Google Slides Link

https://docs.google.com/presentation/d/1wtq0cUCGtjrROVQmu2qx6MEfxbe3cRcKeRxInHNW57o/edit?usp=sharing

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

12/8/2025

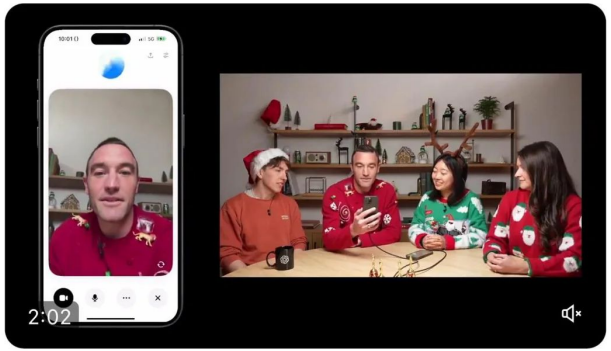
ROB 498 /599: Computational Human-Robot Interaction Fall 2025

Presenter: Zheng Wang, Kevin Zhou

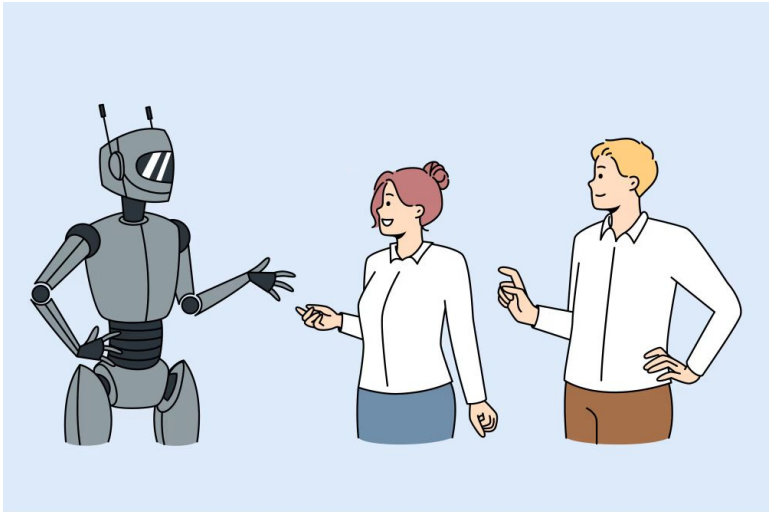
Motivation

OpenAI 
@OpenAI

Just in time for the holidays, video and screensharing are now starting to roll out in Advanced Voice in the ChatGPT mobile app.

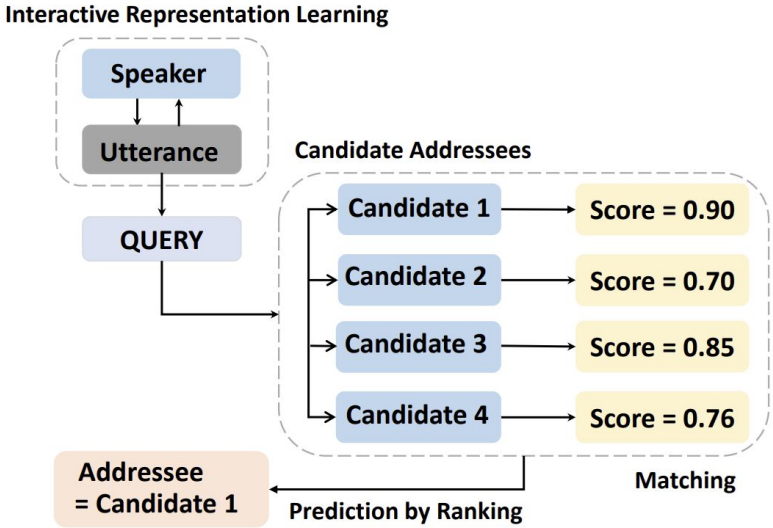


2:38 PM · 12/12/24



Prior Works

- Text-Based NLP Approaches
  - Who is speaking to whom? learning to identify utterance addressee in multi-party conversations, Le et al., 2019

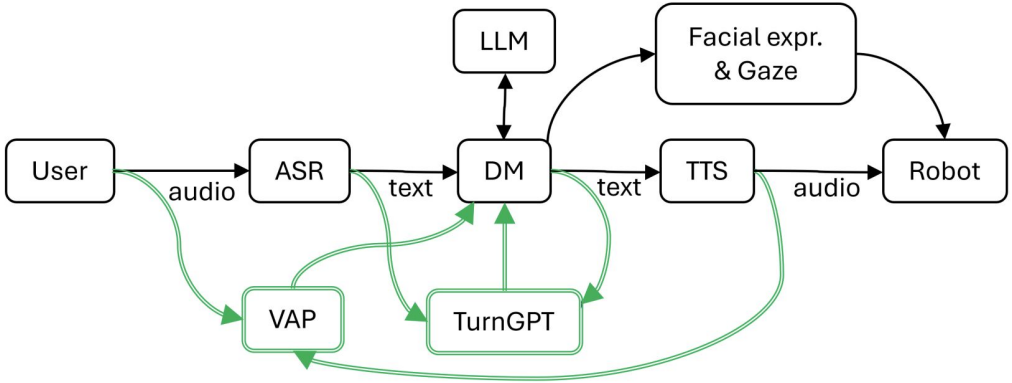


Prior Works

- Multimodal Fusion Approaches
  - Multimodal Turn Analysis and Prediction for Multi-party conversations, Lee et al., 2023.

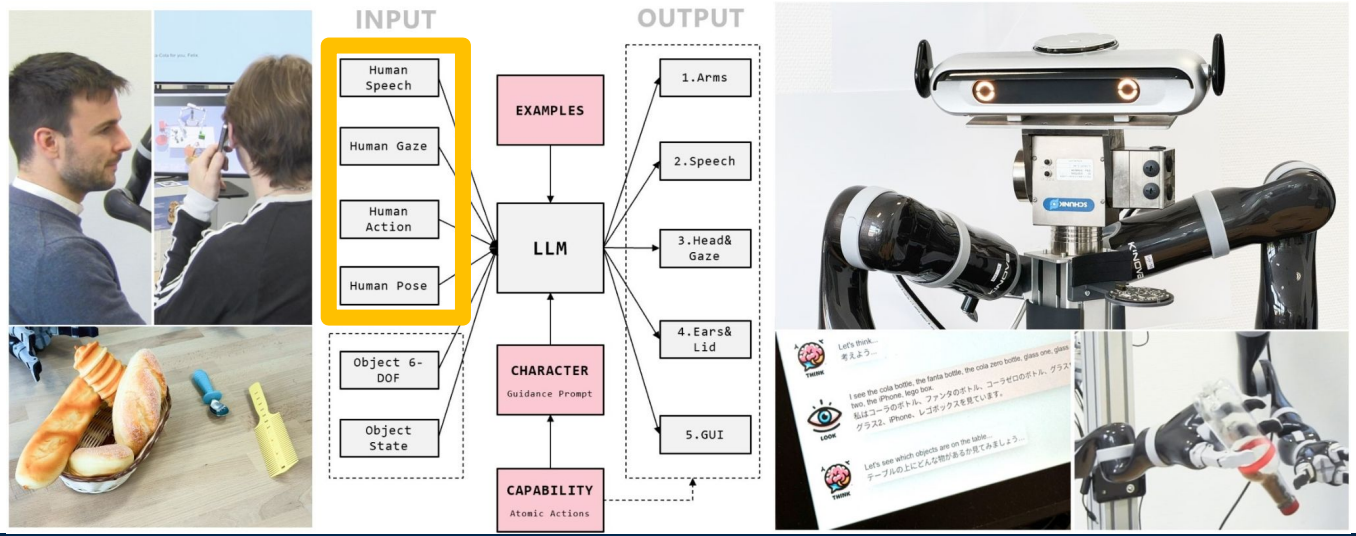
Prior Works

- Multimodal Fusion Approaches
  - Applying General Turn-taking Models to Conversational Human-Robot Interaction, Skantze et al., 2025.



Prior Works

[LaMI: Large Language Models for Multi-Modal Human-Robot Interaction] - human speech, gaze, action, pose

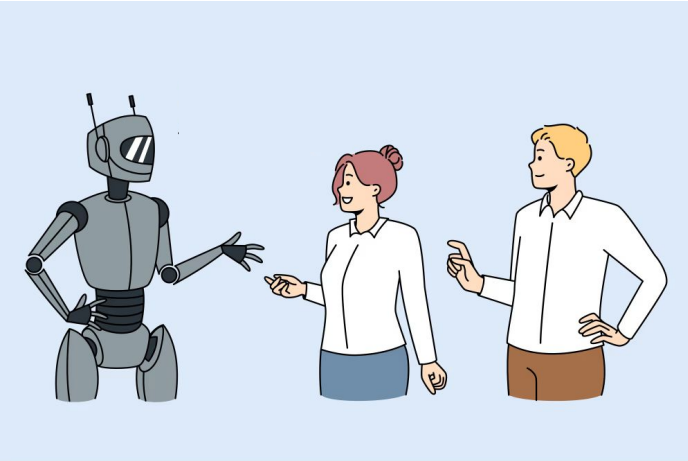


Key Insight

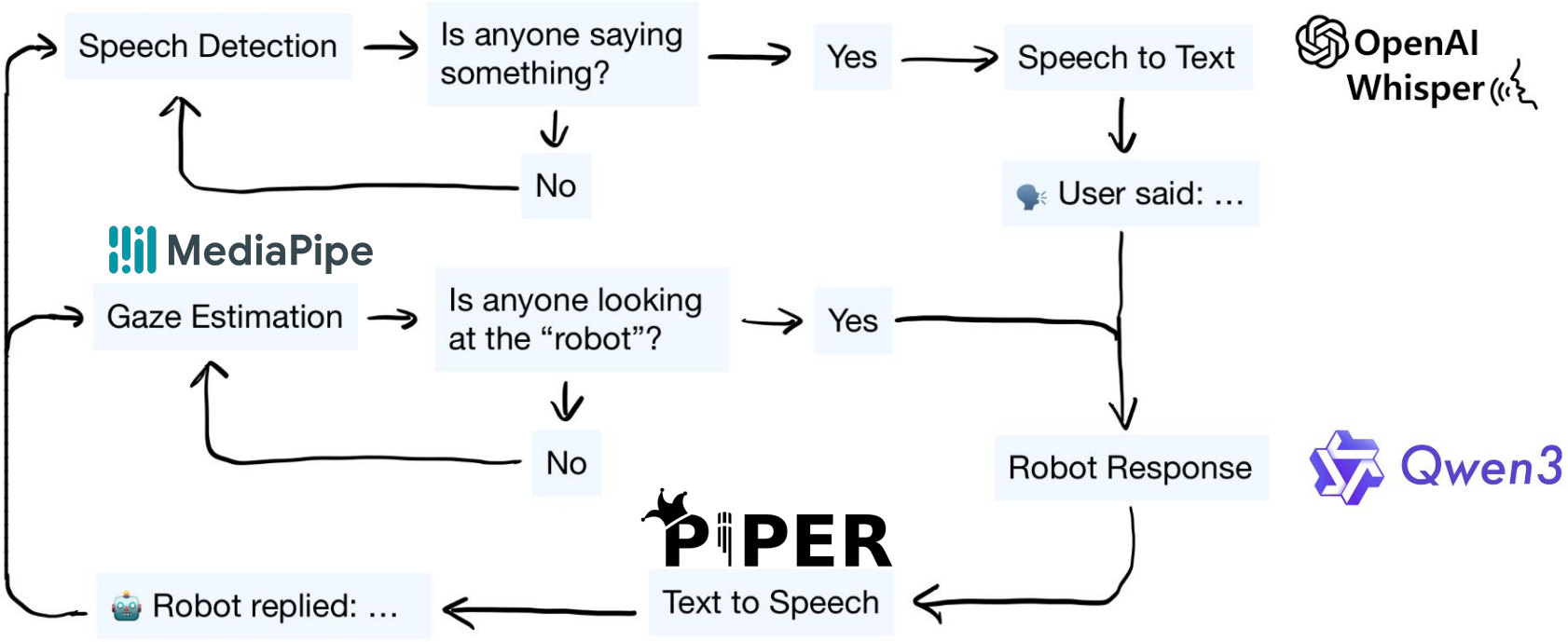
Who Are You Talking To?

A Gaze-Aware Conversational Robot for Multi-User Interaction

- No heavy models
- No special hardware
- Speech activity + gaze estimation



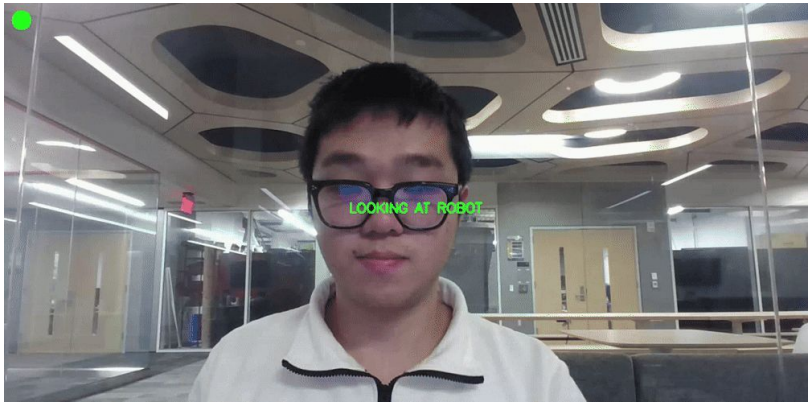
System Overview

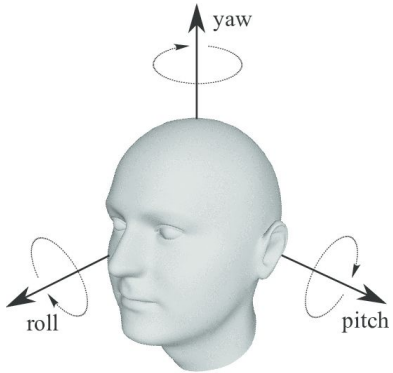


Technical Approach

Gaze Estimation: [MediaPipe Face Mesh]

- THRESHOLD_YAW = 7
- THRESHOLD_PITCH = 10





Technical Approach

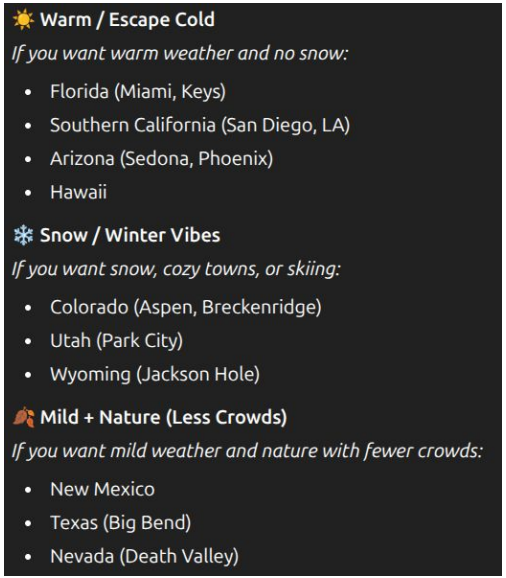
- Speech Detection: SpeechRecognition 3.14.4

```
pip install SpeechRecognition
```

- Speech to Text: OpenAI Whisper
- Text to Speech: PIPER

Technical Approach

Robot Response: [Qwen3-4B-Instruct]



"Keep your responses brief and concise - limit your answers to 1-3 sentences. This is for a voice interface, so be conversational but brief. NO EMOJIS ALLOWED."

If you want warm weather and no snow, go to Florida, Southern California, Arizona, or Hawaii.
If you want snow, cozy towns, or skiing, Colorado, Utah, and Wyoming are great winter picks.
If you prefer mild weather with beautiful nature and fewer crowds, consider New Mexico, Big Bend in Texas, or Death Valley in Nevada.


SPEAKING...
help?
Assistant: Ask me anything? Questions, problems, or just want to chat. What's on your mind?

Evaluation

- 4 conversations, 4-8 rounds of discussions in each, 24 in total
- 11 rounds are supposed to be human-human
- 13 rounds are supposed to be human-robot

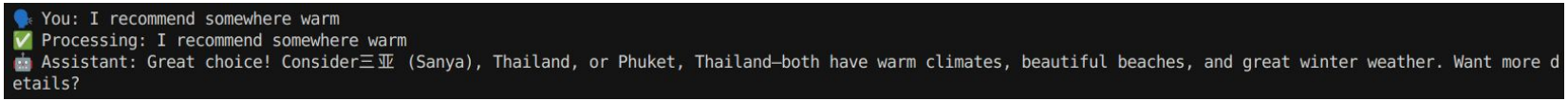
	Our system	ChatGPT
Inappropriate interrupts (max: 11)	1	11
Missed responses (max: 13)	3	0
Latency of responses	2.24 seconds	1.98 seconds
Quality of responses (Third-party ratings)	7/10	9.5/10
Preference (Third-party ratings)	1	1

Strengths

- Explicit addressee awareness in multi-user settings
- Lightweight, runs locally
- No external hardware needed
- Relatively low latency
- Has the potential to be deployed on real robots

Weakness

- Gaze estimation not always reliable
  - Not accurate when users are in the corners of the frame
  - Sensitive to light conditions
- Doesn’t know if person A or person B is speaking
  - Need another model to do this
- User cannot interrupt the robot when it is speaking
- LLM & Speech to Text makes mistakes



Connection with HRI

- Verbal communication
- Nonverbal communication
- Groups and teams

Thank you!

Questions?