A Multimodal Dialogue System to Lead Consensus Building with Emotion-Displaying

Shinnosuke Nozue[1], Yuto Nakano[1], Shoji Moriya[1], Tomoki Ariyama[1,2], Kazuma Kokuta[1,2], Suchun Xie[1], Kai Sato[1], Shusaku Sone[1], Ryohei Kamei[1], Reina Akama[1,2], Yuichiroh Matsubayashi[1,2], Keisuke Sakaguchi[1,2]

[1]Tohoku University, [2]RIKEN

{nozue.shinnosuke.q5, nakano.yuto.t2, shoji.moriya.q7, tomoki.ariyama.s3, kokuta.kazuma.r3, xie.suchun.p7, kai.satou.r8, sone.shusaku.r8, ryohei.kamei.s4}@dc.tohoku.ac.jp, {akama, y.m, keisuke.sakaguchi}@tohoku.ac.jp

Abstract

The evolution of large language models has enabled fluent dialogue, increasing interest in the coexistence of humans and avatars. An essential aspect of achieving this coexistence involves developing sophisticated dialogue systems that can influence user behavior. In this background, we propose an effective multimodal dialogue system designed to promote consensus building with humans. Our system employs a slot-filling strategy to guide discussions and attempts to influence users with suggestions through emotional expression and intent conveyance via its avatar. These innovations have resulted in our system achieving the highest performance in a competition evaluating consensus building between humans and dialogue systems. We hope that our research will promote further discussion on the development of dialogue systems that enhance consensus building in human collaboration.

1 Introduction

The emergence of large language models, such as GPT-4 (OpenAI, 2023), has facilitated highly fluent text-based conversations. Nevertheless, in many practical situations, dialogue systems require the capability to influence users through negotiation, persuasion, and consensus building (Zhan et al., 2024). In these advanced dialogue scenarios, it is essential to impact users' cognitive and emotional responses to induce alternation in their thoughts, opinions, and behaviors, yet research on these skills remains limited (Chawla et al., 2023).

In this context, the Dialogue System Live Competition 6 (DSLC6) was organized to evaluate the current technological limitations and identify necessary components for creating systems capable of consensus building (Higashinaka et al., 2024). The competition set up a dialogue scenario wherein the system had to negotiate with the user in the context of conflicting goals: a system and a user jointly plan a party, but the system desires a grand party while the user actually prefers a modest one. The systems are required to take the user's voice utterance as input and respond with avatar movements and synthesized voice. The demonstrations are evaluated based on three criteria to ensure they are contextually appropriate: relevance of utterance content, suitability of gestures and facial expression, and appropriateness of pause and voice modulation.

This paper presents our system[1] submitted to DSLC6 (Nakano et al., 2023), which aims to build consensus with a user by guiding discussion with subdivided topics and conveying emotions and positions through its avatar. Our focus in designing the consensus building process is on agenda management and the clear communication of emotions and positions. In order to facilitate discussion between speakers with different objectives, our system employs a strategy of dividing a large topic into subtopics and guiding the user step by step. Specifically, to manage discussion flow, we predetermine a list of subtopics as "blanked slots" in the GPT-based system's prompt. Once a particular issue is resolved, the system introduces the next unresolved subtopic. Within each subtopic, negotiations progress through a sequence of proposals and responses (including acceptance, rejection, and counter-proposals) (Maynard, 2010). Throughout this process, the system is designed to articulate its intentions while responding. One crucial element in this effort is the expression of emotions, which is considered to influence the future actions of others in negotiations (Morris and Keltner, 2000; de Melo et al., 2011; Melo et al., 2012). When controlling the avatar, we incorporate facial expressions, voice modulation, and body poses to express emotions. Through these innovations, our system demonstrated the best performance in DSLC6, and

[1]https://github.com/cl-tohoku/hagi-bot

Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 669–673
September 18–20, 2024.

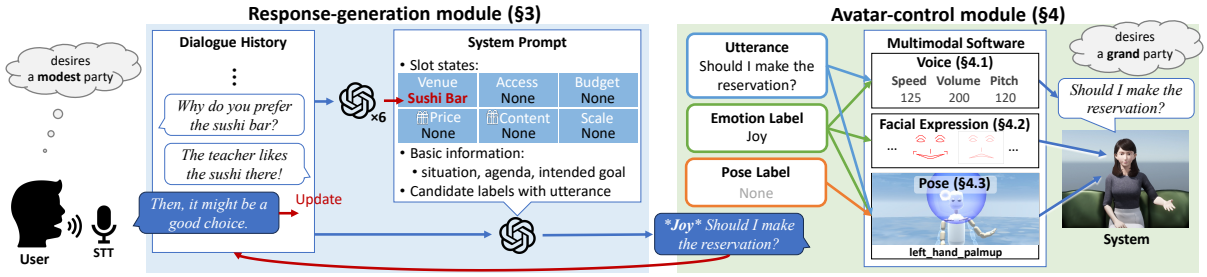


Figure 1: System overview. The user's utterance is added to the dialogue history, and GPT-4 generates the system response based on the history and the system prompt. In the avatar-control module, the system's voice, facial expressions, and poses are controlled according to the generated utterances and emotion/pose labels. Slots are updated in parallel with other processes after the generated response is added to the dialogue history.

feedback from the evaluators confirmed their effectiveness.

2 System Overview

Our system consists of two modules: the **response-generation module** and the **avatar-control module** (Figure 1). The response-generation module (Section 3) generates the system's utterance and emotion/pose labels in response to the user's utterance recognized by a speech-to-text (STT) function. The generated responses are forwarded to the avatar-control module (Section 4) that controls the voice, facial expressions, and poses according to predefined rules based on the received labels.

3 Response-Generation Module

For the response-generation module, we employed GPT-4 to generate outputs. We provided the module with a system prompt including basic information regarding the dialogue setting and the current status of slots (Section 3.1). Subsequently, the module generated an utterance and corresponding emotion/pose labels based on the provided prompt. Additionally, we ensured smooth turn-taking through pre- and post-module processing (Section 3.2).

3.1 Prompt Engineering

In the prompt, we detailed the basic situation of the dialogue, the agenda, and the intended objective of the discussion. We also provided some example utterances to instruct the output format of emotion/pose labels. Finally, we included the slot states of predefined subtopics. In the preliminary testing, we observed that the system often readily accepted user proposals that conflicted with its designated goal. To mitigate this tendency, we modified the prompt to include specific instructions: "If opinions differ from those of the user, please engage in a discussion to make a decision while showing empathy." This adjustment encouraged the system to express opposing views when necessary.

Slot-Filling We employed slot-filling-based dialogue state management to enable our system to lead discussions. We adopted GPT-4 as a slot-filling module for each subtopic individually. Each module dynamically updated the slot for the targeted subtopic with the determined content using the dialogue history (Figure 2). This approach enabled the system to start the discussion on an unfilled subtopic and facilitated smooth transitions to the next subtopic upon reaching a conclusion.

Emotion and Pose Label Generation In dialogue systems, emotion classification is typically performed independently from response generation (Moriya et al., 2023; Yamazaki et al., 2023). Consequently, errors in the classifier can lead to inconsistencies between utterance content and the avatar's emotional expression. To address this issue, our system simultaneously generates utterance and emotion/pose labels, thereby ensuring coherence between utterance content and avatar expressions.[2] The candidate emotion labels were derived from Plutchik's basic eight emotions (Plutchik, 2001), *Joy*, *Sadness*, *Anticipation*, *Surprise*, *Anger*, *Fear*, *Disgust*, and *Trust*, as utilized in the Japanese WRIME dataset (Kajiwara et al., 2021), with an additional *Neutral* label. Furthermore, we incorporated four pose labels—*Bowing*, *Nodding*, *Shaking head*, and *Pondering*—to represent the system's intention and prevent misunderstanding.

[2]This was achieved by in-context learning using the prompt by providing an utterance example, such as "**Pondering** Hmm, I see your point... **Joy** In that case, it should be fine!"

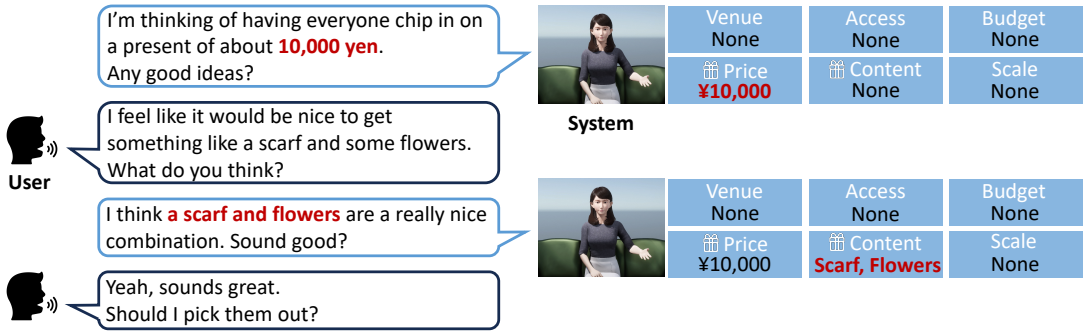


Figure 2: Example of dialogue and filling slots.

3.2 Natural Turn-Taking

It is crucial to avoid responses that are either too quick, which may interrupt the user's speech, or too slow, which may lead the user to perceive the system as unresponsive (Skantze, 2021). Our system determines the end of the user's turn based on the duration of silence. Through qualitative evaluations, a duration of 1.6 sec. was deemed suitable and adopted as the threshold. Additionally, in preliminary experiments, it was found that GPT-4 can take more than 10 sec. to generate a response when multiple sentences are involved. To address this issue, we employ the streaming mode. Specifically, we generated one character at a time while retaining and sending them to the text-to-speech (TTS) function upon generating a delimiter. This approach effectively reduced the response time from approximately 13 sec. to 3 sec.

4 Avatar-Control Module

In the avatar-control module, the system's emotions and position are effectively conveyed by reflecting rules corresponding to emotion/pose labels or utterance content in the avatar's voice, facial expressions, and poses. The avatar utilizes the resources provided by DSLC6 (Higashinaka et al., 2024).

4.1 Effect of Emotion Expressions

Emotion expressions serve as media to communicate the sender's internal states, enabling the receiver to infer these states from the sender's emotional displays (de Melo et al., 2023, 2011, 2014; Gratch and de Melo, 2019). This process significantly influences the thoughts and behaviors of the receiver, proving effective in various stages of negotiation, including trust development and consensus building. (de Melo et al., 2023; Morris and Keltner, 2000). Furthermore, it has been found that the impact of such emotional conveying processes is also effective in human-machine interactions (Melo et al., 2012; de Melo et al., 2011). In this paper, we aim to achieve effective negotiation and consensus building by using competition-regulated tools to express emotions through voice, facial expressions, and poses (posture and gesture).

		speed	volume	pitch
Emotion Label	Joy	125	200	120
	Anticipation	120	150	117
	Sadness	120	100	105
	Surprise	125	250	115
	Anger	120	230	100
	Fear	125	250	115
	Disgust	120	100	95
	Trust	120	100	117
	Neutral	120	100	115
End of Sentence	"!"	+0	+50	+0
	"..."	-30	-50	-10

Table 1: Voice parameters. Our system utilizes AmazonPollyServer for its TTS function.

4.2 Voice

The avatar's voice was modulated by predefined parameters such as speed, volume, and pitch corresponding to the emotion. This concept is adopted from the work of Togo et al. (2022). As a voice control guideline, we employed the two-dimensional arrangement of emotions in Russell's Circumplex Model of Affect (Russell, 1980). We mapped the arousal-sleep dimension to the speed and volume and the pleasure-displeasure dimension to the pitch. Furthermore, when the utterance ends with "!" or "...," we adjusted the parameters based on the polarity and intensity of the emotion. Table 1 shows the specific parameter values defined above. Additionally, in conversations, pauses express emotions and give listeners time to understand nuances (Nakamura, 2009). Accordingly, we introduced short

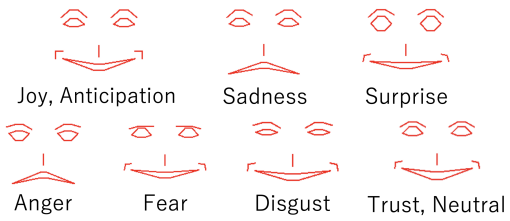
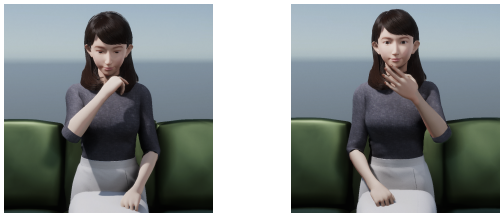


Figure 3: Facial expression set based on emotion labels. The expressions are configured using the JointMapperPlusUltraSuperFace preset.



(a) Pondering (b) Surprised

Figure 4: Examples of pose control.

silences after punctuation, like periods, and longer silences after ellipses.

4.3 Facial Expressions

We predefined the avatar's facial expressions for each of the nine emotion labels (Figure 3). The corresponding preset is referenced from the received emotion label and reflected in the avatar.

4.4 Control of Poses

The poses are determined by both the labels and the utterance content. We implemented fundamental behaviors such as nodding, but this paper focuses on the most crucial aspect: expressing emotions and intentions. To represent pondering, we utilized a hand-on-chin gesture, slightly lowering the face in response to the pose label *Pondering* or utterances such as "Hmm." or "..." (Figure 4a). Similarly, placing a hand over the mouth represents a gesture of surprise, triggered by the emotion label *Surprise* or utterances indicating astonishment such as "Really?" (Figure 4b). Based on the Aoba_v3 bot (Moriya et al., 2023), which aimed to behave like a human, we represented emotions such as *Sadness*, *Joy*, and *Anticipation*. For example, when emotions such as *Joy* or *Anticipation* were expressed, we defined the pose of lightly reaching out to the interlocutor. Conversely, when a label indicated *Sadness*, we expressed it by bowing the avatar's head.

	Content	Expressions	Voice	Average
Ours	**3.93**	**3.41**	**3.01**	**3.45**
System A	3.32	3.32	2.94	3.19
System B	3.58	3.27	2.66	3.17

Table 2: Results of the final round in DSLC6. Scores range from 1 to 5. Three evaluation criteria are utterance content, gestures and facial expressions, and pause and voice modulation.

Evaluators' comments and scores
I sensed the system's intention. The system progressed to the next, and the dialogue pace was good. [Content, Expressions, Voice] = [5, 5, 5]
The conversations, gestures and pauses were very human. [Content, Expressions, Voice] = [4, 4, 4]
I appreciated that the system shared its opinions and listened to mine. However, it spoke too quickly and seemed to talk quite a bit. [Content, Expressions, Voice] = [3, 3, 3]
Her expressions seem exaggerated, which can be tiring. [Content, Expressions, Voice] = [4, 3, 3]

Table 3: Evaluators' comments with scores.

5 Human Evaluation

Three top-performing systems, selected from an initial pool of ten systems in the preliminary round, participated in the final round of DSLC6. At the final round, each system engaged in a five-minute conversation in Japanese with a user who is a humanities researcher. The interactions were evaluated by an audience of 80 attendees. Each system was evaluated twice. The assessments were on the basis of contextual appropriateness, considering the following three criteria: (i) Content: relevance of utterance content, (ii) Expressions: suitability of gestures and facial expressions, and (iii) Voice: appropriateness of pause and voice modulation. Using a 5-point Likert scale, the systems' performances were ranked based on the average scores of all criteria.

The results in Table 2 indicate our system achieved the best performance across all criteria. Table 3 contains some of the feedback from the preliminary evaluators. According to the positive feedback, the system was effective in leading discussions and naturally conveying intentions through the avatar. However, some users provided negative feedback regarding the non-verbal aspects, such as rapid speaking and excessive movement, indicating that there is still room for improvement.

6 Conclusion

This paper presented the top-performing system in a dialogue competition focused on the consensus building process between systems and humans. Our proposed system incorporated two strategies for smooth consensus building: topic control through a slot-filling approach and conveying intent through emotional expression via an avatar.

Our system can be adapted for various situations. The only task-specific components are the prompts and slots. The prompts include a general instruction format necessary for the consensus building process, such as context, personas, subtopics, and the system's objectives, thereby reducing the effort required for prompt engineering when applied to different tasks. Each topic slot is managed by a separate model, allowing easy instantiation and repurposing once the subtopics are defined.

We hope that this system will contribute insights to research on dialogue systems to build consensus with humans in the context of conflicting goals.

Acknowledgements

We would like to express our gratitude to the organizers of DSLC6 for providing the software necessary for the development of the multimodal dialogue system, as well as to Prof. Kentaro Inui and Ms. Fuka Narita of the Tohoku NLP Group for their cooperation in the development of this system. This research was supported in part by JSPS KAKENHI Grants JP22K17943 and JP21K21343.

References

Kushal Chawla et al. 2023. Social influence dialogue systems: A survey of datasets and models for social influence tasks. In *EACL*, pages 750–766.

Celso M. de Melo et al. 2011. The effect of expression of anger and happiness in computer agents on negotiations with humans. In *The 10th International Conference on Autonomous Agents and Multiagent Systems - Volume 3*, AAMAS '11, page 937–944.

Celso M. de Melo et al. 2014. Reading people's minds from emotion expressions in interdependent decision making. *Journal of personality and social psychology*, 106 1:73–88.

Celso M. de Melo et al. 2023. Social functions of machine emotional expressions. *Proceedings of the IEEE*, 111(10):1382–1397.

Jonathan Gratch and Celso M. de Melo. 2019. *Inferring Intentions from Emotion Expressions in Social Decision Making*, pages 141–160. Springer International Publishing, Cham.

Ryuichiro Higashinaka et al. 2024. Dialogue system live competition goes multimodal: Analyzing the effects of multimodal information in situated dialogue systems. In *IWSDS*.

Tomoyuki Kajiwara et al. 2021. WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations. In *NAACL: Human Language Technologies*, pages 2095–2104.

Douglas W. Maynard. 2010. Demur, defer, and deter: Concrete, actual practices for negotiation in interaction. *Negotiation Journal*, 26(2):125–143.

Celso Melo, Peter Carnevale, and Jonathan Gratch. 2012. The impact of emotion displays in embodied agents on emergence of cooperation with people. *PRESENCE: Teleoperators and Virtual Environments*, 20:449–465.

Shoji Moriya, Daiki Shiono, et al. 2023. Aoba_v3 bot: a multimodal chatbot system combining rules and various response generation models. *Advanced Robotics*, 37(21):1392–1405.

Michael W. Morris and Dacher Keltner. 2000. How emotions work: The social functions of emotional expression in negotiations. *Research in Organizational Behavior*, 22:1–50.

Toshie Nakamura. 2009. Psychological study of 'ma' (a synonym of 'pause') in communication. *Journal of the Phonetic Society of Japan*, 13(1):40–52.

Yuto Nakano, Shinnosuke Nozue, et al. 2023. Hagi bot: A multimodal dialogue system for smooth discussion with human-like behavior and llm-based dialogue state tracking. *In JSAI SIG-SLUD*, 99:102–107.

OpenAI. 2023. GPT-4 technical report.

Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. In *American Scientist, Vol. 89, No. 4 (JULY-AUGUST)*, pages 344–350.

James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language*, 67:101178.

Eitetsu Togo et al. 2022. A multimodal dialogue system with control based on emotion classification of system utterances. *In JSAI SIG-SLUD*, 96:206–210.

Takato Yamazaki et al. 2023. An open-domain avatar chatbot by exploiting a large language model. In *SIGDIAL*, pages 428–432.

Haolan Zhan et al. 2024. Let's negotiate! a survey of negotiation dialogue systems. *Preprint*, arXiv:2402.01097.