

CHiME-8 challenge: Task 3 MMCSG

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*

MMCSG Dataset

Multi-Modal Conversations in Smart Glasses

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

Manual annotations

- Speech transcriptions
- Speaker activity



Aria Smart Glasses

Sensors & Modalities

Audio

- 7x spatialized microphones
- 48 kHz

Video

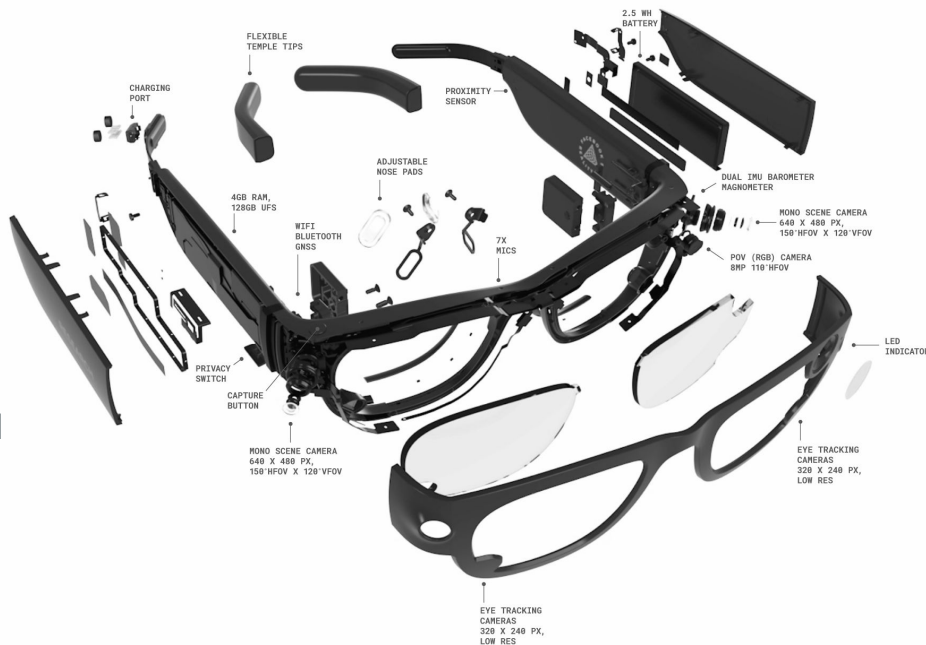
- RGB camera, 720x720, 15fps → faces blurred
- 2x SLAM Cameras, 640x480
- 2x Eye Tracking Cameras, 320x240

IMU

- Accelerometer, 1kHz
- Gyroscope, 1kHz
- Barometer

Other

- GPS, Wifi, Bluetooth



Project Aria

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



Research Challenges

TASK: Speaker-attributed speech recognition

Transcription of
multi-party
conversations

Speaker activity
detection

Non-static
microphone array
processing

Limited
computation

Multimodal
processing

Evaluation

Systems should provide, for each word:

- attribution to SELF/OTHER
- timestamp

1.62	1.62	2.18	2.18	2.74	2.74	3.3	5.53	6.10	6.10	6.65
SELF	SELF	SELF	SELF	SELF	SELF	SELF	OTHER	OTHER	OTHER	OTHER
ehm	i	have	a	deer	oh	yeah	how	was	it	good

Multi-talker word error rate

Reference:

SELF: I had a beer
OTHER: oh yes? how was it?
SELF: good great!

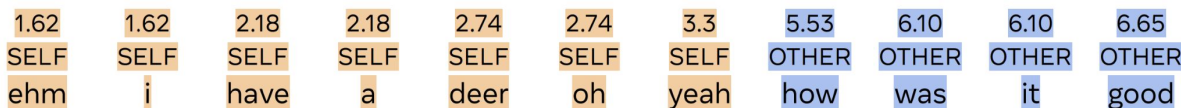
Alignment:

REF	-	i	had	a	beer	oh	yes	how	was	it	good	great
HYP	ehm	i	have	a	deer	oh	yeah	how	was	it	good	-
	INS	CORR	SUB	CORR	SUB	ATTR	SUB ATTR	CORR	CORR	CORR	ATTR	DEL

Evaluation

Systems should provide, for each word:

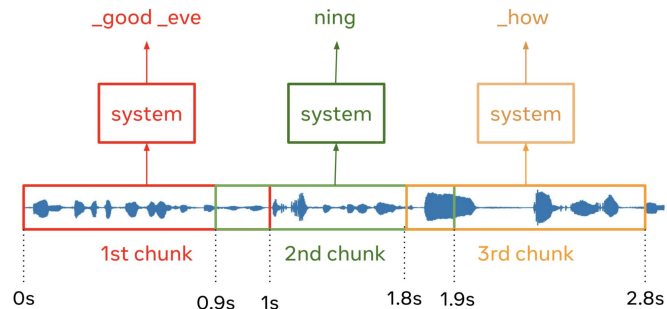
- attribution to SELF/OTHER
- timestamp



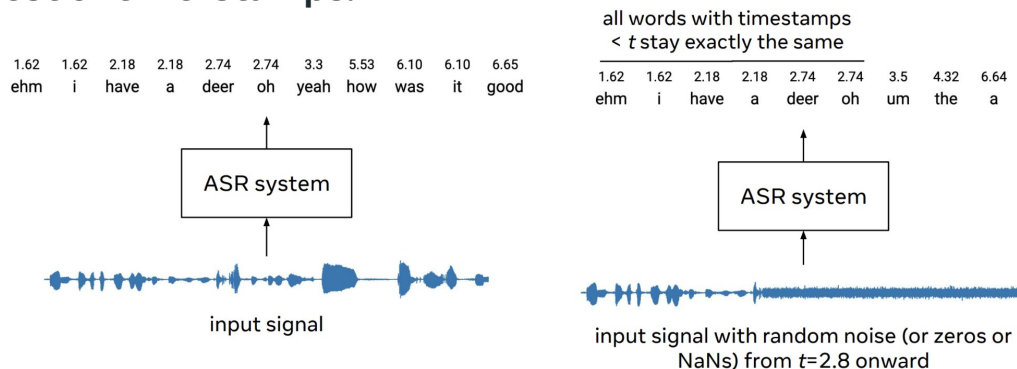
The system had processed 6.10 seconds of the input recording when finished emitting the entire word "was".

4 categories of systems, thresholded by: 1000ms, 350ms, 150ms

Example for chunk-based system:

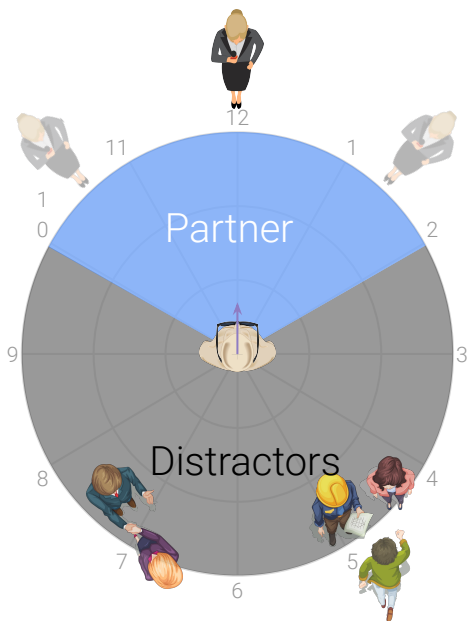


Test of time-stamps:



MCAS Dataset & Tools

Multichannel Audio Conversation Simulator



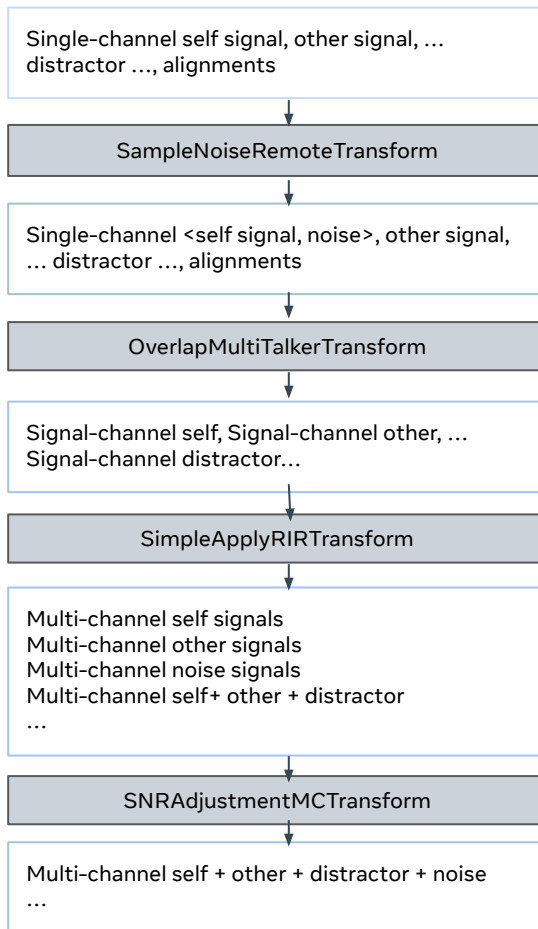
GOAL: Simulate the conversation scenarios

Alignments: Make speech sound naturally in terms of the phone or word at the boundary when cutting into small segments

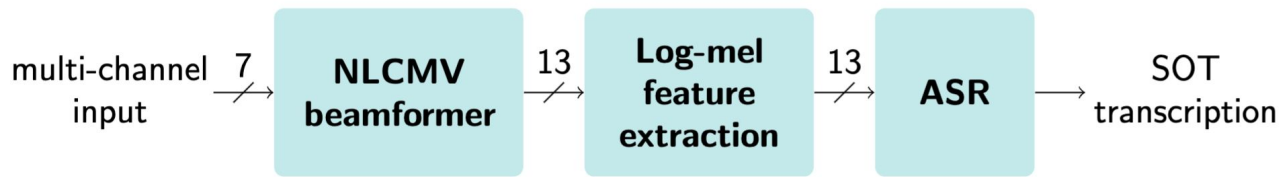
Advantages:

- Can be used both in audio task (output clean and mixed audio) and ASR
- Leverage word alignment to make it sound more natural
- Support both on-the-fly and offline simulation

Processing 14k hours data
offline takes around 4 hours
(using internal Infra).



Baselines



Two baseline systems:

1. Starting from public pre-trained model

- fine-tuned on MMCSG dataset

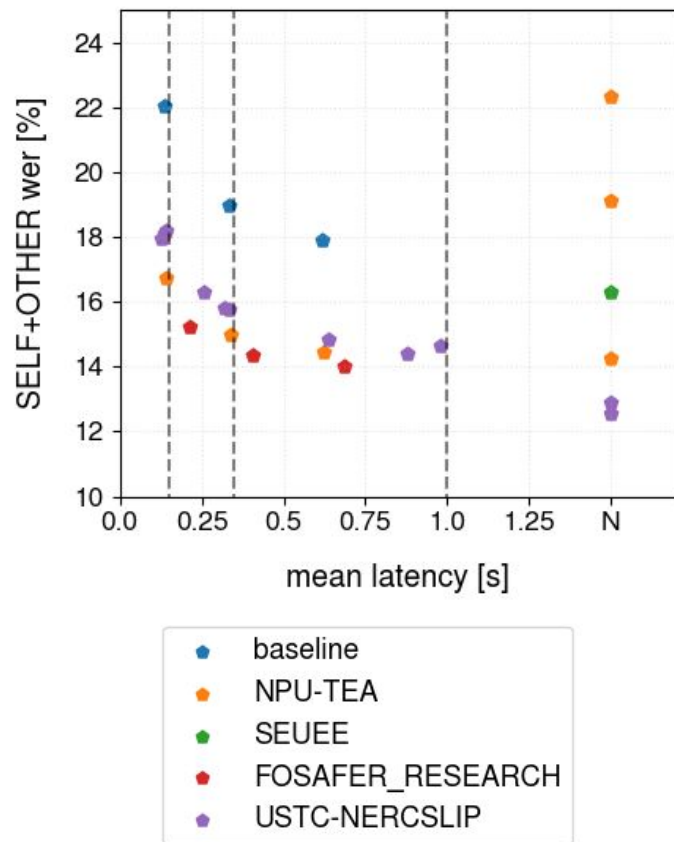
	SELF	OTHER
Latency [s]	WER	WER
mean	[%]	[%]
0.15	17.9	24.4
0.34	15.0	21.4
0.62	14.3	20.3

2. Trained from scratch

- no pre-trained models used
- trained on simulated data with MCAS tool

	SELF	OTHER
Latency [s]	WER	WER
mean	[%]	[%]
0.08	29.1	37.6
0.27	24.9	33.3
0.55	23.5	31.7

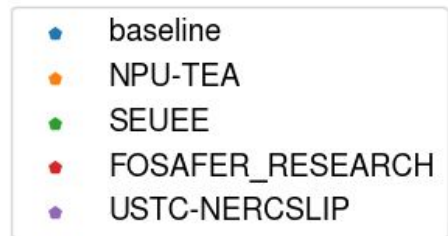
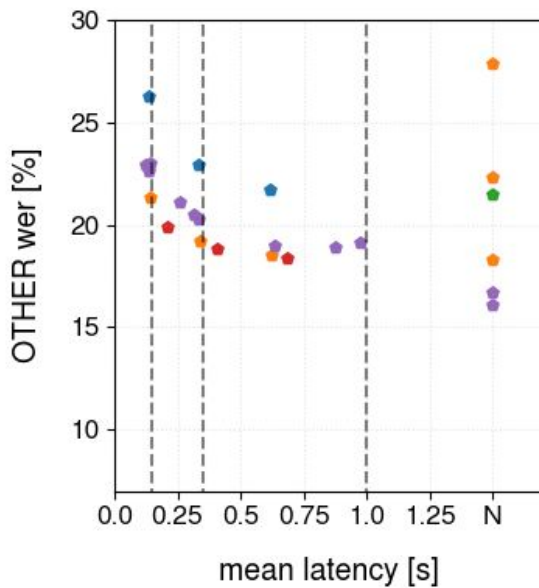
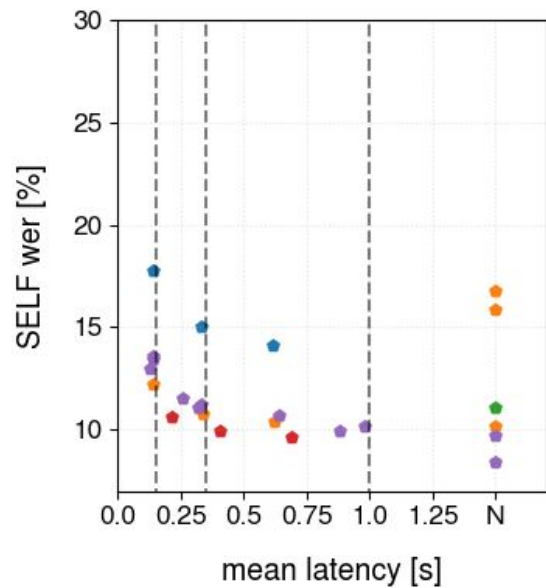
Results: overall



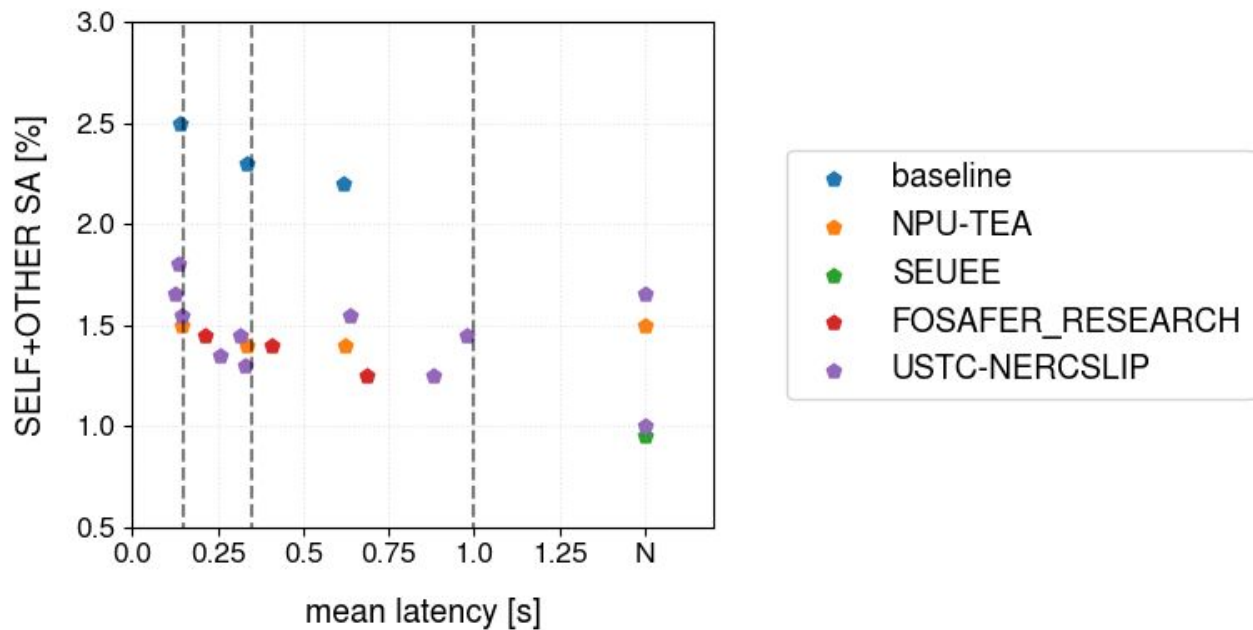
Highlights

- importance of data (*USTC-NERCSLIP*, *NPU-TEA*, *FOSA FER_RESEARCH*)
- modular system following Task-2 baseline (*NPU-TEA*)
- exploration of multi-channel NN speech separation (*SEUEE*)
- exploration of IMU modality (*USTC-NERCSLIP*)
- no use of visual modality

Results: SELF vs OTHER



Results: speaker attribution



Winning system by category

mean latency < 150ms

The NPU-TEA System Report for the CHiME-8 MMCSG Challenge

Kaixun Huang; Wei Rao; Yue Li; Hongji Wang; Yannan Wang; Shen Huang; Lei Xie

150ms < mean latency < 350ms

The NPU-TEA System Report for the CHiME-8 MMCSG Challenge

Kaixun Huang; Wei Rao; Yue Li; Hongji Wang; Yannan Wang; Shen Huang; Lei Xie

350ms < mean latency < 1000ms

THE FOSA FER SYSTEM FOR THE CHiME-8 MMCSG CHALLENGE

Shangkun Huang

1000ms < mean latency (non-streaming)

The USTC-NERCSLIP Systems for the CHiME-8 MMCSG Challenge

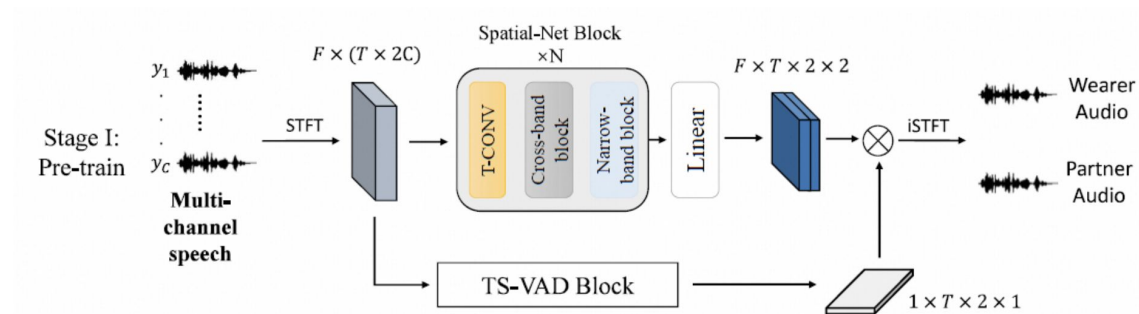
Ya Jiang; Jun Du; Qing Wang; Hongbo Lan; Shutong Niu

Jury award

The SEUEE System for the CHiME-8 MMCSG Challenge

Cong Pang; Feifei Xiong; Ye Ni; Lin Zhou; Jinwei Feng

- for investigation of original speech separation method and its successful application to the egocentric data
- best speaker attribution performance
- efforts towards a fully streaming method



Conclusions

Potential future directions

- more acoustically challenging conditions (noise, distractor speech)
- better support of visual modality
- multilinguality? translation task?

Big thanks to

- all collaborators
- steering committee
- task-1/2 organizers
- **participants!**



[MMCSG dataset](#)



[MCAS dataset](#)



[CHiME-8 Challenge](#)