Jingjing Jiang

Nagoya University
Nagoya, Aichi
Japan

jiang.jingjing.k6
@s.mail.nagoya-u.ac.jp

1 Research interests

My ultimate goal is to develop human-like dialogue systems capable of expressing emotions through multimodal communication, incorporating not only spoken responses but also non-verbal cues such as facial expressions and physiological signal visualizations.

My current research interests lie in the areas of **multimodal emotion recognition** and **non-verbal cue generation**, with a focus on emotion-enriched human-computer interaction.

1.1 Multimodal emotion recognition

My current research involves multimodal emotion recognition in face-to-face conversations.

In face-to-face human interactions, people convey emotions and intentions through a combination of verbal cues and non-verbal cues, including facial expressions, gaze, gestures, and physiological responses. Capturing these multimodal cues is essential for building effective emotion recognition systems, and the development of such systems demands high-quality multimodal dialogue datasets that reflect the complexity of natural human interaction. However, existing multimodal datasets such as IEMOCAP (Busso et al., 2008), Hazumi (Komatani and Okada, 2021), and EEVR (Singh et al., 2024) present certain limitations, such as limited sensor modalities, constrained diversity of emotional expressions, and the absence of spontaneous dialogue scenarios that reflect everyday conversations. To address these limitations, my colleagues and I have constructed a Japanese multimodal dialogue dataset that captures a wide range of modalities in dyadic face-to-face conversations. This dataset includes synchronized recordings of textual transcriptions, speech, facial video, physiological signals, body movement, and self-assessments of emotional valence (Jiang et al., 2024).

Leveraging this dataset, I developed a multimodal emotional valence recognition model that performs binary valence classification (high vs. low) on 15-second conversational segments (Jiang et al., 2025). The model integrates modality-specific representations: textual embeddings extracted using Japanese BERT[1], speech representations obtained through Japanese HuBERT (Sawada et al., 2024), and physiological signals including electrodermal activity (EDA), blood volume pulse (BVP), photoplethysmography (PPG), and pupil diameter, which are encoded into time-series embeddings using the self-supervised Ts2Vec encoder (Yue et al., 2022). Experimental results demonstrate that the multimodal fusion approach achieves superior classification performance compared with single-modality baselines.

My future research will leverage additional modalities, such as visual and motion information, to enhance the performance of emotional valence recognition.

1.2 Non-verbal cue generation

Additionally, to enable more comprehensive and emotionally expressive human-computer interactions, I am working on the generation of non-verbal cues, including observable non-verbal behaviors and internal physiological signals for embodied conversational agents (ECAs).

Observable non-verbal behaviors, including facial expressions, gestures, and gaze, play crucial roles in natural human communication. Incorporating these cues into the ECA design would enhance users' perception of social presence and improve human-computer interaction quality. Current methodologies include two primary approaches: rule-based systems that use hand-crafted rules to generate specific behaviors (Cassell et al., 1994) and data-driven approaches that utilize machine learning techniques to learn expressive behaviors from multimodal data (Admoni and Scassellati, 2014). Recent advances have achieved remarkable progress in audio-driven talking face generation, particularly in producing realistic lip synchronization with precise temporal accuracy (Prajwal et al., 2020; Xing et al., 2023; Kim et al., 2025).

Physiological signals, including heart rate (HR), breathing rate (BR), and EDA, provide rich insights into humans' internal and emotional states. Recent studies have demonstrated that visualizing users' physiological

[1] https://github.com/cl-tohoku/bert-japanese

The 21th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 15–17
Aug 28–29, 2025.

signals through virtual agents, such as with dynamic 3D heart icons that change size and pulse rate based on HR and animated breath bubble icons representing BR cycles, enhances expressiveness and fosters empathy and communication (Lee et al., 2022; Sasikumar et al., 2024). Building upon these findings, I am particularly interested in the generation and visualization of physiological signals that reflect emotional states for ECAs, seeking to develop systems that can generate appropriate physiological responses to create more human-like emotional interactions.

My future research aims to explore techniques for achieving natural and empathetic generation of nonverbal cues in ECAs, including emotion-to-behavior mapping algorithms and real-time visualization algorithms.

2 Spoken dialogue system (SDS) research

Regarding the future vision for SDSs, recent breakthroughs in SDS research have demonstrated the potential for real-time, simultaneous bidirectional communication that more closely mimics natural human conversation patterns (Défossez et al., 2024; Ohashi et al., 2025). These advances represent a significant leap from traditional turn-based dialogue systems by enabling continuous, overlapping speech interactions with minimal latency.

Looking toward the near future, full-duplex SDSs are expected to expand into multimodal interaction. Future systems will likely integrate multiple communication channels, including speech, visual cues, and physiological signals. By combining these modalities, next-generation SDSs will be capable of delivering immersive and emotionally expressive conversational experiences. Ultimately, these systems will not only interpret users' multimodal inputs but also generate appropriate responses across multiple modalities simultaneously.

3 Suggested topics for discussion

- What novel architectural approaches show the most promise for developing more efficient and scalable multimodal encoders and decoders?
- How can we establish standardized protocols for multimodal dialogue data collection, annotation, and sharing to ensure reproducibility and comparability across different datasets?
- Compared to industry, what unique advantages does academia have in SDS research, and on which aspects of research should academia focus more?

Acknowledgments

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

References

Henny Admoni and Brian Scassellati. 2014. Data-Driven Model of Nonverbal Behavior for Socially Assistive Human-Robot Interactions. In *Proc. ICMI*. page 196–199.

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.

Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. 1994. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proc. SIGGRAPH*. page 413–420.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037* .

Jingjing Jiang, Ao Guo, and Ryuichiro Higashinaka. 2024. Estimating the Emotional Valence of Interlocutors Using Heterogeneous Sensors in Human-Human Dialogue. In *Proc. SIGDIAL*. pages 718–727.

Jingjing Jiang, Ao Guo, and Ryuichiro Higashinaka. 2025. Integrating Physiological, Speech, and Textual Information Toward Real-Time Recognition of Emotional Valence in Dialogue. In *Proc. SIGDIAL*.

Jisoo Kim, Jungbin Cho, Joonho Park, Soonmin Hwang, Da Eun Kim, Geon Kim, and Youngjae Yu. 2025. DEEPTalk: Dynamic Emotion Embedding for Probabilistic Speech-Driven 3D Face Animation. In *Proc. AAAI*. volume 39, pages 4275–4283.

Kazunori Komatani and Shogo Okada. 2021. Multimodal Human-Agent Dialogue Corpus with Annotations at Utterance and Dialogue Levels. In *Proc. ACII*. pages 1–8.

Sueyoon Lee, Abdallah El Ali, Maarten Wijntjes, and Pablo Cesar. 2022. Understanding and designing avatar biosignal visualizations for social virtual reality entertainment. In *Proc. CHI*. pages 1–15.

Atsumoto Ohashi, Shinya Iizuka, Jingjing Jiang, and Ryuichiro Higashinaka. 2025. Towards a Japanese

Full-duplex Spoken Dialogue System. In *Proc. INTERSPEECH*.

K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. 2020. A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild. In *Proc. MM*. page 484–492.

Prasanth Sasikumar, Ryo Hajika, Kunal Gupta, Tamil Selvan Gunasekaran, Yun Suen Pai, Huidong Bai, Suranga Nanayakkara, and Mark Billinghurst. 2024. A User Study on Sharing Physiological Cues in VR Assembly Tasks. In *Proc. IEEE VR*. pages 765–773.

Kei Sawada, Tianyu Zhao, Makoto Shing, Kentaro Mitsui, Akio Kaga, Yukiya Hono, Toshiaki Wakatsuki, and Koh Mitsuda. 2024. Release of Pre-Trained Models for the Japanese language. In *Proc. LREC-COLING*. pages 13898–13905.

Pragya Singh, Ritvik Budhiraja, Ankush Gupta, Anshul Goswami, Mohan Kumar, and Pushpendra Singh. 2024. EEVR: A Dataset of Paired Physiological Signals and Textual Descriptions for Joint Emotion Representation Learning. In *Proc. NeurIPS*. volume 37, pages 15765–15778.

Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. 2023. Codetalker: Speech-Driven 3D Facial Animation with Discrete Motion Prior. In *Proc. CVPR*. pages 12780–12790.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. 2022. Ts2vec: Towards Universal Representation of Time Series. In *Proc. AAAI*. volume 36, pages 8980–8987.

4 Biographical sketch

Jingjing Jiang is a first-year PhD 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.