Jingjing Jiang

Nagoya University
Nagoya, Aichi
Japan
jiang.jingjing.k6
@s.mail.nagoya-u.ac.jp

1 Research interests

The ultimate goal of my research is to develop human-like chat-oriented dialogue systems that establish long-term connections with users by satisfying their need for communication and affection. To achieve this, dialogue systems need to accurately understand the user's mental state and generate appropriate responses. However, most of the current dialogue systems interact with users relying solely on text or speech, which is insufficient for estimating the user's mental state.

Therefore, to enable dialogue systems to accurately capture the user's mental state, we focus on two areas: **construction and utilization of multimodal datasets** in human communication and **real-time multimodal affective computing**.

1.1 Construction and utilization of multimodal datasets

The primary interests of our group include processing and analyzing multimodal information in dialogue.

In face-to-face human communication, we inherently use verbal information, such as language and speech, and non-verbal information, including facial expressions, gaze, and gestures, to convey our intentions and ideas. The combination of these multiple modes of information is termed multimodal information. By comprehensively processing and analyzing multimodal information, human intentions and emotional states can be accurately comprehended during communication.

Several multimodal dialogue datasets have been constructed to date. For example, IEMOCAP (Busso et al., 2008) is a script-based human-human dialogue dataset containing speech, video, and facial motion capture. RECOLA (Ringeval et al., 2013) is a dataset that includes audio, visual, and physiological recordings regarding a collaborative dialogue task. Hazumi (Komatani and Okada, 2021) is a human-agent multimodal dialogue corpus containing audio, visual, and physiological data. However, these datasets lack comprehensive multimodal information during dialogue, which limits the scope and depth of research that can be conducted.

Consequently, our research group has collected a Japanese human-human dialogue dataset comprising a wide range of modalities, including speech, video, physiological signals, gaze, and body movement, as well as subjective evaluations of the interlocutor's emotional valences. All data are synchronized with timestamps. Furthermore, we analyzed the relationships between various physiological signals and subjective evaluations (Jiang et al., 2024). In future work, we plan to extend the analysis beyond physiological signals to understand and model various phenomena that occur in natural human communication.

1.2 Real-time multimodal affective computing

We also focus on developing a model that can detect or predict the interlocutors' emotional state in real-time for spoken dialogue systems.

In the field of affective computing, sentiment analysis and emotion recognition are combined to detect and model human emotional behavior. Most multimodal affective computing approaches in dialogue use text, speech, and video to identify the emotional state of the interlocutor. However, relying solely on this observable information makes it challenging to correctly recognize subtle emotional changes when the interlocutors do not explicitly express their emotions.

To address this limitation, extensive research (Katada et al., 2020; Keren et al., 2017) has been conducted on identifying an interlocutor's emotional state using physiological signals, such as heart rate and electrodermal activity. However, these studies typically conduct affective computing at the utterance or sentence level and do not consider the real-time nature of the user's mental state, which is essential for dialogue systems.

Therefore, we seek an effective method for data processing and multimodal feature fusion to construct a real-time emotion estimation model that leverages audiovisual information and physiological signals. Such a model needs to identify patterns that are generalizable across users. However, emotional expression varies significantly among individuals and is influenced by factors such as culture and personality. These individual differences are critical in affective computing and cannot be disregarded. Future studies should also consider models that adapt to individual users while preserving generalizability across diverse users.

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 60–61
September 16–17, 2024. ©2024 Association for Computational Linguistics

2 Spoken dialogue system (SDS) research

Recently, large language models (LLMs) have significantly improved the understanding of user input, retaining long-term contextual information and generating more fluent responses. They have also demonstrated the capability to recognize human emotions (Tak and Gratch, 2023). Building on these developments, multimodal LLMs (MM-LLMs) that process multimodal inputs and outputs across various modalities such as text, audio, and video, have undergone widespread development. Most current MM-LLMs are primarily employed to assist users with specific tasks such as image editing (Zhang et al., 2024); their potential to understand human emotions is also promising.

In the future, MM-LLMs capable of establishing long-term connections with users may be developed, which can access users' intentions and emotional states and respond accordingly. Furthermore, future dialogue systems might be widely integrated into humanoid robots. This integration could significantly enrich user-system interactions, for instance, by incorporating haptics, which has the potential to enhance user immersion and engagement during interactions with dialogue systems (Minato et al., 2023).

3 Suggested topics for discussion

I would like to discuss the following topics:

- What are the possible applications of a dialogue system that can acquire physiological signals?
- What would a dialogue system that builds long-term relationships with humans look like? What kind of appearance, interaction interface, and qualities would it possess?
- Assuming the existence of a dialogue system that can establish a long-term connection with humans, which would be more desirable: a dialogue system that has its personality and emotions, including the possibility of getting angry, or a dialogue system that solely focuses on satisfying all your needs without expressing its own emotions?

Acknowledgments

This work was supported by JST Moonshot R&D Grant Number JPMJMS2011.

References

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation* 42:335–359.

Jingjing Jiang, Ao Guo, and Ryuichiro Higashinaka. 2024. Estimating the Emotional Valence of Interlocutors Using Heterogeneous Sensors in Human-Human Dialogue. In *Proc. SIGDIAL*.

Shun Katada, Shogo Okada, Yuki Hirano, and Kazunori Komatani. 2020. Is She Truly Enjoying the Conversation? Analysis of physiological signals toward adaptive dialogue systems. In *Proc. ICMI*. page 315–323.

Gil Keren, Tobias Kirschstein, Erik Marchi, Fabien Ringeval, and Björn Schuller. 2017. End-to-end learning for dimensional emotion recognition from physiological signals. In *Proc. ICME*. pages 985–990.

Kazunori Komatani and Shogo Okada. 2021. Multimodal human-agent dialogue corpus with annotations at utterance and dialogue levels. In *Proc. ACII*. pages 1–8.

Takashi Minato, Ryuichiro Higashinaka, Kurima Sakai, Tomo Funayama, Hiromitsu Nishizaki, and Takayuki Nagai. 2023. Design of a competition specifically for spoken dialogue with a humanoid robot. *Advanced Robotics* 37:1349–1363.

Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *Proc. 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. pages 1–8.

Ala N. Tak and Jonathan Gratch. 2023. Is GPT a Computational Model of Emotion? In *Prco. ACII*. pages 1–8.

Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. 2024. MM-LLMs: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601* .

Biographical sketch

Jingjing Jiang is a master's student at the Graduate School of Informatics, Nagoya University, under the supervision of Prof. Ryuichiro Higashinaka. She is interested in dialogue systems, multimodal interaction, and affective computing. During her PhD course, she aspires to broaden her perspectives by collaborating with other research institutions on multimodal interaction in dialogue.