ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10889243

Multi-modal Streaming ASR in Cross-talk Scenario for Smart Glasses

Ya Jiang†, Hongbo Lan†, Qing Wang†,*, Shutong Niu†
†NERC-SLIP, University of Science and Technology of China (USTC), Hefei, China
{yajiang, lhb1900, niust}@mail.ustc.edu.cn, {jundu, qingwang2}@ustc.edu.cn

***Abstract*—In the MMCSG task of the CHiME-8 Challenge, achieving real-time speaker-attributed transcriptions with limited multi-modal data presents significant challenges. To cope with the problem, we propose a novel ASR framework that leverages both audio-only and multi-modal inputs in a streaming fashion. For the audio-only modality, analyzing and emulating the characteristics of real audio, we utilize a multi-channel simulation to generate the augmented dataset, which efficiently reduces the model training deviation between real and simulated data. Additionally, we integrate the IMU data with audio data in the network structure, demonstrating that the functional filtered and encoded IMU data can assist audio information in achieving better real-time speech recognition performance with ablation experiments. Notably, our explorations based on the above schemes not only secured first place in the MMCSG sub-track but also represented the first investigation into the effectiveness of leveraging IMU data for this task.**

***Index Terms*—streaming ASR, multi-modal data, IMU data, cross-talk, smart glasses, MMCSG, CHiME**

I. Introduction

End-to-end speech recognition technology integrates the acoustic model, language model, and pronunciation lexicon of traditional speech recognition models into a unified system, providing a more efficient solution for automatic speech recognition (ASR) tasks. To address the problem of inconsistency between the lengths of the speech sequence and the output sequence, connectionist temporal classification (CTC) [1], recurrent neural network transducer (RNN-T) [2], and attention-based approaches [3]–[6] are the most prominent approaches in this field. In recent years, models based on the Transformer [7] structure have shown excellent performance in a range of tasks such as natural language understanding, machine translation, and speech recognition [8]–[10].

In various practical application scenarios like cross-talk conversations with people wearing smart glasses, transcribing and presenting the speaker's content in real-time is an intriguing implementation, providing prior information for subsequent tasks such as translation and comprehension. The RNN-T-based ASR system has demonstrated outstanding performance in streaming and online applications and has been successfully deployed in production systems [11], [12]. Nevertheless, attention-based encoder-decoder architectures represent the most effective end-to-end ASR systems, yet their deployment in a streaming context remains challenging, impeding their broader adoption in practice. To address this limitation, alternative streaming ASR approaches based on attention systems have been proposed, including Neural Transducer (NT) [13], Monotonic Chunk Attention (MoChA) [14], and Trigger Attention (TA) [15]. With the advancement and development of multi-modal investigations, research has substantiated that incorporating video input into ASR systems benefits the model's recognition performance [16]–[18]. Recognizing that humans naturally integrate information from multiple modalities to gain a deeper understanding of their surroundings, researchers are increasingly integrating a variety of sensors [19], [20] in devices to collect multi-modal data in addition to video information, aiming to supplement the audio-only modality for better recognition performance.

The CHiME-8 organizers introduced an engaging task, ASR for multi-modal conversations in smart glasses (MMCSG), centering on the cross-talk between two participants recorded with smart glasses, Project Aria [21]. Project Aria, equipped with RGB cameras and non-visual sensors, namely two inertial measurement units (IMUs), microphones, and so on, provides valuable multi-modal information for speech recognition. The goal of the task is to obtain speaker-attributed transcriptions in a streaming fashion with multiple input modalities, where the wearer of the glasses is referred to as SELF and the conversation partner is referred to as OTHER. To tackle this challenge, we explore the practical use of multi-modal information for streaming ASR on smart glasses. Specifically, our key contributions are summarized as follows:

- We propose two approaches for the CHiME-8 MMCSG task: audio-only streaming ASR and multi-modal streaming ASR. These methods offer advanced solutions for real-time transcription of two-person conversations in the smart glasses scenario.
- We adopt a Fast-Conformer-based end-to-end neural transducer as our audio-only ASR system, and perform specific multi-channel simulation and data augmentation strategy, which significantly improves the recognition performance of the streaming model and won first place in the sub-track of the MMCSG task.
- In the multi-modal system, we innovatively integrate IMU data as an auxiliary input to the audio-only system, resulting in effective improvement over the audio-only baseline, where the filtering and encoding of the IMU data play a crucial role in this enhancement.

*corresponding author

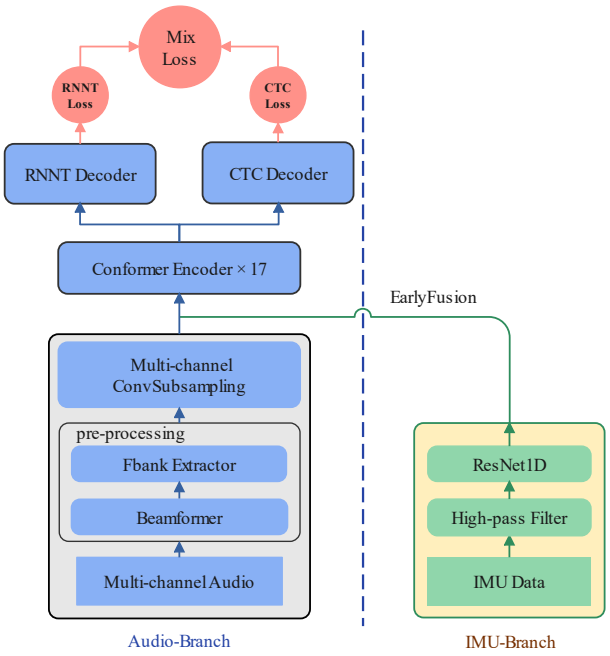

Fig. 1. The proposed audio-only and multi-modal streaming ASR architecture.

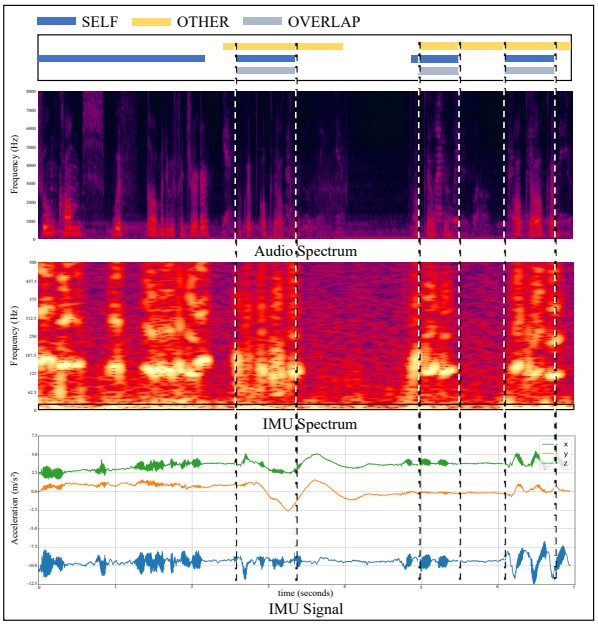

Fig. 2. The speech spectrogram of SELF and OTHER speakers from a randomly selected sample of the training set.

II. Proposed Method

A. Data Augmentation

Given the scarcity of real training audio (only 8.5 hours), we employed a multi-channel audio simulation to generate additional data. We utilize the Librispeech dataset [22] as a clean speech corpus, randomly pairing single-channel audio, designating one as SELF and the other as OTHER. As background noise, we select the Deep Noise Suppression (DNS) Challenge noise dataset [23]. The clean speech and noise were convolved with the provided real room impulse responses (RIRs) to simulate a realistic multi-channel spatial environment. By adjusting the signal-to-noise ratio (SNR), we generate multi-channel audio with both near-field (SELF) and far-field (OTHER) characteristics. Additionally, to mitigate the performance saturation of the model on large datasets, we design various simulated overlap rates tailored to the realistic scenario to reduce the training bias between real and simulated data in experiments.

B. Audio-only ASR Architecture

Our audio-only ASR model, depicted as the left branch of Fig. 1, features a front-end with multiple super-directive beamformers followed by an ASR module. The N = 7 channels of raw audio are processed by the Linearly Constrained Minimum Variance (NLCMV) beamformers [24] with predetermined beamformer weights into K = 12 horizontal steering directions plus one towards the speaker's mouth, resulting in 13-channel beamformed outputs. Mel filter bank (Fbank) features are then extracted and concatenated from each direction. We subsample the Fbank features with a multi-channel convolutional downsampling module and feed them into the ASR module. The pre-processing of the audio branch can be formulated as follows:

$$\mathbf{X}_{bf} = \text{Beamformer}\left(\mathbf{X}_{audio}\right) \tag{1}$$

$$\mathbf{X}_{fbank} = \text{FbankExtractor}\left(\mathbf{X}_{bf}\right) \tag{2}$$

$$\mathbf{F}_{a} = \text{MultichannelCNN}\left(\mathbf{X}_{fbank}\right) \tag{3}$$

where $\mathbf{X}_{audio}$ is the audio sequence, $\mathbf{X}_{bf}$ is the multi-channel beamformed outputs, and $\mathbf{F}_{a}$ is the high-level representations of the audio modality fed into the backend ASR module.

The ASR module is adapted to receive multi-channel audio inputs from single-channel ASR systems [25] and utilize serialized output training (SOT) [26], [27] to detect and separate speech signals from various directions. We use a neural transducer as the end-to-end ASR module, consisting of an encoder, a prediction network, and a joint network. Specifically, a Fast-Conformer network [5] is employed as the encoder, which processes input sequences into acoustic representations and enhances training and inference efficiency using a novel downsampling schema and limited context attention. The prediction network functions as an internal language model or decoder, while the joint network combines outputs from the encoder and prediction network to create the joint representation. In this paper, we employ the RNN-T decoder and its associated losses.

C. Multi-modal ASR Architecture

Although previous studies have demonstrated that IMU data exhibits synergy and complementarity with the sensor wearer's activities, enabled to support tasks like activity recognition [28] and posture detection [29], its integration into ASR systems remains insufficiently explored, particularly in streaming applications. Remarkably, IMU sensors with high

sampling rates can cover the fundamental frequency band of the human voice (85-255 Hz) [30]. As highlighted in the synchronized audio and IMU spectrums of Fig. 2, IMU sensors on smart glasses capture the SELF speaker's speech characteristics, complementing information when the SELF speaker is interfered with. Simultaneously, the IMU data recorded with the accelerometer and gyroscope implicitly reveals the SELF speaker's facial orientation, which facilitates the multi-speaker ASR model in distinguishing between the SELF and OTHER speakers combined with the omnidirectional beamformed audio features, especially in high-overlap segments, thereby reducing speaker substitution errors. Drawing from these analyses, we propose supplementing audio modality with IMU data collected by the smart glasses in the MMCSG task, as described in Fig.1.

According to [31], human physical activities occur mostly below 20 Hz and the sensitive IMU devices are typically prone to noise interference, as confirmed by the low-frequency noise highlighted in the IMU spectrum in Fig. 2. Therefore we first employ a high-pass filter to remove components below 20 Hz, eliminating extraneous noise while preserving essential motion information. Due to the differences in sampling rates and characteristics between IMU and audio modalities, the direct summation or addition fusion methods are inappropriate. We design several IMU pre-encoders, including a causal convolutional module, a stacked unidirectional LSTM, and a 1D version of the standard ResNet-18 architecture referred to [32] to explore their performances. The encoded IMU features are added with the audio representations along the temporal dimension and fed into the conformer module to learn profound interactions and semantic information within modalities, advancing the recognition performance through multi-modal fusion. The other parts of the multi-modal ASR model are the same as the audio-only ASR model and the multi-modal fusion can be formulated as follows:

$$\mathbf{X}_{hf} = \mathrm{HighpassFilter}\left(\mathbf{X}_{imu}\right) \tag{4}$$

$$\mathbf{F}_i = \mathrm{ResNet1D}\left(\mathbf{X}_{hf}\right) \tag{5}$$

$$\mathbf{F}_m = \mathrm{Conformer}\left(\mathrm{Add}\left(\mathbf{F}_a, \mathbf{F}_i\right)\right) \tag{6}$$

where $\mathbf{X}_{imu}$ is the sequence of the IMU signal and high-pass filtering operation in the frequency domain yields de-noised IMU information $\mathbf{X}_{hf}$. $\mathbf{F}_i$ is the encoded IMU features by a pre-encoder (exemplified by a $\mathrm{ResNet1D}$ encoder in Eq. 5) and F_a is the encoded audio features. $\mathbf{F}_m$ is the multi-modal representations output from the conformer module.

D. Streaming ASR Mechanism

Since the model is non-autoregressive in the training stage while streaming ASR is not allowed to access the global information, we adopt various methods to keep consistency in the training and inference process. Firstly, normalization is avoided to process the Fbank features and we replace the convolutional and downsampling layers in the system with corresponding causal modules. The original BatchNorm [33] layers are replaced by LayerNorm [34] layers. Moreover, we use the chunk-aware look-ahead approach [35] to limit the length of the context for the self-attention layers. This approach helps the model reduce unnecessary repetitive computation and accelerates the inference speed. The larger look-ahead contributes to higher accuracy with higher latency. During the inference stage, we utilize the caching mechanism in [35] to convert the non-autoregressive Fast-Conformer encoder into an autoregressive recurrent model by caching intermediate activations computed from the processing of previous time steps. This mechanism avoids duplicated computations and leads to a more efficient streaming inference.

III. Experiments

A. Experimental setups

The MMCSG dataset is a multi-modal dataset of in-room dialog scenes recorded with Project Aria glasses, consisting of 172 recordings for training, 169 recordings for development, and 189 recordings for evaluation. Each recording contains a conversation between two participants wearing Aria glasses with potential background noise. We operate experiments with the following modalities: 7-channel audio recorded at 48 kHz and IMU data (including accelerometer and gyroscope signals) recorded at 1 kHz. For the audio-only system, we apply multi-channel simulation in Section II-A with the Librispeech clean speech corpus and the DNS noise dataset in addition to the real audio dataset, while we only adopt real multi-modal dataset for the multi-modal system. In both the audio-only and multi-modal systems, we utilize a Fast-Conformer pre-trained model [36] to initialize the network. We use only RNN-T loss instead of a hybrid RNN-T/CTC loss, and train all models with Adam [37] optimizer and Noam learning rate scheduler [7] with an initial learning rate of 0.5 using Pytorch. The sub-tracks divide the latency of the submitted systems into four ranges according to different thresholds: 150 ms, 350 ms, and 1000 ms, by calculating the average delay time for each system to correctly recognize words according to the task evaluation rules, and report the average multi-speaker word error rate (WER) of SELF and OTHER speakers on the development set to select the optimal models, where the multi-speaker WER includes errors in recognizing speakers in addition to the standard insertion, deletion, and substitution errors.

B. Results and analysis

*1) **Audio-only Streaming ASR**:* We first applied data augmentation (e.g., speed perturbation) to increase the diversity of real data, depicted as 'Real Data + DA' in Table I. Subsequently, we divided the Librispeech clean speech into three equal parts to generate the simulated audio dataset 'SD1' (comprising split1, split2, and split3), totaling approximately 1200 hours of cross-conversation scenarios. However, when we progressively added simulated data to the real data by splits, the recognition performance of the model initially improved but gradually stagnated, especially for the OTHER speaker the optimization was negligible with only a 1.4 drop of WER at an attention context size of $(70, 13)$ compared to the 'Real Data' baseline.

TABLE I
ABLATION EXPERIMENTS ON AUDIO-ONLY STREAMING ASR MODEL USING MULTI-CHANNEL SIMULATION DATA.

Dataset Duration	Attention Context Size	SELF WER [%]	OTHER WER [%]	OVERALL WER [%]
Real Data	(70, 1)	17.9	24.4	21.15
	(70, 6)	15.0	21.4	18.20
	(70, 13)	14.3	20.3	17.30
Real Data + DA	(70, 1)	18.3	23.3	20.80
	(70, 6)	15.1	20.4	17.75
	(70, 13)	14.1	19.6	16.85
Real Data + DA + SD1	(70, 1)	14.6	22.0	18.30
	(70, 6)	12.4	19.7	16.05
	(70, 13)	11.7	18.9	15.30
Real Data + DA + SD2	(70, 1)	13.1	22.0	17.55
	(70, 6)	11.0	19.5	15.25
	(70, 13)	10.4	18.8	14.60

Recognizing the high overlap rate of about 11% exhibited in conversations in training set, the OTHER speaker's speech (far-field) is often overwhelmed by that of the SELF speaker (near-field), as reflected in the "overlap" segments in Fig. 2. We regenerated split4 and split5 with pluralistic overlap rates from 5% to 30% to cope with the performance degradation. Merging the new sets split4-5 with split1 leads to a multi-overlapping simulation dataset 'SD2'. As indicated in rows 10∼12 of Table I, the model trained with varying overlap rates outperformed that trained with a single overlap rate by an overall 15.6% advancement at (70, 13) attention context size. Incorporating varied overlap rates provides more representative training examples, closely mirroring real-world scenarios. Based on a threshold of up to 1000 ms for a streaming system, we increased the attention context size to (84, 20), contributing to a substantial improvement by 19.1% with only 14.00 for OVERALL WER. In addition, with our proposed optimized multi-overlapping simulation dataset, modifying the causal modules in the model to regular modules for non-streaming inference secures first place in the MMCSG sub-track.

*2) **Multi-modal Streaming ASR**:* As illustrated in Table II, we first investigate the effects of causal CNN, LSTM, and ResNet18-1D IMU pre-encoders, respectively. Among them, ResNet18-1D is adopted since it offers the best performance with a light improvement compared to the audio-only baseline (listed as "real data" in Table I), attributed to the fact that the high amount of low-frequency noise in IMU data hampers the model to extract valuable information. To this end, we applied a high-pass filter to remove the noise from IMU data. Table III shows the ASR results of implementing the high-pass filter with different cut-off frequencies. It is evident that the cut-off frequency of 20 Hz yields an optimal recognition performance, achieving marked enhancements over the audio-only ASR model baseline (listed as "real data") across all attention context sizes. Particularly, the muti-modal ASR model achieves an improvement of the OVERALL WER 9.7% relatively: 12.3% for the SELF speaker and 7.8% for the OTHER speaker at (70, 1) attention context size. It is noteworthy that the improvement in WER for the SELF speaker is generally more pronounced than for the OTHER speaker, which aligns with the IMU primarily capturing speech spectral cues from the SELF speaker. Furthermore, as the cut-off frequency increases, WER performance deteriorates, suggesting that most human activity frequencies are below 20 Hz. Filtering out interference below 20 Hz in the IMU data provides more relevant auxiliary cues, whereas increasing the cut-off frequency further may exclude speech-related signals.

TABLE II
ABLATION EXPERIMENTS ON MULTI-MODAL STREAMING ASR MODEL WITH DIFFERENT IMU PRE-ENCODERS.

IMU Pre-encoder	Attention Context Size	SELF WER [%]	OTHER WER [%]	OVERALL WER [%]
Casual CNN	(70, 1)	22.6	30.1	26.35
	(70, 6)	18.6	25.8	22.20
	(70, 13)	17.5	24.3	20.90
LSTM	(70, 1)	20.8	29.0	24.90
	(70, 6)	17.7	25.2	21.45
	(70, 13)	16.6	23.7	20.15
ResNet18-1D	(70, 1)	17.9	24.3	21.10
	(70, 6)	15.1	21.1	18.10
	(70, 13)	14.1	20.0	17.05

TABLE III
ABLATION EXPERIMENTS ON STREAMING ASR MODEL WITH VARYING CUT-OFF FREQUENCIES USING THE RESNET18-1D PRE-ENCODER.

Cut-off Frequency	Attention Context Size	SELF WER [%]	OTHER WER [%]	OVERALL WER [%]
20 Hz	(70, 1)	15.7	22.5	19.10
	(70, 6)	13.6	19.9	16.75
	(70, 13)	13.0	19.1	16.05
40 Hz	(70, 1)	16.1	22.2	19.15
	(70, 6)	14.0	19.9	16.95
	(70, 13)	13.3	19.0	16.15
60 Hz	(70, 1)	18.1	22.5	20.30
	(70, 6)	15.1	19.5	17.30
	(70, 13)	14.1	18.6	16.35
80 Hz	(70, 1)	18.5	23.9	21.20
	(70, 6)	15.5	20.9	18.20
	(70, 13)	14.6	19.7	17.25

IV. CONCLUSION

In this paper, we introduce a novel multi-modal streaming ASR system on smart glasses, combining audio-only and multi-modal inputs to transcribe the two participants' speech within cross-talk scenarios in real-time. The audio-only streaming ASR model, built upon a Fast-Conformer end-to-end neural transducer architecture, is trained with a large-scale augmented dataset covering varied overlap rates, enabling improved generalization across simulated and real-world data and achieving first place in the MMCSG sub-track. Our multi-modal system, integrating IMU and audio data, effectively captures the synergy and complementarity between modalities through productive noise filtering and encoding. As the only team to explore the integration of IMU data, we showcase its promising potential to enhance real-time multi-modal ASR.

V. ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant No. 62401533.

REFERENCES

[1] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 369–376.

[2] Alex Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[3] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, "Attention-based models for speech recognition," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[4] Zhengkun Tian, Jiangyan Yi, Ye Bai, Jianhua Tao, Shuai Zhang, and Zhengqi Wen, "Synchronous transformers for end-to-end speech recognition," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7884–7888.

[5] Dima Rekesh, Nithin Rao Koluguri, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna Puvvada, Ankur Kumar, Jagadeesh Balam, et al., "Fast conformer with linearly scalable attention for efficient speech recognition," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–8.

[6] Albert Zeyer, Robin Schmitt, Wei Zhou, Ralf Schlüter, and Hermann Ney, "Monotonic segmental attention for automatic speech recognition," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 229–236.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[8] Anthony Gillioz, Jacky Casas, Elena Mugellini, and Omar Abou Khaled, "Overview of the transformer-based models for nlp tasks," in *2020 15th Conference on computer science and information systems (FedCSIS)*, 2020, pp. 179–183.

[9] Jacob Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[10] Linhao Dong, Shuang Xu, and Bo Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5884–5888.

[11] Jinyu Li, Rui Zhao, Hu Hu, and Yifan Gong, "Improving rnn transducer modeling for end-to-end speech recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 114–121.

[12] Jay Mahadeokar, Yuan Shangguan, Duc Le, Gil Keren, Hang Su, Thong Le, Ching-Feng Yeh, Christian Fuegen, and Michael L. Seltzer, "Alignment restricted streaming recurrent neural network transducer," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 52–59.

[13] Tara N Sainath, Chung-Cheng Chiu, Rohit Prabhavalkar, Anjuli Kannan, Yonghui Wu, Patrick Nguyen, and ZhiJeng Chen, "Improving the performance of online neural transducer models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5864–5868.

[14] Chung-Cheng Chiu* and Colin Raffel*, "Monotonic chunkwise attention," in *International Conference on Learning Representations*, 2018.

[15] Niko Moritz, Takaaki Hori, and Jonathan Le Roux, "Triggered attention for end-to-end speech recognition," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5666–5670.

[16] Xueyuan Chen, Yuejiao Wang, Xixin Wu, Disong Wang, Zhiyong Wu, Xunying Liu, and Helen Meng, "Exploiting audio-visual features with pretrained av-hubert for multi-modal dysarthric speech reconstruction," in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12341–12345.

[17] Pingchuan Ma, Stavros Petridis, and Maja Pantic, "End-to-end audio-visual speech recognition with conformers," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7613–7617.

[18] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," in *International Conference on Learning Representations*, 2022.

[19] Zhongjie Ba, Tianhang Zheng, Xinyu Zhang, Zhan Qin, Baochun Li, Xue Liu, and Kui Ren, "Learning-based practical smartphone eavesdropping with built-in accelerometer.," in *NDSS*, 2020, vol. 2020, pp. 1–18.

[20] Running Zhao, Jiangtao Yu, Hang Zhao, and Edith CH Ngai, "Radio2text: Streaming speech recognition using mmwave radio signals," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 3, pp. 1–28, 2023.

[21] Kiran Somasundaram, Jing Dong, Huixuan Tang, Julian Straub, Mingfei Yan, Michael Goesele, Jakob Julian Engel, Renzo De Nardi, and Richard Newcombe, "Project aria: A new tool for egocentric multi-modal ai research," *arXiv preprint arXiv:2308.13561*, 2023.

[22] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev 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.

[23] Harishchandra Dubey, Ashkan Aazami, Vishak Gopal, Babak Naderi, Sebastian Braun, Ross Cutler, Alex Ju, Mehdi Zohourian, Min Tang, Mehrsa Golestaneh, and Robert Aichner, "ICASSP 2023 deep noise suppression challenge," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 725–737, 2024.

[24] Tiantian Feng, Ju Lin, Yiteng Huang, Weipeng He, Kaustubh Kalgaonkar, Niko Moritz, Li Wan, Xin Lei, Ming Sun, and Frank Seide, "Directional source separation for robust speech recognition on smart glasses," *arXiv preprint arXiv:2309.10993*, 2023.

[25] Ju Lin, Niko Moritz, Ruiming Xie, Kaustubh Kalgaonkar, Christian Fuegen, and Frank Seide, "Directional speech recognition for speaker disambiguation and cross-talk suppression," in *INTERSPEECH 2023*, 2023, pp. 3522–3526.

[26] Naoyuki Kanda, Jian Wu, Yu Wu, Xiong Xiao, Zhong Meng, Xiaofei Wang, Yashesh Gaur, Zhuo Chen, Jinyu Li, and Takuya Yoshioka, "Streaming multi-talker asr with token-level serialized output training," in *Interspeech 2022*, 2022, pp. 3774–3778.

[27] Xuankai Chang, Niko Moritz, Takaaki Hori, Shinji Watanabe, and Jonathan Le Roux, "Extended graph temporal classification for multi-speaker end-to-end asr," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7322–7326.

[28] Muhammad Awais, Luca Palmerini, and Lorenzo Chiari, "Physical activity classification using body-worn inertial sensors in a multi-sensor setup," in *2016 IEEE 2nd International Forum on Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI)*, 2016, pp. 1–4.

[29] Hristijan Gjoreski, Aleksandra Rashkovska, Simon Kozina, Mitja Lustrek, and Matjaž Gams, "Telehealth using ecg sensor and accelerometer," in *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2014, pp. 270–274.

[30] Ingo R Titze and Daniel W Martin, "Principles of voice production," 1998.

[31] Ming Gao, Yajie Liu, Yike Chen, Yimin Li, Zhongjie Ba, Xian Xu, and Jinsong Han, "Inertiear: Automatic and device-independent imu-based eavesdropping on smartphones," in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, 2022, pp. 1129–1138.

[32] Sachini Herath, Hang Yan, and Yasutaka Furukawa, "RoNIN: Robust neural inertial navigation in the wild: Benchmark, evaluations, & new methods," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 3146–3152.

[33] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, 2015, pp. 448–456.

[34] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[35] Vahid Noroozi, Somshubra Majumdar, Ankur Kumar, Jagadeesh Balam, and Boris Ginsburg, "Stateful conformer with cache-based inference for streaming automatic speech recognition," in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12041–12045.

[36] NVIDIA-NeMo, "*FastConformer Hybrid Large Streaming Multi (en-US)*," 2023, https://huggingface.co/nvidia/stt_en_fastconformer_hybrid_large_streaming_multi.

[37] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.