Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Gaze-Enhanced Multimodal Turn-Taking Prediction in Triadic Conversations

Seongsil Heo[1,2*], *Calvin Murdock*[2], *Michael Proulx*[2], *Christi Miller*[2]

[1]University of California, Santa Cruz, USA
[2]Meta Reality Labs Research, USA

sheo1@ucsc.edu, {cmurdock,michaelproulx,christim}@meta.com

Abstract

Turn-taking prediction is crucial for seamless interactions. This study introduces a novel, lightweight framework for accurate turn-taking prediction in triadic conversations without relying on computationally intensive methods. Unlike prior approaches that either disregard gaze or treat it as a passive signal, our model integrates gaze with speaker localization, structuring it within a spatial constraint to transform it into a reliable predictive cue. Leveraging egocentric behavioral cues, our experiments demonstrate that incorporating gaze data from a single-user significantly improves prediction performance, while gaze data from multiple-users further enhances it by capturing richer conversational dynamics. This study presents a lightweight and privacy-conscious approach to support adaptive, directional sound control, enhancing speech intelligibility in noisy environments, particularly for hearing assistance in smart glasses.

Index Terms: turn-taking prediction, conversational AI, multimodal interaction, human-computer interaction

1. Introduction

Turn-taking is a fundamental mechanism in triadic conversations, enabling smooth speaker transitions and preventing interruptions. Accurate turn-taking prediction enhances conversational flow by ensuring timely responses and sustained engagement [1, 2, 3].

Non-verbal cues, particularly eye gaze, play a crucial role in regulating turn transitions [4, 5, 6, 7]. As shown in Fig. 1, (a) the future speaker often gazes at the current speaker before the transition, then either maintains or averts their gaze, becoming disengaged from the current speaker after the transition. It indicates complex response planning, or maintaining gaze for a simpler, immediate reply. Meanwhile, (b) the current speaker may direct their gaze toward the future speaker during transitions, signaling an intention to yield the floor. (c) Observers also tend to shift their gaze from the current speaker to the future speaker after the transition, anticipating the next turn.

Existing machine learning models for turn-taking prediction have primarily relied on modalities such as natural language processing (NLP) or image data [8, 9, 10], which require high computational resources and raise privacy concerns. In contrast, gaze-based approaches offer a lightweight, privacy-conscious alternative. Previous studies have explored gaze in turn-taking, but often relied on hand-crafted rules or statistical analyzes [4, 11], or models trained on discretized intervals [12, 13], limiting real-world applicability.

To address these challenges, we propose a gaze-enhanced turn-taking prediction model that integrates minimal contextual cues, including spatial heatmaps representing detected speaker locations and binary voice activity detection (VAD). These models provide a lightweight solution for continuous turn-taking prediction, enabling more dynamic and context-aware applications in triadic conversations.

Our findings demonstrate that (1) egocentric gaze and voice activity are meaningful cues, as voice activity alone lacks sufficient context, (2) multi-user gaze data further improves prediction performance by capturing richer conversational dynamics, (3) minimal complementary inputs, such as spatial heatmaps and voice activity, enable lightweight yet efficient modeling, and (4) evaluating performance from multiple perspectives, including transition-level and group-level analyzes, offers deeper insights into turn-taking dynamics. These models were evaluated on a naturalistic, triadic conversational dataset, highlighting their effectiveness in real-world scenarios.

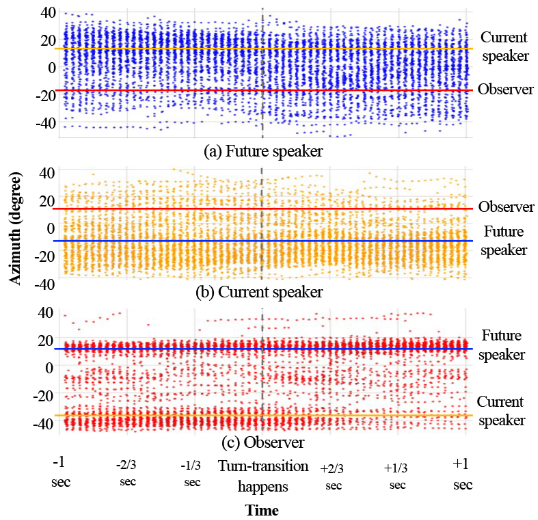

Figure 1: *Gaze behavior during turn transitions in triadic conversations is shown in peri-stimulus time histogram, with gaze azimuth coordinates around the turn-transition moment (dotted vertical lines), accumulating data around these events. This visualization is based on one session from our collected dataset, spanning 1 second before and after the transition. Horizontal bold colored lines indicate other users' positions. For detailed descriptions of (a)-(c), please refer to the main text.*

2. Related Work

2.1. Gaze-driven approaches to turn-taking prediction

Several studies have explored the integration of gaze data for turn-taking prediction. Lee et al. [12] developed a binary

* Work done during an internship at Meta.

10.21437/Interspeech.2025-167

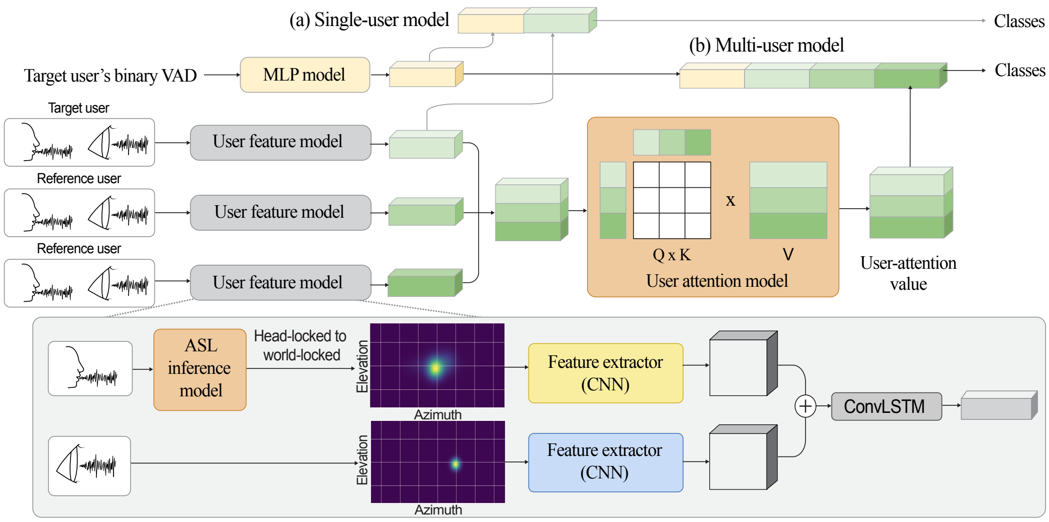

Figure 2: *Architecture of the single- and multi-user models. (a) The single-user model only uses the target user's feature. (b) The multi-user model incorporates features from all users with a user attention mechanism to capture the dynamics multi-user interactions.*

classification model that distinguishes between turn-taking and turn-keeping using gaze target vectors and prior conversational states. However, its dependence on pre-labeled information limits its practicality in real-world applications.

Similarly, Wang et al. [13] proposed a model that extracts features from speech, gaze, and personality data. While this method enriches representation, it increases complexity. Moreover, their evaluation is confined to specific data segments rather than continuous sequences, which restrict its applicability.

2.2. Audio active speaker localization

Jiang et al. [14] introduced an active speaker localization model that utilizes both audio and visual data, generating heatmaps to represent the spatial positions of speakers. It demonstrated robust performance across various channel configurations, including scenarios with audio-only inputs. Although it effectively identifies active speakers, it has not been applied to turn-taking prediction, to the best of our knowledge.

Integrating this model for spatial context can enhance gaze-based turn-taking prediction. By determining whether individuals are looking at the active speaker, the model can provide crucial conversational context, improving turn-taking prediction while maintaining a lightweight framework.

3. Dataset

We used a subset of the Reality Labs Research Conversations for Hearing Augmentation Technology (RLR-CHAT) dataset [15, 16, 17], collected using Aria glasses [18]. The dataset consists of 30 hours triadic conversations engaged in free-form conversations around a round table. Participants aged 20 to 60 years, including those without and with mild hearing loss. Each session lasted approximately 1 hour.

Our study focused on audio, gaze, head movement, and VAD. Audio was recorded using a 7-channel microphone array (48 kHz), while gaze data was recorded with an eye-tracking camera at 30 Hz. Visual-inertial odometry (VIO) was used to gather head orientation measurements at 1 kHz, which also provided input for SLAM-based 3-DoF gaze correction. This correction removed head-motion effects to ensure that gaze direction accurately reflected participants' intended focus points. Binary VAD was derived by isolating own-voice audio through nearfield beamforming, followed by voice classification to generate a binary indicator of speaking activity.

The target user is defined as the participant wearing the glasses for whom the turn-taking prediction is performed, while reference users are all other participants. To ensure a precise analysis of turn-taking behaviors, it is essential to synchronize users' data streams with minimal error. Even slight discrepancies can alter turn-taking labels, affecting dataset reliability. To our knowledge, no existing gaze datasets in multi-party conversational settings, other than ours, achieves this level of synchronization accuracy.

Turn-taking was labeled using a role-based and behavior-based framework [13]. Labels were initially generated using a binary VAD time-based algorithm, with 4 out of the 10 sessions manually re-labeled to enhance labeling accuracy. To assign labels, we analyzed Inter-Pausal Units (IPUs) extracted from binary VAD data. An IPU is defined as a continuous segment of speech separated by pauses. We applied 0.5 second smoothing to the binary VAD signal [13, 19, 20], merging segments where speech resumed within this threshold.

In the role-based classification, each participant was assigned one of three roles: main speaker (if multiple speakers overlapped, the first to start speaking was assigned this role), non-main speaker (spoke during an IPU but was not the main speaker), and observer (did not speak during an IPU).

For behavior-based classification, we considered the target user's binary VAD status and the presence of a main speaker. Silence was assigned when the VAD indicates silence. Turn-taking occurred when the VAD indicates speaking, either (1) no main speaker immediately before the target user began speaking, and the previous speaker is not target user, or (2) main speaker exists immediately before the target user began speaking, and the target user stopped speaking before the main speaker finished. Turn-keeping was assigned when the VAD indicates speaking, with no prior main speaker immediately before the target user began speaking, and the previous speaker was the target user. Back-channeling occurred when The VAD indicates speaking, with a prior main speaker immediately before the target user began speaking, but the main speaker con-

Table 1: *F1 scores for role-based classification.* ***Bold*** *indicates the best, while* <u>*underline*</u> *represents the second-best.*

Model	All	Main	Non-main	Observer
Target VAD	0.704	0.761	0.351	**1.000**
Single-user	<u>0.746</u>	<u>0.835</u>	<u>0.405</u>	**1.000**
Multi-user	**0.765**	**0.869**	**0.428**	**1.000**

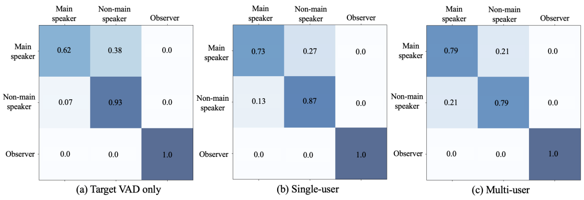

Figure 3: *Confusion matrices of Table 1.*

tinued speaking even after the target user started speaking.

4. Methodology

We propose two models for turn-taking prediction: a single-user model that focuses on the target user's gaze and audio data, and a multi-user model that incorporated gaze features from all participants. An overview of these models is presented in Fig. 2.

4.1. Single-user model

As illustrated in Fig. 2 (a), the single-user model processes multimodal inputs from the target user. To ensure comprehensive turn-taking analysis, each participant is assigned as the target user in turn, allowing the model to learn from multiple perspectives.

Binary VAD features are extracted using a 5-second sliding window and passed through a Multi-layer Perceptron (MLP) for feature extraction. Gaze data, represented as azimuth and elevation angles, are accumulated over a 0.2-second window (6 frames) to construct gaze heatmaps. Audio data, segmented into 0.2-second windows, is processed through an Active Speaker Localization (ASL) inference model, generating a spatial heatmap of the active speaker's location. Both heatmaps are then fed into a CNN-based feature extractor respectively.

Extracted gaze and audio features are combined to enhance the representation of spatiotemporal dynamics and forwarded to a Convolutional Long Short-Term Memory (ConvLSTM) [21]. This produces a comprehensive user feature vector,concatenated with the binary VAD feature vector, and fed into a classifier to predict turn-taking behavior.

4.2. Multi-user model

The multi-user model extends the single-user approach by incorporating gaze and audio features from all participants while keeping the binary VAD specific to the target user. As shown in Fig. 2 (b), gaze and audio data from each user are processed independently through respective feature extractors, generating individual user feature vectors. These vectors pass through a user attention model, computing attention scores to quantify user interactions [22, 23]. The model outputs feature vectors for each user, concatenated with the target user's binary VAD feature. The final representations go through classification layers. To prevent positional bias, training includes both original and reversed reference user orders.

Table 2: *F1 scores for behavior-based classification.* ***Bold*** *marks the best, while* <u>*underline*</u> *the second-best. (a) Original evaluation method. (a.1) Transition-level analysis: F1 scores computed within the 1-second window following a turn change. (a.2) Group-level analysis: evaluates interaction pattern by analyzing sequences of consecutive data points.*

Model	All	Turn-taking	Turn-keeping	Back-channel	Silence
(a) Behavior-based					
Target VAD	0.661	0.490	0.730	0.424	**1.000**
Single-user	<u>0.670</u>	<u>0.518</u>	<u>0.735</u>	<u>0.429</u>	**1.000**
Multi-user	**0.682**	**0.534**	**0.741**	**0.454**	**1.000**
(a.1) Transition-level					
Target VAD	0.559	0.081	<u>0.711</u>	0.443	**1.000**
Single-user	<u>0.595</u>	<u>0.218</u>	<u>0.711</u>	<u>0.452</u>	**1.000**
Multi-user	**0.653**	**0.400**	**0.733**	**0.481**	**1.000**
(a.2) Group-level					
Target VAD	0.705	0.414	0.701	0.703	**1.000**
Single-user	<u>0.715</u>	<u>0.437</u>	<u>0.713</u>	<u>0.711</u>	**1.000**
Multi-user	**0.732**	**0.494**	**0.720**	**0.715**	**1.000**

5. Experiments and Results

5.1. Training setup and evaluation metrics

The user feature model consists of a 2D CNN block (kernel size: 3, 32 dimensions) followed by a ConvLSTM layer (kernel size: 3, 64 dimensions). The binary VAD model employs a two-layer MLP with 128 and 32 units. We used the Adam optimizer (learning rate: 1e-2) and trained the model using a pre-training and fine-tuning strategy: pre-trained on 6 algorithmically labeled sessions, then fine-tuned on 4 manually labeled sessions. Each manually labeled session was split into first 60% for training, the middle 20% for validation, and the final 20% for testing.

As a baseline, we trained a model using only the target user's binary VAD features, comparing its performance with both the single-user and multi-user models to assess the impact of gaze information on turn-taking prediction. To address class imbalance, we employ a weighted cross-entropy loss function [24]. Given the challenges posed by class imbalances, we used the F1 score as the primary metric, applying the macro-averaging method to compute an overall F1 score by averaging per-class scores equally [25].

5.2. Role-based results

Table 1 shows the F1 scores for role-based classification, comparing the target VAD baseline, single-user model, and multi-user model. Incorporating single-user gaze data improves classification for both main and non-main speakers, while multi-user gaze data further enhances performance. Since the observer role is directly determined by binary VAD indicating silence, the F1 score remains 100% across all. A gaze-only model without VAD was excluded as it failed to converge, indicating gaze alone is insufficient to distinguish silence periods.

Figure 3 shows confusion matrices. The target VAD-only model often misclassifies main speakers as non-main, while the single-user gaze model reduces errors, and the multi-user model further refines classification for more balanced performance.

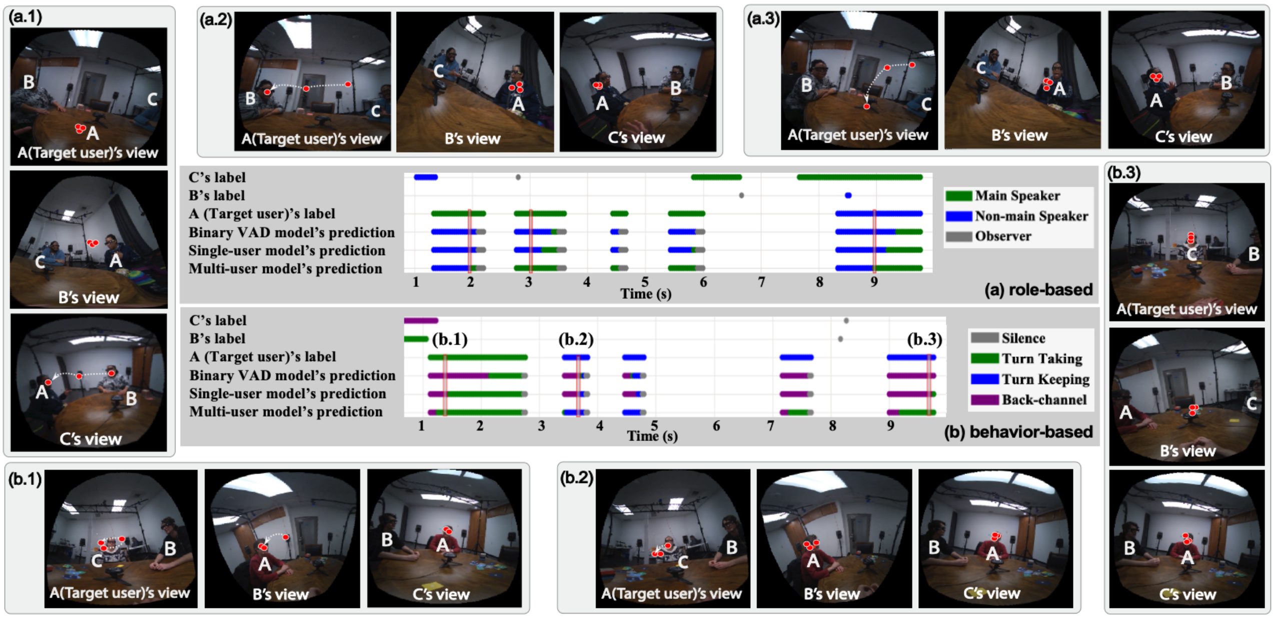

Figure 4: *A 10-second video segment illustrating turn-taking behavior for target user A and reference users B and C. Panels (a) and (b) present role-based and behavior-based classifications, respectively, with color-coded turn-taking labels when voice is detected. The first 3 rows display ground truth labels for A, B and C, while the next 3 rows display predictions from the binary VAD, single-user model, and multi-user model for target user A. Gaze direction (red dots) and gaze shifts (white arrows) provide key turn-taking cues. For detailed descriptions of (a.1)–(b.3), please refer to the main text.*

5.3. Behavior-based results

Table 2 summarized the F1 scores for behavior-based classification. Incorporating single-user gaze data improves performance across turn-taking, turn-keeping, and back-channel, with multi-user model further enhancing results.

To provide a detailed evaluation, we analyze transition-level and group-level performance (Table 2). Transition-level analysis assesses predictions within 1s after a turn change, ensuring timely responses [26, 13]. Group-level analysis aggregates consecutive data points into single events, capturing broader interaction patterns [27, 28]. The predicted label for each event is determined by the most frequent class. Both approaches show improvements with single-user gaze data and even greater gains with multi-user gaze data.

5.4. Analysis of predictions

To further examine the model's behavior, we analyze specific cases of turn transitions, comparing the binary VAD, single-user, and multi-user gaze models (Fig. 4). The following sections highlight key moments from both role-based (panel a) and behavior-based (panel b) classifications, emphasizing the impact of gaze dynamics on predictions.

5.4.1. Role-based

- **(a.1):** Only the multi-user model correctly predicts target user A as the main speaker, likely because both reference users, B and C, direct their gaze toward A. B maintains eye contact with A, while C shifts attention from B to A. However, the correct prediction occurs slightly later, as C's gaze shift is delayed. This joint attention serves as a strong cue, which the binary VAD and single-user models fail to capture.
- **(a.2):** The single-user model predicts A as the main speaker earlier than the binary VAD model, while the multi-user model identifies it even sooner. A's gaze aversion, and the B/C's sustained gaze on A enhance the prediction.
- **(a.3):** The multi-user model appears to underperform, but the issue stems from mislabeling. Around 9 seconds, C starts laughing, unintentionally shifting the conversational turn from C to A. Despite the incorrect label, the multi-user model correctly predicts the transition, capturing nuances that were overlooked in the manual labeling process.

5.4.2. Behavior-based

- **(b.1):** The single-user model detects the transition earlier than the VAD model, while the multi-user model identifies it even earlier. A's gaze aversion, B's shift to A, and C maintaining gaze on A enhances turn-taking detection.
- **(b.2):** When the ground truth label is turn-keeping, only the multi-user model predicts correctly by leveraging B/C's sustained gaze on A, an indicator of continued engagement.
- **(b.3):** All models fail to predict turn-keeping, as B disengages by looking away from the speaker, highlighting gaze data's limitation and the need for additional context.

6. Conclusion

We introduced a lightweight framework for turn-taking prediction, demonstrating that gaze, despite its inherent complexity, is a reliable predictive cue when structured with spatial constraints. Unlike prior work that overlooked or passively used gaze, our method integrates speaker localization and speech activity to extract meaningful predictive signals without computationally intensive models.

Beyond modeling, this approach is computationally efficient and suited for low-latency applications. It enables adaptive directional sound control based on speaker roles, improving hearing assistance in smart glasses through beamforming.

7. References

[1] H. Sacks, "A simplest systematics for the organization of turn taking for conversation," 1978.

[2] S. Duncan Jr, "On the structure of speaker–auditor interaction during speaking turns1," *Language in society*, vol. 3, no. 2, pp. 161–180, 1974.

[3] C. Riest, A. B. Jorschick, and J. P. de Ruiter, "Anticipation in turn-taking: mechanisms and information sources," *Frontiers in Psychology*, vol. 6, p. 89, 2015.

[4] Z. Degutyte and A. Astell, "The role of eye gaze in regulating turn taking in conversations: a systematized review of methods and findings," *Frontiers in Psychology*, vol. 12, p. 616471, 2021.

[5] K. Jokinen, H. Furukawa, M. Nishida, and S. Yamamoto, "Gaze and turn-taking behavior in casual conversational interactions," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 3, no. 2, pp. 1–30, 2013.

[6] K. H. Kendrick, J. Holler, and S. C. Levinson, "Turn-taking in human face-to-face interaction is multimodal: gaze direction and manual gestures aid the coordination of turn transitions," *Philosophical Transactions of the Royal Society B*, vol. 378, no. 1875, p. 20210473, 2023.

[7] L. Hirvenkari, J. Ruusuvuori, V.-M. Saarinen, M. Kivioja, A. Peräkylä, and R. Hari, "Influence of turn-taking in a two-person conversation on the gaze of a viewer," *PloS One*, vol. 8, no. 8, p. e71569, 2013.

[8] E. Ekstedt and G. Skantze, "Turngpt: a transformer-based language model for predicting turn-taking in spoken dialog," *arXiv preprint arXiv:2010.10874*, 2020.

[9] S.-y. Chang, B. Li, T. N. Sainath, C. Zhang, T. Strohman, Q. Liang, and Y. He, "Turn-taking prediction for natural conversational speech," *arXiv preprint arXiv:2208.13321*, 2022.

[10] F. Kurata, M. Saeki, S. Fujie, and Y. Matsuyama, "Multimodal turn-taking model using visual cues for end-of-utterance prediction in spoken dialogue systems," *Proc. Interspeech 2023*, pp. 2658–2662, 2023.

[11] R. Ishii, K. Otsuka, S. Kumano, and J. Yamato, "Prediction of who will be the next speaker and when using gaze behavior in multiparty meetings," *ACM Transactions on Interactive Intelligent Systems (TIIS)*, vol. 6, no. 1, pp. 1–31, 2016.

[12] M.-C. Lee, M. Trinh, and Z. Deng, "Multimodal turn analysis and prediction for multi-party conversations," in *Proceedings of the 25th International Conference on Multimodal Interaction*, 2023, pp. 436–444.

[13] P. Wang, E. Han, A. Queiroz, C. DeVeaux, and J. N. Bailenson, "Predicting and understanding turn-taking behavior in open-ended group activities in virtual reality," *arXiv preprint arXiv:2407.02896*, 2024.

[14] H. Jiang, C. Murdock, and V. K. Ithapu, "Egocentric deep multi-channel audio-visual active speaker localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 544–10 552.

[15] Y. Yin, I. Ananthabhotla, V. K. Ithapu, S. Petridis, Y.-H. Wu, and C. Miller, "Hearing loss detection from facial expressions in one-on-one conversations," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 5460–5464.

[16] C. Murdock, I. Ananthabhotla, H. Lu, and V. K. Ithapu, "Self-motion as supervision for egocentric audiovisual localization," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 7835–7839.

[17] H. Yun, R. Gao, I. Ananthabhotla, A. Kumar, J. Donley, C. Li, G. Kim, V. K. Ithapu, and C. Murdock, "Spherical world-locking for audio-visual localization in egocentric videos," in *European Conference on Computer Vision*. Springer, 2025, pp. 256–274.

[18] J. Engel, K. Somasundaram, M. Goesele, A. Sun, A. Gamino, A. Turner, A. Talattof, A. Yuan, B. Souti, B. Meredith *et al.*, "Project aria: A new tool for egocentric multi-modal ai research," *arXiv preprint arXiv:2308.13561*, 2023.

[19] S. Liu, Y. Nakajima, L. Chen, S. Arndt, M. Kakizoe, M. A. Elliott, and G. B. Remijn, "How pause duration influences impressions of english speech: Comparison between native and non-native speakers," *Frontiers in Psychology*, vol. 13, p. 778018, 2022.

[20] E. Campione and J. Véronis, "A large-scale multilingual study of silent pause duration," in *Speech Prosody 2002, International Conference*, 2002.

[21] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[22] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[23] Y. Kim, C. Denton, L. Hoang, and A. M. Rush, "Structured attention networks," *arXiv preprint arXiv:1702.00887*, 2017.

[24] Y. S. Aurelio, G. M. De Almeida, C. L. de Castro, and A. P. Braga, "Learning from imbalanced data sets with weighted cross-entropy function," *Neural Processing Letters*, vol. 50, pp. 1937–1949, 2019.

[25] J. Opitz and S. Burst, "Macro f1 and macro f1," *arXiv preprint arXiv:1911.03347*, 2019.

[26] R. Ishii, K. Otsuka, S. Kumano, and J. Yamato, "Using respiration to predict who will speak next and when in multiparty meetings," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 6, no. 2, pp. 1–20, 2016.

[27] S. Hoppe and A. Bulling, "End-to-end eye movement detection using convolutional neural networks," *arXiv preprint arXiv:1609.02452*, 2016.

[28] C. Elmadjian, C. Gonzales, R. L. d. Costa, and C. H. Morimoto, "Online eye-movement classification with temporal convolutional networks," *Behavior Research Methods*, vol. 55, no. 7, pp. 3602–3620, 2023.