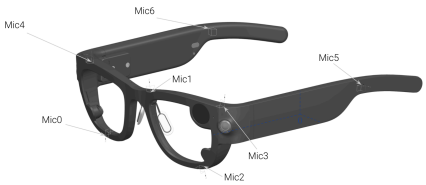
# The CHiME-8 MMCSG Challenge: Multi-modal conversations in smart glasses

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

## MMCSG Dataset

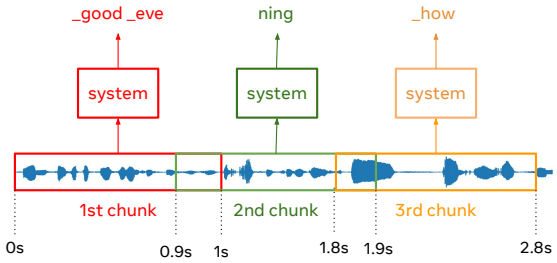
- Two-sided conversations in Aria smart glasses with small amount of noise
- Modalities:
  - 7-channel audio
  - video
  - IMU (accelerometer, gyroscope)
- 8.5 / 8.4 / 9.4 hours for train / dev / eval
- MCAS dataset with Aria RIRs




Aria smart glasses

## Task

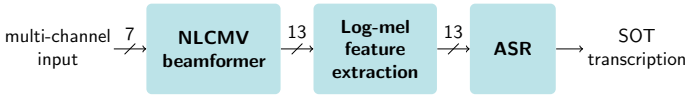
- Streaming ASR system, evaluated using multi-talker WER
- Four categories based on mean per-word latency
  - Thresholds 150ms, 350ms, 1000ms
  - >1000ms includes non-streaming systems
- The outputs of the systems have to include per-word speaker attribution and timestamp




good: 1s, evening: 1.8s, how: 2.8s

## Baseline

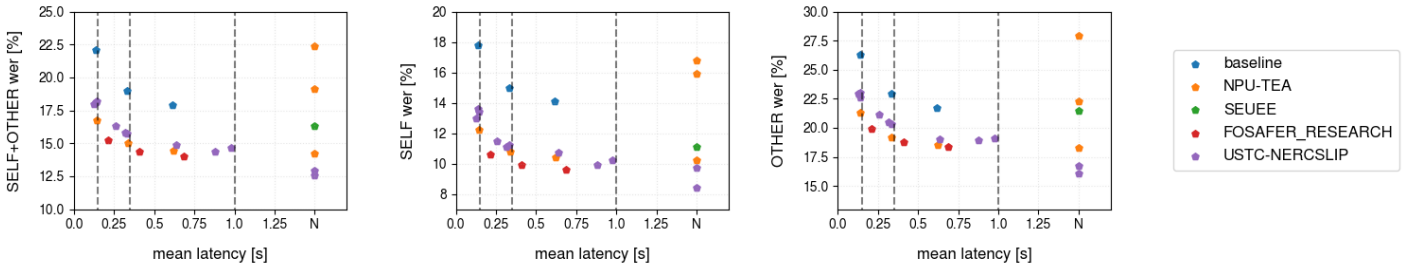
1. Baseline starting from a publicly available pre-trained model fine-tuned on in-domain MMCSG dataset
2. Baseline trained from scratch using simulated data and fine-tuned on the MMCSG dataset

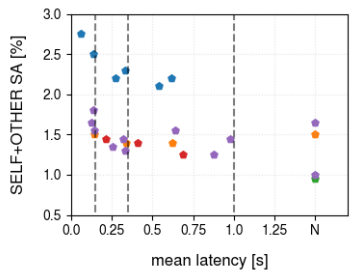



ASR pre-trained model:

- FastConformer RNN-T architecture
- Trained on “NeMo ASRSET” (>10k hours)
- Single-channel, single-speaker
- Configurable attention context in test-time

## Results




**Highlights:**

- Importance of data, especially in-domain (USTC-NERCLIP, FOSAFER)
- Separation based approach achieving the best speaker-attribution error (SEUEE)
- Modular system following NOTSOFAR baseline (NPU-TEA)

MMCSG dataset

MCAS dataset

Challenge website