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

**Ya Jiang[1†], Hongbo Lan[1†], Jun Du[1*], Qing Wang[1], Shutong Niu[1]**

[1]University of Science and Technology of China, China

## Background

**Analysis of Multimodal Conversations in Smart Glasses (MMCSG) Task in CHiME-8 Challenge**

- **Two-person conversation scenarios:** only one person wear the Aria glasses and having **natural conversation**.
- **Streaming speaker-attributed transcriptions system:** both **transcription** and **diarization** need to happen in a **streaming** fashion within the setting latency.
- **Multimodal dataset:** about **8.5 hours multi-channel** audio, **faces blurred** video, **IMU** data (accelerometer and gyroscope recordings)**.**

## The Proposed Approach

### Overall Framework

- **Data simulation:** we used real room impulse response(RIR) data to simulate multi-channel audio to make it more closely approximates the real data.
- **Training strategy**: we proposed a balanced training strategy to prevent real data from being overwhelmed by simulated data and reduce the risk of overfitting.
- **Multi-modal information fusion:** we use high-pass filtering to denoise the IMU modal data and capture the SELF speaker's speech-related information to aid the model's inference.

### Audio-only Architecture

- **Audio data augmentation**

We use real RIR information and adjust overlap ratios and SNR values to make the simulated 7-channel audio data more realistic. We also use the speed perturbation technique to augment the real data.

- **NLCMV beamforming**

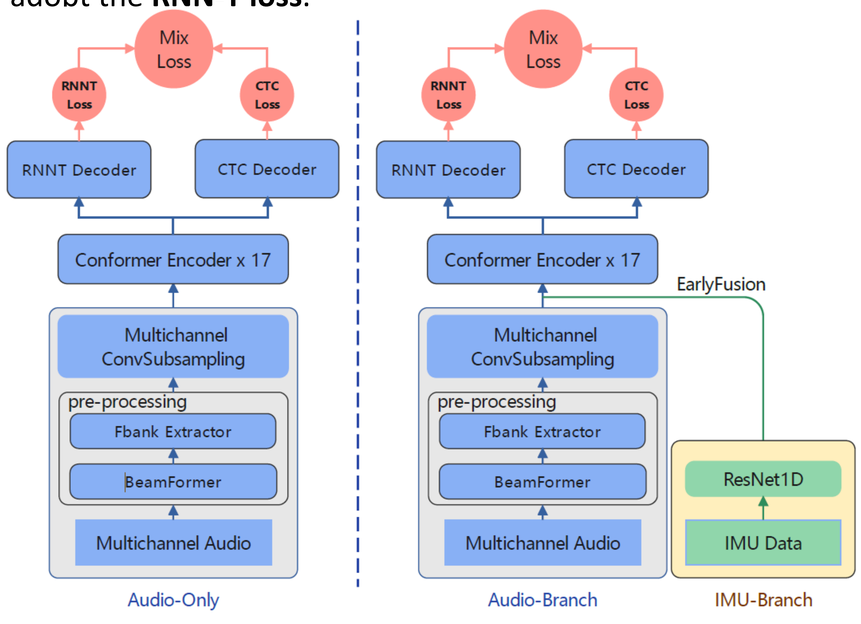
The multiple **NLCMV beamformers** pre-process the raw 7-channel audio to 13-channel beamformed outputs for extracting the **log-Mel features** for each of the 13 beams.

- **Streaming ASR System**

The Fast-Conformer Hybrid Transducer-CTC model processes the multi-channel features and estimates serialized-output-training (SOT) transcriptions.

- **Loss function**

We mainly use the RNN-T decoder to train our models and adopt the **RNN-T loss**.




### Multimodal Architecture

- **IMU preprocess**

Given the IMU unit's sensitivity to different frequencies, we apply **high-pass filtering** to remove components below 20 Hz, which typically contain bias, drift and noise.

- **Feature extraction and fusion**

We use a **1D version of ResNet-18** to encode the IMU data and concatenate the encoded IMU features with CNN-subsampled Fbank features, then feed them into the model.

## Experiments

- **Dataset**

**Development and Test set**: the official development and test set of the MMCSG dataset in CHiME-8 challenge.

- **Experimental setup**

**Evaluation Metric:** multi-talker word error rate (WER) for the latency thresholds: 1000ms, 350ms, 150ms.

**The latencies and WERs on the dev dataset of the final submission systems, one of which won first place in the sub-track of the MMCSG task.**

| System Number | Latency Mean [s] | Attention Context Size | SELF WER [%] | OTHER WER [%] | OVERALL WER [%] |
|---|---|---|---|---|---|
| 1 | 0.130 | [70, 1] | 14.0 | 21.3 | 17.65 |
| 2 | 0.126 | [70, 1] | 13.6 | 21.7 | 17.65 |
| 3 | 0.144 | [70, 1] | 13.7 | 21.5 | 17.60 |
| 4 | 0.254 | [70, 4] | 11.8 | 19.9 | 15.85 |
| 5 | 0.323 | [70, 6] | 11.4 | 19.3 | 15.35 |
| 6 | 0.332 | [70, 6] | 11.4 | 19.2 | 15.30 |
| 7 | 0.645 | [70, 13] | 10.9 | 17.7 | 14.30 |
| 8 | 0.964 | [88, 21] | 10.4 | 18.1 | 14.25 |
| 9 | 0.871 | [84, 20] | 10.3 | 17.7 | 14.00 |
| 10 | - | - | 9.9 | 15.4 | 12.65 |
| 11 | - | - | 8.6 | 15.7 | 12.15 |

**The latencies and WERs on the dev dataset of the multimodal systems, and we are the only team that has investigated the effectiveness of using IMU unit data.**

| Dataset Modality | Latency Mean [s] | SELF WER [%] | OTHER WER [%] | OVERALL WER [%] |
|---|---|---|---|---|
| Audio + Accelerometer | 0.125 | 18.8 | 25.3 | 21.15 |
| | 0.331 | 15.0 | 22.0 | 18.50 |
| | 0.617 | 13.9 | 20.8 | 17.35 |
| Audio + Gyroscope | 0.126 | 18.2 | 25.4 | 21.80 |
| | 0.342 | 14.8 | 21.9 | 18.35 |
| | 0.620 | 13.7 | 20.8 | 17.25 |
| Audio + Accelerometer + Gyroscope | 0.118 | 18.0 | 24.3 | 21.15 |
| | 0.344 | 14.8 | 20.6 | 17.70 |
| | 0.622 | 13.8 | 19.7 | 16.75 |