Proposal: Who Are You Talking To? A Gaze-Aware Conversational Robot for Multi-User Interaction

Kevin Zhou
yrzhou@umich.edu

Zheng Wang
zzww@umich.edu

***Abstract*—In natural human communication, gaze direction and eye contact play a crucial role in regulating conversational turn-taking and indicating the addressee of speech. However, many conversational robots lack this contextual awareness and respond indiscriminately to all speech inputs, often producing socially inappropriate interruptions. For example, in multi-user scenarios in ChatGPT's video chat mode, the system may respond to speech directed between human participants, creating confusion and disrupting natural conversation flow. This project will develop a lightweight yet effective, gaze-aware virtual robot that will infer whether a user is speaking to it or to another person in a multi-party interaction. The robot will run on standard laptop hardware, using a webcam and microphone for perception, and will be represented as a virtual avatar. By integrating real-time gaze estimation and speech detection, the robot will determine whether to respond and visually orient toward the intended speaker. This prototype will demonstrate how simple multimodal perception can enable socially intelligent, context-aware human-robot interaction without requiring heavy computational resources, with future deployment to physical robotic platforms planned.**

I. Related Work

In natural human communication, non-verbal cues, particularly gaze, are crucial for managing turn-taking and signaling the intended addressee of speech [1, 5]. Many conversational robots, however, lack this contextual awareness and respond indiscriminately to all speech, leading to socially inappropriate interruptions. This project addresses the "who is speaking to whom" (W2W) challenge, which requires a system to perceive and understand the conversational situation before responding.

The literature on this W2W problem can be broadly split into two main approaches, neither of which is suitable for lightweight, real-time, embodied interaction. The first is a text-based, Natural Language Processing (NLP) approach. Le et al. [6] introduced "who-to-whom (W2W)" models to identify addressees in multi-party text corpora. These models are fundamentally non-embodied and often non-causal (using subsequent text to make a prediction), making them incompatible with a real-time robot that must react as interaction happens.

The second approach, found in Human–Robot Interaction (HRI), uses computationally heavy, data-driven models. Haefflinger et al. [3] developed a data-driven end-to-end model trained on a large, bespoke dataset to generate nuanced, human-like robot eye and head movements. Similarly, Gillet et al. [2] learned gaze behaviors for balancing participation in group HRI, showing that learned gaze policies can redistribute speaking opportunities among participants. Tennent et al. [7] introduced Micbot, a peripheral robotic device designed to shape conversational dynamics and improve team performance in small-group settings, demonstrating how nonverbal robot behavior can influence participation balance. More recently, Wang et al. [8] presented LaMI, a large language model framework for multimodal HRI, integrating verbal and nonverbal behaviors such as gaze and gesture for context-aware interaction. While these systems achieve sophisticated multi-user coordination, they rely on complex multimodal pipelines and significant computational resources, making them impractical for deployment on standard hardware.

Other lightweight multimodal work, such as Heo et al. [4] on gaze-enhanced turn-taking prediction, is also emerging. However, this research focuses on predicting transitions between speakers for applications like hearing assistance, rather than solving the W2W problem to determine if a robot should reply.

II. Methodology

The proposed system will enable a virtual robot to detect which human participant in a multi-person conversation is speaking to it and to respond only when it is the intended addressee. The approach will integrate gaze and speech perception to infer conversational intent and trigger socially appropriate responses. The processing pipeline will consist of four main components: gaze detection, speech detection, addressee inference, and dialogue generation.

Gaze Detection. We will implement gaze estimation using the Mediapipe FaceMesh framework for real-time 3D facial landmark tracking. For each participant, the system will compute head orientation based on the relative positions of the nose and eye centers. If the orientation vector lies within a small angular threshold toward the camera, the participant will be considered to be *looking at* the robot. This binary gaze state will be updated continuously at the video frame rate.

Speech Detection. The system will continuously monitor microphone input to determine who is currently speaking. We will employ voice activity detection (VAD) and optional speech-to-text transcription using either PyAudioAnalysis or OpenAI Whisper. Each utterance will be associated with the detected speaker identity and a confidence score, ensuring the system reacts only to deliberate vocal input rather than background noise.

Addressee Inference. The decision layer will fuse gaze and speech signals to estimate whether the robot is being addressed. When a participant is speaking and simultaneously looking at the robot, the system will infer that the utterance is directed toward it. If a participant is speaking while looking elsewhere, the system will assume the conversation is between the humans and remain idle. In ambiguous cases, such as when multiple users look at the robot while speaking, the system will enter a neutral state and may issue a clarification prompt. This rule-based logic will enable robust, interpretable behavior without requiring large training datasets.

Dialogue Generation and Response. Once an addressee is identified, the robot will generate an appropriate verbal and visual response. Responses will be either predefined rule-based phrases or dynamically generated using the OpenAI GPT-4o-mini model. The system will convert text output into speech via a lightweight TTS engine such as `pyttsx3` or OpenAI's built-in TTS API. A simple 2D avatar displayed in a Python interface (e.g., Pygame or Streamlit) will visually orient toward the speaking participant to reinforce mutual attention.

Overall, this multimodal pipeline will allow the virtual robot to maintain socially appropriate engagement, responding only when directly addressed. A key design goal is to achieve both effectiveness in addressee detection and lightweight implementation for practical deployment. By combining efficient computer-vision and audio-processing modules, the system will achieve real-time performance on standard laptop hardware using only a webcam and microphone, without requiring GPU acceleration or cloud-based processing. While the current implementation focuses on a virtual robot interface, the modular design will facilitate future deployment to physical robotic platforms.

III. Evaluation

The evaluation will focus on testing whether the gaze-aware mechanism improves the robot's ability to respond appropriately in multi-party settings. We will primarily perform the evaluation ourselves as the two project team members, using controlled dialogues to simulate different conversational scenarios. If time permits, we will invite a few additional participants to gather more qualitative observations.

Evaluation Setup. Each test session will involve multiple human participants and the virtual robot. We will compare two interaction modes: an *Always-On Robot*, which responds to all detected speech regardless of gaze direction, and a *Gaze-Aware Robot*, which responds only when the user is both looking at and speaking to the robot. During the sessions, we will conduct multiple short conversation trials that include direct communication with the robot, side conversations between participants, and ambiguous situations where multiple users speak simultaneously.

Evaluation Metrics. We will qualitatively assess the system's responsiveness and appropriateness of behavior under each condition. Key observations will include whether the robot responds at the correct moments, how well it ignores unrelated speech, and how natural the gaze-based attention switching appears. Quantitatively, we will log instances of false responses (robot responds when not addressed), missed responses (robot fails to respond when addressed), and average response delay. If additional participants are available, we will also collect brief subjective ratings of perceived naturalness and social appropriateness using 5-point Likert scales.

Expected Outcome. We expect that the gaze-aware robot will demonstrate fewer inappropriate responses and appear more contextually aware than the always-on baseline. Even with a small sample size, these tests will help verify the system's basic functionality and highlight potential improvements for future user studies and deployment to physical robotic platforms.

IV. Timeline

Week 1: We will conduct a literature review on gaze-based engagement and addressee detection. We will finalize the overall system architecture and set up the development environment.

Week 2: We will implement and test the core perception modules, including gaze detection and speech activity detection. We will verify that both modules run in real time and output stable signals.

Week 3: We will integrate the gaze and speech modules into a unified addressee inference pipeline. We will develop the dialogue generation logic and the visualization interface for the virtual robot.

Week 4: We will conduct user testing with team members under both always-on and gaze-aware conditions. We will collect quantitative and qualitative evaluation data and refine system behavior based on feedback.

Week 5: We will analyze the evaluation data, prepare the final report, and finalize the project presentation.

References

[1] Ziedune Degutyte and Arlene Astell. The role of eye gaze in regulating turn taking in conversations: A systematized review of methods and findings. *Frontiers in Psychology*, Volume 12 - 2021, 2021.

[2] Sarah Gillet, Maria Teresa Parreira, Marynel Vázquez, and Iolanda Leite. Learning gaze behaviors for balancing participation in group human-robot interactions. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 265–274, 2022. doi: 10.1109/HRI53351.2022.9889416.

[3] Léa Haefflinger, Frédéric Elisei, and Gérard Bailly. Data-driven control of eye and head movements for triadic human-robot interactions. *International Journal of Social Robotics*, pages 1–22, 2025.

[4] Seongsil Heo, Calvin Murdock, Michael Proulx, and Christi Miller. Gaze-enhanced multimodal turn-taking prediction in triadic conversations, 2025. URL https://arxiv.org/abs/2505.13688.

[5] Kristiina Jokinen, Hirohisa Furukawa, Masafumi Nishida, and Seiichi Yamamoto. Gaze and turn-taking behavior in casual conversational interactions. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 3(2):1–30, 2013.

[6] Ran Le, Wenpeng Hu, Mingyue Shang, Zhenjun You, Lidong Bing, Dongyan Zhao, and Rui Yan. Who is speaking to whom? learning to identify utterance addressee in multi-party conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1909–1919, 2019.

[7] Hamish Tennent, Solace Shen, and Malte Jung. Micbot: a peripheral robotic object to shape conversational dynamics and team performance. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction*, HRI '19, page 133–142. IEEE Press, 2020. ISBN 9781538685556.

[8] Chao Wang, Stephan Hasler, Daniel Tanneberg, Felix Ocker, Frank Joublin, Antonello Ceravola, Joerg Deigmoeller, and Michael Gienger. Lami: Large language models for multi-modal human-robot interaction. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI '24, page 1–10. ACM, May 2024. doi: 10.1145/3613905.3651029. URL http://dx.doi.org/10.1145/3613905.3651029.