker WER, both the reference and hypothesis transcription undergo normalization. This process includes removing punctuation, standardizing capitalization and executing a list of permitted substitutions (e.g., converting "okay" to "ok").

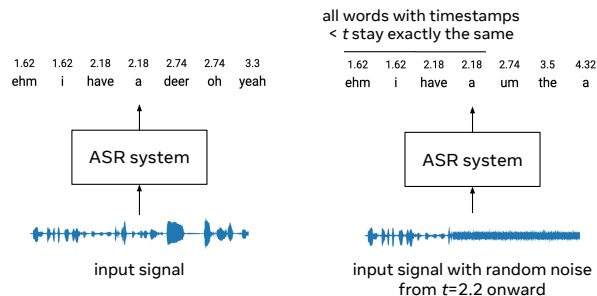


Figure 3: Illustration of the test to verify the accuracy of the word timestamps. The same signal is processed through the system twice; the second time it is perturbed from 2.2 seconds onward. All words with timestamps earlier than 2.2 seconds remain unchanged.

##### 3.1.2. Latency

In this challenge, we emphasize the development of streaming ASR systems for smart glasses, which are required to operate in real-time. Specifically, we assess the algorithmic latency of these systems, measuring how much of the input signal the system uses to emit each of the words. For this challenge, we do not consider computational latency due to the difficulty of fair measurement and comparison. That is, we assume that the computation of the forward pass through the system is instant and we are not asking for actual measurements of the wall-clock time needed to run the system (even though we encourage participants to report it). The submitted systems will be categorized into four groups based on mean latency using the following thresholds: 1000ms, 350ms, 150ms.

To calculate the latency, participants must provide a timestamp for each word, indicating the amount of the input signal that was used to emit that word. The timestamp should account for any look-ahead resulting from the model's architecture, as well as any emission latency that the model learned. The word-timestamp must cover processing of all system components end-to-end and take into account all modalities employed by the system.

To assist the participants in providing accurate timestamps, we created a test script, which can help to uncover cases when the systems violate the streaming assumption. The word timestamp should reflect the portion of the input signal utilized to decode the word, meaning that any alterations to the signal after this timestamp should not affect the decoded word. The test operates by forwarding two similar signals through the system; one signal is perturbed from a certain point in time onward. All words with timestamps preceding this perturbation time should remain unchanged in both forward passes. This is depicted in Figure 3.

#### 3.2. Rules

For the challenge, we have established a set of rules to ensure a fair comparison of the submitted systems. The complete set of the rules is available at the challenge website<sup>4</sup>.

Below is a brief summary:

- Participants are allowed to use the training subset of the MMCSG dataset and a predefined list of other pub-

<sup>4</sup><https://www.chimechallenge.org/current/task3/rules>

Table 3: Results of the baseline systems on the MMCSG development and evaluation sets.

System	Subset	Latency mean [s]	WER [%]	SELF				OTHER				
				ins [%]	del [%]	sub [%]	attr [%]	WER [%]	ins [%]	del [%]	sub [%]	attr [%]
Baseline 1 (from pre-trained model)	dev	0.15	17.9	1.7	4.2	10.5	1.6	24.4	2.6	7.3	12.3	2.2
		0.34	15.0	1.4	3.9	8.4	1.4	21.4	2.2	7.2	10.1	1.8
		0.62	14.3	1.3	3.8	7.9	1.3	20.3	2.1	7.1	9.6	1.6
Baseline 2 (trained from scratch)	dev	0.08	29.1	3.0	6.0	18.4	1.7	37.6	4.2	9.1	20.9	3.4
		0.27	24.9	2.6	5.2	15.8	1.3	33.3	3.6	8.6	18.4	2.7
		0.55	23.5	2.5	5.0	14.8	1.1	31.7	3.4	8.2	17.5	2.5
Baseline 1 (from pre-trained model)	eval	0.14	17.8	1.7	3.9	9.7	2.5	26.3	3.1	7.5	13.2	2.5
		0.33	15.0	1.3	3.4	7.8	2.4	22.9	2.5	7.4	10.8	2.2
		0.62	14.1	1.3	3.4	7.1	2.3	21.7	2.4	7.2	10.0	2.1
Baseline 2 (trained from scratch)	eval	0.06	28.0	3.5	5.2	17.3	1.9	38.9	5.1	8.7	21.5	3.6
		0.27	23.1	2.8	4.4	14.2	1.6	34.4	4.4	8.5	18.7	2.8
		0.54	22.0	2.7	4.3	13.5	1.6	32.8	4.2	8.0	17.9	2.6

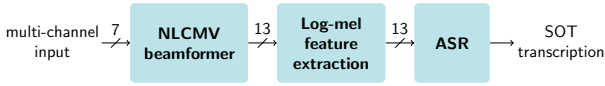


Figure 4: High-level scheme of the baseline system.

licly available datasets, such as AMI, LibriSpeech [10], TEDLIUM [11], among others. This list was specified prior to the challenge, and participants were given the opportunity to propose additional datasets.

- The development set of the MMCSG dataset is available for system evaluation throughout the challenge period.
- A fixed list of pre-trained models is available for use. This list was also specified before the challenge began, with an opportunity for participants to suggest new models.
- Systems must process recordings in sequential order and should not utilize global information. The systems must also provide accurate word time-stamps reflecting this order.
- Each recording must be considered independently during the evaluation process.

## 4. Baseline system

We provide two baseline systems for the challenge:

1. Baseline starting from a publicly available pre-trained model, which is fine-tuned on the in-domain MMCSG dataset.
2. Baseline trained from scratch using simulated data and fine-tuned on the in-domain MMCSG dataset. This baseline showcases the tools to simulate multi-channel multi-talker recordings and can be a good starting point for participants who prefer not to use a pre-trained model.

Both baseline are open-sourced at the challenge Github page<sup>5</sup>. The results for both baselines on development and evaluation set are shown in Table 3.

<sup>5</sup><https://github.com/facebookresearch/MMCSG>

### 4.1. Baseline 1: using pre-trained model

The baseline system follows the framework introduced in [15] and illustrated in Figure 4. The initial component is a fixed Nonlinearly Constrained Minimum Variance (NLCMV) beamformer [1], which employs 13 beams targeting 12 uniformly spaced directions around the wearer, plus one for the mouth of the wearer. The beamformer coefficients are derived using the ATFs recorded in anechoic room with Aria glasses. We have released both the beamforming coefficients and the original ATFs. The beamformer outputs are then used to extract log-Mel features and these features from all 13 beams are concatenated at the input of an ASR model. This model predicts the serialized-output-training (SOT) transcriptions [16].

The ASR model is based on a publicly available pre-trained streaming model, FastConformer Hybrid Transducer-CTC model<sup>6</sup> [17]. It is trained with multiple sizes of attention context, which makes it possible to switch between several latency configurations during test-time. This model is initially designed as a single-speaker, single-channel model. We adapted it by prepending the beamformer, extending the input of the model to multiple channels, extending the tokenizer with speaker tokens for SELF and OTHER, and fine-tuning the model to provide the SOT transcriptions. The fine-tuning is done on the training subset of the MMCSG dataset.

### 4.2. Baseline 2: trained from scratch

The second baseline system employs the same framework and ASR architecture as the first baseline but differs in the training process. It is initially trained from scratch using simulated data and subsequently fine-tuned on the in-domain MMCSG training dataset. For simulating the training data, we have used Librispeech [10] and TEDLIUM [11] for speech, DNS [18] for noise, and real RIRs from the MCAS dataset. We simulated two-speaker conversations with background noise levels ranging from -5 to 80 dB SNR, producing approximately 1 million conversations, each lasting about 17 seconds. Unlike the first baseline, which uses a single-channel single-speaker seed model, this system is trained from scratch on multi-channel simulated data to predict SOT transcriptions. However, the pre-

<sup>6</sup>[https://huggingface.co/nvidia/stt\\_en\\_fastconformer\\_hybrid\\_large\\_streaming\\_multi](https://huggingface.co/nvidia/stt_en_fastconformer_hybrid_large_streaming_multi)

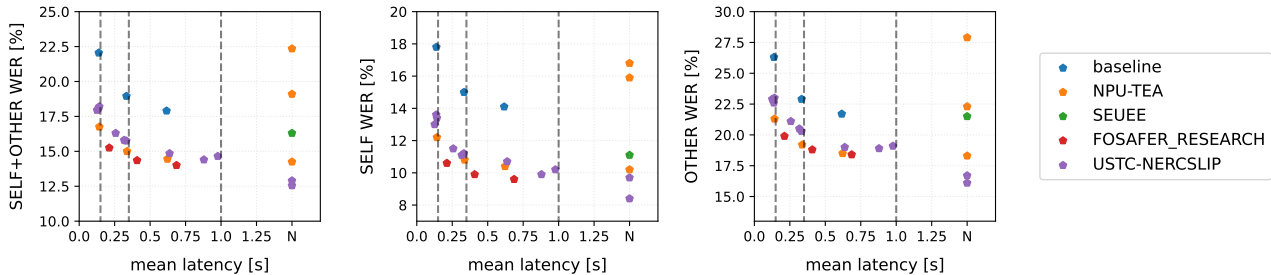


Figure 5: Comparison of overall performance and mean latency across submitted systems.  $\text{mean latency} = N$  indicates non-streaming systems. Dashed vertical lines represent the boundaries between different latency categories.

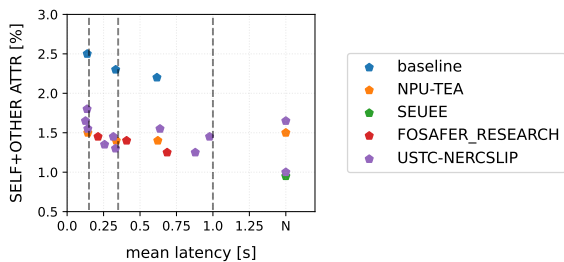


Figure 6: Comparison of speaker attribution error rates and mean latency across submitted systems.  $\text{mean latency} = N$  indicates non-streaming systems. Dashed vertical lines represent the boundaries between different latency categories.

trained model in Baseline 1 used significantly more training data, likely contributing to its better performance. One viable way to improve the performance of Baseline 2 would be to incorporate larger and more diverse speech datasets for simulation, such as VoxCeleb [19] or Gigaspeech [20].

## 5. Challenge results

The challenge obtained submissions from four teams [21, 22, 23, 24], which collectively submitted a total of 21 systems. Of these, 15 systems provided per-word timestamps necessary for latency computation, while six were submitted as non-streaming systems. Many of the submissions adhered to the structure of the baseline system, making improvements upon it. The overall results are summarized in Figure 5. For enhanced clarity in the figure, only Baseline 1 is included, as it achieved better results. Most systems demonstrated significant improvements over the baseline. As anticipated, there was a noticeable correlation between system latency and performance. The break-down of the results into SELF and OTHER highlights the remaining gap between the recognition accuracy between these two types of speakers, confirming the challenges associated with recognizing far-field speech compared to close-talk speech. However, the rating of the systems did not significantly differ between SELF and OTHER.

Figure 6 further zooms in on the attribution error rate. Interestingly, the trends in attribution error rate do not always align with the overall trends observed. A notable outlier is the system developed by SEUEE [22] which achieves the best speaker attribution rate despite its lower ranking in overall WER.

While we refer to the individual system descriptions for detailed information, we highlight several noteworthy points:

- Many submissions demonstrated the importance of training data. The most substantial improvements were achieved by expanding the dataset used to fine-tune the systems with a larger set of simulated data, as shown by the submissions from USTC-NERCSLIP [21], NPU-TEA [23], and FOSAFER\_RESEARCH [24].
- The submissions from NPU-TEA [23] explored a modular approach that includes separate modules for speech separation, diarization, and recognition. This contrasts with the end-to-end baseline system and roughly aligns with the baseline of the NOTSOFAR Task [25] from the CHiME-8 challenge.
- The submission from SEUEE [22] implemented a method focusing on a multi-channel neural-network-based speech separation front-end, which achieved a superior attribution error rate. Furthermore, while this system was submitted into the non-streaming category, a significant portion of its design supports streaming inference.
- The visual modality in the dataset was largely unexplored, possibly due to the blurring of faces in the videos. However, the USTC-NERCSLIP [21] team experimented with the use of IMU sensors, which led to modest improvements.

## 6. Conclusions

In this paper, we have introduced the MMCSG dataset and associated challenge, designed to support advancements of multi-modal speech recognition systems tailored for smart glasses technology. We have provided a dataset that includes multi-channel audio, video, and IMU measurements, supplemented by the MCAS dataset and simulation tools for generating training data. We hope that this challenge can inspire innovative approaches to designing efficient streaming multi-modal speaker-attributed speech recognition systems.

The submitted systems emphasized the critical role of training data in achieving competitive results. Additionally, several interesting analyses and approaches were proposed, including the utilization of IMU modality and leveraging of multi-channel neural-network-based speech separation. Participants explored systems across a spectrum of algorithmic latencies, a factor that is frequently overlooked in other speech recognition challenges.

Looking ahead, there are numerous opportunities for expanding this challenge in future, such as including greater number of speakers, variety of devices, multiple languages, and potentially broadening the scope to include translation tasks.

Finally, we thank the CHiME steering board for their feedback, advice, and oversight in organizing the challenge.

## 7. References

- [1] 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.
- [2] R. Arakawa, M. Parvaix, C. Lai, H. Erdogan, and A. Olwal, "Quantifying the effect of simulator-based data augmentation for speech recognition on augmented reality glasses," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 726–730.
- [3] T. Deppisch, N. Meyer-Kahlen, and S. V. A. Garí, "Blind identification of binaural room impulse responses from smart glasses," 2024. [Online]. Available: <https://arxiv.org/abs/2403.19217>
- [4] F. Ryan, H. Jiang, A. Shukla, J. M. Reh, and V. K. Ithapu, "Egocentric auditory attention localization in conversations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 14 663–14 674.
- [5] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, M. Martin, T. Nagarajan, I. Radosavovic, S. K. Ramakrishnan, F. Ryan, J. Sharma, M. Wray, M. Xu, E. Z. Xu, C. Zhao, S. Bansal, D. Batra, V. Cartillier, S. Crane, T. Do, M. Doulaty, A. Erappalli, C. Feichtenhofer, A. Fragomeni, Q. Fu, A. Gebreselasie, C. González, J. Hillis, X. Huang, Y. Huang, W. Jia, W. Khoo, J. Kolář, S. Kottur, A. Kumar, F. Landini, C. Li, Y. Li, Z. Li, K. Mangalam, R. Modhugu, J. Munro, T. Murrell, T. Nishiyasu, W. Price, P. Ruiz, M. Ramazanov, L. Sari, K. Somasundaram, A. Southerland, Y. Sugano, R. Tao, M. Vo, Y. Wang, X. Wu, T. Yagi, Z. Zhao, Y. Zhu, P. Arbeláez, D. Crandall, D. Damen, G. M. Farinella, C. Fuegen, B. Ghanem, V. K. Ithapu, C. V. Jawahar, H. Joo, K. Kitani, H. Li, R. Newcombe, A. Oliva, H. S. Park, J. M. Reh, Y. Sato, J. Shi, M. Z. Shou, A. Torralba, L. Torresani, M. Yan, and J. Malik, "Ego4d: Around the world in 3,000 hours of egocentric video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 18 995–19 012.
- [6] J. Donley, V. Tourbabin, J.-S. Lee, M. Broyles, H. Jiang, J. Shen, M. Pantic, V. K. Ithapu, and R. Mehra, "Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments," 2021. [Online]. Available: <https://arxiv.org/abs/2107.04174>
- [7] J. P. Barker, R. Marxer, E. Vincent, and S. Watanabe, *The CHiME Challenge: Robust Speech Recognition in Everyday Environments*. Cham: Springer International Publishing, 2017, pp. 327–344. [Online]. Available: [https://doi.org/10.1007/978-3-319-64680-0\\_14](https://doi.org/10.1007/978-3-319-64680-0_14)
- [8] 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.
- [9] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj, D. Snyder, A. S. Subramanian, J. Trmal, B. B. Yair, C. Boeddeker, Z. Ni, Y. Fujita, S. Horiguchi, N. Kanda, T. Yoshioka, and N. Ryant, "CHiME-6 Challenge: Tackling Multispeaker Speech Recognition for Unsegmented Recordings," in *Proc. 6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)*, 2020, pp. 1–7.
- [10] 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.
- [11] F. Hernandez, V. Nguyen, S. Ghannay, N. Tomashenko, and Y. Estève, "Ted-lium 3: Twice as much data and corpus repartition for experiments on speaker adaptation," in *Speech and Computer*, A. Karpov, O. Jokisch, and R. Potapova, Eds. Cham: Springer International Publishing, 2018, pp. 198–208.
- [12] K. Somasundaram, J. Dong, H. Tang, J. Straub, M. Yan, M. Goele, J. J. Engel, R. De Nardi, and R. Newcombe, "Project aria: A new tool for egocentric multi-modal ai research," *arXiv preprint arXiv:2308.13561*, 2023.
- [13] N. Raina, G. Somasundaram, K. Zheng, S. Saarinen, J. Messiner, M. Schwesinger, L. Pesqueira, I. Prasad, E. Miller, P. Gupta *et al.*, "Egoblur: Responsible innovation in aria," *arXiv preprint arXiv:2308.13093*, 2023.
- [14] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 351–355.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] H. Dubey, A. Aazami, V. Gopal, B. Naderi, S. Braun, R. Cutler, H. Gamper, M. Golestaneh, and R. Aichner, "Icassp 2023 deep noise suppression challenge," in *ICASSP*, 2023.
- [19] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620.
- [20] G. Chen, S. Chai, G.-B. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang, M. Jin, S. Khudanpur, S. Watanabe, S. Zhao, W. Zou, X. Li, X. Yao, Y. Wang, Z. You, and Z. Yan, "GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio," in *Proc. Interspeech 2021*, 2021, pp. 3670–3674.
- [21] Y. Jiang, J. Du, Q. Wang, H. Lan, and S. Niu, "The USTC-NERCSLIP systems for the CHiME-8 MMCSG challenge," in *CHiME Workshop on Speech Processing in Everyday Environments*, 2024.
- [22] C. Pang, F. Xiong, Y. Ni, L. Zhou, and J. Feng, "The SEUEE System for the CHiME-8 MMCSG Challenge," in *CHiME Workshop on Speech Processing in Everyday Environments*, 2024.
- [23] K. Huang, W. Rao, Y. Li, H. Wang, Y. Wang, S. Huang, and L. Xie, "The NPU-TEA System Report for the CHiME-8 MMCSG Challenge," in *CHiME Workshop on Speech Processing in Everyday Environments*, 2024.
- [24] S. Huang *et al.*, "The FOSAFER system for the CHiME-8 MMCSG challenge," in *CHiME Workshop on Speech Processing in Everyday Environments*, 2024.
- [25] A. Vinnikov, A. Ivry, A. Hurvitz, I. Abramovski, S. Koubi, I. Gurvich, S. Pe'er, 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 *CHiME Workshop on Speech Processing in Everyday Environments*, 2024.