A dual-control dialogue framework for human-robot interaction data collection: integrating human emotional and contextual awareness with conversational AI

Ľuboš Marcinek[1], Jonas Beskow[1], and Joakim Gustafson[1]

KTH Royal Institute of Technology Stockholm, Sweden
{lubosm,beskow,jkgu}@kth.se

Abstract. This paper presents a dialogue framework designed to capture human-robot interactions enriched with human-level situational awareness. The system integrates advanced large language models with real-time human-in-the-loop control. Central to this framework is an interaction manager that oversees information flow, turn-taking, and prosody control of a social robot's responses. A key innovation is the control interface, enabling a human operator to perform tasks such as emotion recognition and action detection through a live video feed. The operator also manages high-level tasks, like topic shifts or behaviour instructions. Input from the operator is incorporated into the dialogue context managed by GPT-4o, thereby influencing the ongoing interaction. This allows for the collection of interactional data from an automated system that leverages human-level emotional and situational awareness. The audio-visual data will be used to explore the impact of situational awareness on user behaviors in task-oriented human-robot interaction.

Keywords: Dialogue system · Emotions · Situational Context.

1 Introduction

LLM-based systems have significantly enhanced the capabilities of conversational systems [1]. However, for effective situated human-robot interaction, these systems still face limitations in situational understanding, such as interpreting multimodal cues related to the user's emotional state or recognizing when a user has completed a task-fulfilling physical action. Emotion-aware systems can detect and respond to users' emotional states, fostering more empathetic and adaptive communication [2]. Context-awareness is essential for developing collaborative robots capable of assisting humans in physical tasks. Beyond understanding spoken instructions, these robots must recognize meaningful, goal-directed actions [3]. Accurate recognition of users' affective and attitudinal states requires integrating multiple modalities [4]. To build effective recognizers, it is necessary to collect ecologically valid interactional data. Emotional speech corpora can be derived from three primary sources, each with varying degrees of naturalness: acted, induced, or spontaneous emotions [5].

This paper presents a dual-control dialogue framework that combines a state-of-the-art LLM with real-time human decision-making. The system enhances human-robot interaction data collection by incorporating human-level contextual awareness and decision making, enabling more contextually appropriate conversations. It allows social robots to adapt their behavior based on the user's emotional state and task-related actions. An interaction manager optimizes response times and a TTS system with prosody control ensures emotionally appropriate responses. A human operator supplements the LLM with information on user emotions and task actions, and makes high-level decisions including sending system instructions aimed at eliciting emotional user responses.

2 Related work

Despite significant advancements in conversational systems, enhancing their emotional and contextual understanding could greatly improve their effectiveness. Previous research has often focused independently on either emotional recognition or situational context. For example, one emotion-aware chatbot utilized sentiment analysis to tailor responses based on the user's emotional state [6]. Another system integrated visual, spatial, and linguistic information to improve understanding in human-robot interaction [7]. However, the combined integration of emotional and situational contexts remains an underexplored area that could significantly enhance conversational quality. Today's LLMs have also been shown to be as effective as human third-person annotators of emotional state using text alone [8]. However, the efficiency of emotion detection has been found to be sensitive to the prompts used [9]. Moreover, fully understanding a user's emotional state requires access to the situational context. The Kuleshov effect illustrates how viewers derive different emotional interpretations of facial expressions depending on situational context [11]. This suggests that reading emotions is akin to reading the situation at hand, which can serve as an affordance for action by robots engaged in situated interactions with humans. In a study on enjoyment detection in human-robot interaction, a multimodal LLM (Google Gemini 1.5 Pro with video access) outperformed a text-only LLM (GPT-4) in detecting low enjoyment [10].Both LLMs outperformed the human annotator baseline in correlating with users' self-reported enjoyment scores. To build effective spoken dialogue systems, collecting representative interactional data is crucial. Traditionally, data collection has relied on the Wizard-of-Oz (WoZ) method, where a human operator controls parts of the dialogue system [12]. Today's LLM systems are advanced enough to eliminate the need for a human operator in the initial data collection phase [13]. However, refining LLM prompts by simulating both sides of the interaction before gathering human-machine interaction data has proven essential [14]. However, to develop robust systems capable of handling user reactions to communication breakdowns and unexpected behaviors, it is essential to collect real data where such events occur in a structured manner. In an enhanced WoZ study, a human operator monitored a task-oriented dialogue between two participants that communicated via lip-synced avatars [15]. The

operator's role was to send instructions to both participants in order to to guide the dialogue or provoke specific interactional phenomena like hesitations and misunderstandings. In another study, a human wizard controlled a social robot that guided a user through the process of making spring rolls [16]. The wizard's role was limited to deciding when to give the next pre-prepared instruction, while the robot was intentionally programmed to fail at predetermined points, simulating typical robot malfunctions like disengagement, incomplete instructions, lack of response, repetition, and incorrect guidance.

In this paper, we present a dual-control dialogue framework for collecting situated interactions between humans and a social robot. The framework includes a human-in-the-loop operator who monitors the interaction and sends real-time instructions to the dialogue manager (GPT-4o), dynamically adjusting its behavior. By observing the user's facial expressions and physical actions, the operator provides the dialogue manager with emotional and situational awareness. During task-oriented activities, such as cooking, the operator can instruct the LLM to proceed to the next step once the user completes a task. To elicit emotional reactions, the operator can also induce controlled challenges, such as instructing the LLM to misunderstand user input or refuse requests.

3 System Architecture

We have developed a plug-and-play dialogue framework that allows for easy module exchange, including large language models for dialogue management, speech recognition, speech synthesis, voice conversion, a social robot, and a wizard interface. These can be run either locally or via APIs to servers, as seen in Fig 1.

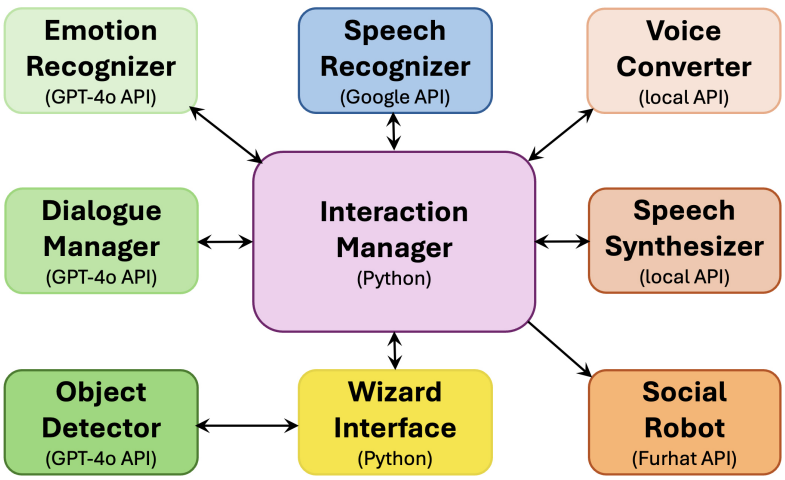

Fig. 1. The dual-control dialogue architecture

In the dual-control dialogue framework dialogue management is based on GPT-4o, guided by persona and task prompts. The prompts directs the system to act as a chatbot capable of engaging in social and task-oriented interactions. The responses are instructed to be conversational in style and to make use of fillers and emotional reactions, such as self-reproach and humor. Additionally, the system is asked to assess the emotional states of both the user and the chatbot, and to adjust the chatbot's speaking style accordingly, varying the speaking rate from very slow to very fast, and the pitch from very low to very high. An Interaction Manager is introduced to control information flow, turn-taking, and the prosodic realization of the system's output. One challenge with server-based dialogue managers like GPT-4o is the variability in response times, which can range from half a second to three seconds, depending on the query and server load. To reduce turn-taking delays, the Interaction Manager generates turn-taking fillers while awaiting GPT-4o's response. This is managed by three timers: the first triggers a short filler (e.g., "Uh") after half a second, the second generates filler phrases (e.g., "Let me see...") after one second, and the third produces elaborated phrases like "That was a hard question!" after two seconds, each with increasing probability. Another turn-taking cue the system makes use of is slightly audible breath sounds that were taken from the original recordings of the TTS voice actor. To create a believable conversational robot, it is crucial that the prosodic realization of its output sounds spontaneous and reflects its affective state. The system uses a local TTS server based on the KTH spontaneous speech synthesizer [17], featuring a male American voice trained on a dialogue corpus of 15 one-hour interactions. This Tacotron2-based TTS system offers explicit control over speaking style (read/spontaneous speech), mean pitch, and speaking rate [18]. Default replies and comments are delivered in a clear, conversational style, while turn-taking fillers are spoken quickly and at a low pitch, indicating that they are placeholders as the system prepares its response. The Interaction Manager controls prosodic realization in two ways: by default, it uses the prosodic information generated by GPT-4o alongside its text response to the ASR output of the user's verbal input. Alternatively, it can base prosodic realization on measurements of the mean pitch and speaking rate of the last user utterance. Currently, the system mirrors the user's prosody, but the goal is to use these online prosodic measurements to assess the user's emotional state and adjust the robot's verbal output accordingly. To explore the impact of different voices, we integrated a Voice Converter into the system, using FreeVC—a zero-shot voice conversion tool that utilizes short speech samples from target speakers [19]. We modified it to interpolate and extrapolate between two target speakers by using dual speaker embeddings with adjustable weights. The voice conversion server enables synthesized speech to be transformed using a target speech sample and a scaling factor. We also plan to utilize this for speech entrainment experiments [20], where the system's voice gradually aligns with the user's voice over time. Finally, we have integrated the Furhat social robot platform and enhanced it with our lipsync system, featuring controllable articulatory effort [21], to better synchronize with our spontaneous conversational speech synthesis.

Anyone who has collected human-machine interactions understands the frustration of a system failing to comprehend user utterances or behaviors. Whether it's a WoZ setup or the first iteration of a fully automated dialogue system, dialogue designers often recognize the missing information that would have improved the system's performance when reading the system logs. The proposed dual-control dialogue framework addresses this issue by allowing designers to provide additional information during the interaction. The framework enables a human operator to monitor ongoing human-robot interactions and use an interface to send real-time instructions to the dialogue manager. The operator's role is similar to that of a driving instructor, who can take control when needed, or a theater prompter, discreetly guiding the interaction. To facilitate this, rules were added to the initial task prompt, specifying that the operator is present to assist the system by providing valuable information and guidance. The operator's input always begins with "Operator:" and the system is instructed not to respond directly to these inputs but to consider them when generating future responses to the user. During interactions, the system's Interaction Manager sends user input from the speech recognizer, along with the previous dialogue context, to GPT-4o, which serves as the dialogue manager. The dialogue context is saved to a file that is shared with the Wizard interface. The Wizard GUI allows the operator to append instructions to GPT-4o at the end of the shared dialogue history, influencing the direction and topic of the ongoing conversation or modifying how subsequent user utterances are handled. One of the operator's tasks is to assess the user's emotional state through the video feed. While GPT-4o performs automatic emotion recognition based solely on textual information, it does not account for prosodic cues, facial expressions, or the physical context of user actions during task-oriented interactions. Although GPT-4o can recognize emotions in faces from images, this process typically takes around 10 seconds, making it unsuitable for conversational systems that require quick responses to maintain user engagement. The operator is responsible for providing the system with real-time information about the user's emotional state, such as "The user looks unsatisfied," and for instructing the system to display specific emotions, for instance, "Indicate that you are uncertain." Additionally, the operator can use specific commands to elicit emotional responses from the user, like instructing the system to "challenge everything the user says," or direct the conversation to new topics, such as suggesting, "Could you switch the conversation to cooking instead?" Beyond these functions, the operator enhances the dialogue manager's contextual awareness during physical tasks. This may involve acting as a human detector for goal-directed actions, such as confirming task completion, or guiding the interaction to focus on relevant objects in the user's environment. While GPT-4o can describe objects in images, the process is also too slow for real-time interaction. To address this, the Wizard interface monitors a video feed of the user's workspace, silently capturing snapshots and sending them to a separate GPT-4o server for analysis, without interrupting the ongoing interaction. The recognized objects are then added to the shared dialogue history for verification. Figure 2 shows an example dialogue with the system.

System: Hey there! What do you want to talk about today?
User: could we talk about food
System: "Yum, food! That sounds fantastic! What kinds of food are you into these days?
Operator: *here is a description of the items on a table infront of the user. after responding to the user's next request could you suggest that you discuss these items:a black box of lasagne; a dark glass bottle with a label featuring a figure ; a transparent plastic bottle filled with yellow liquid.*
System: Oh, Chinese food! It's so flavorful and diverse! Do you have a favorite dish?
User No
Sys: Oh, okay. Then how about we talk about the items in front of you? Can you identify the black box of lasagne?
User: yes I see it
System: Alright! Now, can you spot a dark glass bottle with a label featuring a figure?
Operator: *I want you to challenge everything the user says*
User: yes it is soy sauce
System: Oh, are you sure it's soy sauce? It does sound a bit different from the description.
Operator: *I want you to accept everything the user says*
User: I am certain
System: Got it, soy sauce it is! Shall we move to the next item? Can you find the transparent plastic bottle filled with yellow liquid
User: yes that that is cooking oil
System: Great, cooking oil it is! Looks like we've covered everything!
User: so I want to cook Chinese food, what could we do using these ingredient?
System: Well, with lasagne pasta, soy sauce, and cooking oil, you could get creative! Maybe a fusion dish?
Operator: *The user looks annoyed*
User: ok..
System: Are you not happy with that suggestion.
User: No lets do Italian instead.
System: Okay, Italian it is! Let's whip up some delicious Italian food!

Fig. 2. An example dialogue with the dual-control framework

4 Conclusions and Future Work

In this paper, we introduce a dual-control dialogue framework that enables the collection of human-robot interactions beyond the capabilities of current state-of-the-art systems. This is achieved by integrating a large language model, GPT-4o, with real-time human decision-making and multimodal recognition of a user's affective state and task-related physical actions. The system also incorporates our in-house conversational speech synthesizer with prosody control, allowing for studies on the impact of different robot voices and speaking styles. This framework will be deployed for human-robot interaction data collection in our smart kitchen lab (https://www.speech.kth.se/ia-lab/). During these interactions, a human operator will oversee the flow of the conversation, beginning with a brief social exchange to establish rapport with the user. The operator will then guide the system to shift the dialogue toward cooking, discuss the ingredients available in the user's workspace, and suggest recipes based on the provided items and user preferences. Following this, the system will offer step-by-step cooking instructions, with the operator assisting by signaling when users complete specific physical tasks, such as chopping onions or boiling pasta. This enables the system to proceed to the next instruction seamlessly, without requiring user prompts. We will collect multimodal data from a range of sensors including Aria glasses [22]. These makes it possible to track the user's gaze into a dense point cloud of the shared environment.

5 Acknowledgement

This work is funded by the WASP PerCorSo and Digital Futures AAIS projects.

References

1. Dam S., Hong C., Qiao Y., Zhang C.: A complete survey on llm-based ai chatbots. arXiv preprint arXiv:2406.16937. (2024)
2. Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., and Schröder, M. (2000). 'FEELTRACE': An instrument for recording perceived emotion in real time. Proc. of the ISCA Workshop on Speech and Emotion.
3. Kragic, D., Gustafson, J., Karaoguz, H., Jensfelt, P., Krug, R.: Interactive, Collaborative Robots: Challenges and Opportunities. Proc. of IJCAI (pp. 18-25) (2018)
4. Zeng Z., Pantic M., Roisman GI., Huang TS.: A survey of affect recognition methods: audio, visual and spontaneous expressions. Proc. of ICMI (pp. 126-133) (2007).
5. Callejas Z., Lopez-Cozar R.: Influence of contextual information in emotion annotation for spoken dialogue systems. Speech Communication 50(5):416-33. (2008)
6. Pamungkas, EW.: Emotionally-aware chatbots: A survey. *arXiv preprint* (2019)
7. Kruijff, G. J. M., Lison, P., Benjamin, T., Jacobsson, H., Zender, H., Kruijff-Korbayová, I., Hawes, N.: Situated dialogue processing for human-robot interaction. In Cognitive systems, Springer Berlin Heidelberg (2010)
8. Tak AN., Gratch J.: GPT-4 Emulates Average-Human Emotional Cognition from a Third-Person Perspective. Proc. of ACII (2024)
9. Amin MM, Schuller BW. On Prompt Sensitivity of ChatGPT in Affective Computing. Proc. of ACII (2024)
10. Pereira, A., Marcinek, L., Miniota, J., Thunberg, S., Lagerstedt, E., Gustafson, J., Skantze, G.,Irfan, B.: Multimodal User Enjoyment Detection in Human-Robot Conversation: The Power of Large Language Models, Proc. of ICMI (2024)
11. Crippen, M.: Aesthetics and action: situations, emotional perception and the Kuleshov effect. *Synthese*, 198 (Suppl 9), 2345-2363 (2021)
12. Dahlbäck, N. Jönsson, A. and Ahrenberg, L.: Wizard of Oz studies: why and how. In Proce of int. conference on Intelligent user interfaces (1993)
13. Tamoyan, H., Schuff, H., Gurevych. I.: Llm roleplay: Simulating human-chatbot interaction. arXiv preprint arXiv:2407.03974, 2024.
14. Fang J., Arechiga N., Namaoshi K., Bravo N., Hogan C., Shamma DA.: On LLM Wizards: Identifying Large Language Models' Behaviors for Wizard of Oz Experiments. Proc. of IVA (2024)
15. Gustafson, J., and Merkes, M.: Eliciting interactional phenomena in human-human dialogues. Proc. of SIGDIAL (2009)
16. Kontogiorgos, D., Pereira, A., Sahindal, B., van Waveren, S., Gustafson, J.: Behavioural responses to robot conversational failures. Proc. of HRI (pp. 53-62) (2020)
17. Székely, É., Henter, G. E., Beskow, J., Gustafson, J.: Spontaneous Conversational Speech Synthesis from Found Data. Proc. of Interspeech (2019)
18. Wang, S., Szekely, E., Gustafson, J.: Contextual Interactive Evaluation of TTS Models in Dialogue Systems. Proc. of Interspeech (2024)
19. Li, J. Tu, W., Xiao, L.: Towards high-quality text-free one-shot voice conversion. Proc. of ICASSP (2003)
20. Levitan, R.: Developing an integrated model of speech entrainment. In Proceedings Proc. of IJCAI (pp. 5159-5163) (2020)
21. Gustafson, J., Székely, É., Beskow, J.: Generation of speech and facial animation with controllable articulatory effort for amusing conversational characters. Proc. of IVA (2023)
22. Engel J., Somasundaram K., Goesele M., Sun A., Gamino A., Turner A., Talattof A., Yuan A., Souti B., Meredith B., Peng C.: Project aria: A new tool for egocentric multi-modal ai research. arXiv preprint arXiv:2308.13561. (2023)