Vision and Speech Language Models for Emotional Interpretation

Jonathan Liu, Princeton University, jl0796@princeton.edu

1 Introduction and Problem Statement

For remote workers who are blind or have low vision, recognizing social cues for emotional interpretation and feedback is a major challenge, since the majority of social cues such as smiling, nodding, or head-shaking are visual. Furthermore, unlike in-person interactions, audible reactions such as laughter are likely not transmitted since online participants are typically muted.

This paper reports the author's latest progress to address these needs, based on the framework presented in *Hear the Room*, winner of the 2024 MIT AI for Accessibility Hackathon (HackDisability). Here, we implement the solution using the latest large language models for understanding coordinated speech and vision.

The preliminary results are positive, with approximately 93% accuracy for cue identification.[1] Further research is needed to avoid hallucination and improve the tool's ease of access.

2 Background and Methods

Social cues convey two types of information: they reflect how people feel and also serve as responses to specific actions or prompts. Identification of cues is inherently multimodal because the prompt for a social cue is text- or speech-based, yet the cues themselves—especially in an online setting—are primarily visual. This project aims to make visual social cues accessible to blind or low-vision individuals, enhancing their ability to interpret emotional feedback in online meetings.

Figure 1 illustrates the multistep pathway the author proposes to obtain the necessary causal relationships between speech and visual social cues over time. First, a vision language model is used to identify the participants in the meeting. Then, the speech is transcribed with an audio large language model to obtain a segmented transcript of the video. After syncing the captions with the video, we use the vision model to identify the participant that is speaking in each segment. Finally, we evaluate each segment of the original video using the vision model to identify social cues that are expressed by listeners in each segment.

[1]The code is available at:
https://github.com/liujonathan24/Facial-Cue-Identification

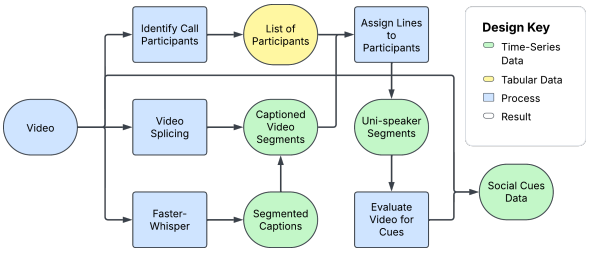

Figure 1: The Data Processing Pipeline: We preprocess the video to obtain the List of Participants and Single Speaker Captioned Video Segments, which are then evaluated for Social Cue Data from listeners.

3 Experiment and Results

We test our data pipeline on a 4-minute clip of a Zoom call from Gitlab's "Code Review Weekly Workshop" (GitLab, 2022). For our vision language and audio language models, we chose Qwen2.5-VL 7B (Bai et al., 2025) and Faster-Whisper (Radford et al., 2022). For the purpose of real-time computation, we first downsample the video to 5 frames per second with a resolution of 756 by 448. For simplicity, we choose nodding as an initial social cue to evaluate.

By applying the experimental procedure, the participant names are correctly extracted from the video. One notable strength of the model is its ability to consistently extract names from Zoom titles, which may contain metadata such as roles or affiliations. Then, when given the correct speaker for each segment, the model performs with 93% accuracy in detecting listener nods and associating them with the corresponding prompt, running in real-time.

4 Future Work and Discussion

Ongoing experiments segment the video feed using bounding boxes, to isolate each meeting participant's camera feed. This allows us to control individual-specific prompts, reducing the possibility of hallucination. Bounding boxes also assign spatial locations to each of the members of the call, enabling precise, spatially correct audio output.

In addition, fine-tuning language models with PEFT or LoRA/QLoRA and comparing different vendors may be necessary to increase performance to an operational level (Zhao et al., 2023).

References

Shuai Bai, Keqin Chen, and Xuejing Liu et al. 2025. Qwen2.5-VL Technical Report. *Preprint*, arXiv:2502.13923.

GitLab. 2022. Code Review Weekly Workshop - Sep 23, 2022. Accessed 2025-03-15.

HackDisability. 2024 Hackathon: AI for Accessibility. MIT CSAIL and Perkin's School for the Blind, February 22-25th, 2024. Sponsored by Amazon. News coverage at: https://www.bostonglobe.com/2024/03/19/business/accessible-apps-blind-deaf-perkins-school/.

Alec Radford, Jong Wook Kim, and Tao Xu et al. 2022. Robust Speech Recognition via Large-Scale Weak Supervision, OpenAI. *Preprint*, arXiv:2212.04356.

Ruochen Zhao, Hailin Chen, and Weishi Wang et al. 2023. Retrieving Multimodal Information for Augmented Generation: A Survey. *Preprint*, arXiv:2303.10868.