Estimating the Emotional Valence of Interlocutors Using Heterogeneous Sensors in Human-Human Dialogue

Jingjing Jiang, Ao Guo, Ryuichiro Higashinaka
Graduate School of Informatics, Nagoya University, Japan
jiang.jingjing.k6@s.mail.nagoya-u.ac.jp
guo.ao.i6@f.mail.nagoya-u.ac.jp
higashinaka@i.nagoya-u.ac.jp

Abstract

Dialogue systems need to accurately understand the user's mental state to generate appropriate responses, but accurately discerning such states solely from text or speech can be challenging. To determine which information is necessary, we first collected human-human multimodal dialogues using heterogeneous sensors, resulting in a dataset containing various types of information including speech, video, physiological signals, gaze, and body movement. Additionally, for each time step of the data, users provided subjective evaluations of their emotional valence while reviewing the dialogue videos. Using this dataset and focusing on physiological signals, we analyzed the relationship between the signals and the subjective evaluations through Granger causality analysis. We also investigated how sensor signals differ depending on the polarity of the valence. Our findings revealed several physiological signals related to the user's emotional valence.

1 Introduction

Most current user-adaptive dialogue systems rely on text or speech to estimate the user's state and generate appropriate responses. However, the user's state that can be inferred solely from text or speech is limited. Consequently, there has been active research on estimating the user's state from multimodal data, particularly focusing on user emotions and engagement through the analysis of facial expressions, gestures, and gaze (Mittal et al., 2020; Yu et al., 2015). In recent years, the application of physiological signals in dialogue systems has also gained popularity. For example, studies have been conducted to estimate a user's sentiment and emotions from physiological signals (Katada et al., 2020, 2023; Saffaryazdi et al., 2022). However, these studies have typically utilized a limited range of sensors and have not dealt with the real-time nature of the user's mental state, which is essential for dialogue systems.

Therefore, in this study, we used heterogeneous sensors to collect a variety of data during human-human dialogues, including speech, video, physiological signals, gaze information, and motion information. After each dialogue, for each time step of the data, interlocutors conducted immediate subjective evaluations of their emotional valence while watching recordings of the dialogues. As an analysis, we used Granger causality analysis to investigate the relationship between the information obtained from the heterogeneous sensors and the subjective evaluation annotations. We also conducted a statistical test to examine how sensor signals differ depending on the polarity of valence. Various sensor signals were collected, but in this paper, we focus on physiological signals, as these data are believed to be closely related to mental states (Russell, 2003). Our key contributions in this work are as follows:

- We created a Japanese multimodal human-human dialogue dataset using heterogeneous sensors, including various types of sensor signals and subjective evaluations of the interlocutors' emotional valence.
- We analyzed the relationship between various sensor signals and subjective evaluations and investigated how sensor signals vary with the polarity of emotional valence.
- Our findings revealed several physiological signals associated with emotional valence.

2 Related Work

Several multimodal dialogue corpora have been constructed that include information such as the interlocutor's movements and gaze in addition to speech. For example, HUMAINE (Douglas-Cowie et al., 2011) is a multimodal dialogue corpus covering various topics aimed at eliciting user emotions. The IEMOCAP dataset (Busso et al., 2008)

Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 718–727
September 18–20, 2024. ©2024 Association for Computational Linguistics

Data version	Data2312	Data2402
Collection time	December 2023	February 2024
Overview	Multimodal dialogues between two human interlocutors	
Dialogue topic	Chit-chat, Narrative, Discussion	
Dialogue duration	10 min (Average of 180 utterances) per dialogue	
No. of dialogues	27	33
Total utterances	4854	5956
Interlocutors	18 (9 male, 9 female)	22 (12 male, 10 female)
	Aged 20 to 50	Aged 20 to 60
	9 groups (3 groups of male pairs, 3 groups of female pairs, 3 groups of both sexes)	11 groups (4 groups of male pairs, 3 groups of female pairs, 4 groups of both sexes)
Questionnaires	Pre-experiment (Demographic information and personality traits scored on 7-point scales: 14 items) Post-dialogue (Impressions of the dialogue scored on 7-point scales: 24 items) Follow-up (Impressions of the experiment through free-form: 3 items)	
Annotations	Subjective evaluations of the interlocutor's emotional valence at each time step of dialogue (Continuous values of 0 to 10 represent negative to positive emotional valence)	
Language	Japanese	

Table 1: Summary of collected dataset.

is a script-based human-human dialogue dataset containing speech, video, and facial information. SEMAINE (McKeown et al., 2011) is a corpus containing dialogues between computer graphics (CG) agents with different personalities and human subjects. D64 (Oertel et al., 2013) is a multi-party dialogue corpus designed to capture the natural reactions and emotions of the interlocutors.

The physiological signals are measured and quantified by sensors for physiological phenomena (such as heartbeat, brain waves, pulse, respiration, and perspiration) and can deal with the real-time state of the interlocutor. Several multimodal dialogue corpora have been constructed that include the physiological signals of the interlocutor in a dialogue. For example, RECOLA (Ringeval et al., 2013) is a human dialogue dataset that includes physiological signals during a collaborative dialogue task. Electrocardiogram (ECG) and electrodermal activity (EDA) are utilized as physiological signals in RECOLA. The PEGCONV dataset (Saffaryazdi et al., 2022) comprises discussion dialogues and includes galvanic skin response (GSR) and photoplethysmography (PPG) as physiological signals during the dialogue. Hazumi (Komatani and Okada, 2021) is a multimodal dialogue corpus containing dialogues between a human and a CG agent. The physiological signals include EDA, blood volume pulse (BVP), skin temperature (TEMP), and heart rate (HR) data (Katada et al., 2023).

Although several corpora have been constructed in this way, none of the corpora contain data that comprehensively includes movement, gaze, and a variety of physiological signals. Moreover, to the best of our knowledge, there has been no research on estimating the real-time user state required by dialogue systems from sensor signals in dialogues.

3 Data Collection

The data were collected in two periods, with the first beginning in December 2023 and the second in February 2024. To distinguish the two datasets, we used the year and month of data collection for naming: "Data2312" for the data collected in December 2023 and "Data2402" for the data collected in February 2024.

In these two sets of data collection experiments, a total of 40 interlocutors (21 male, 19 female), all native Japanese speakers, participated. They were recruited from the general public by a recruiting agency, and each participated in only one data collection experiment. Two interlocutors were paired into one group and engaged in 10-minute dialogues on three different topics: "Chit-chat", "Narrative", and "Discussion". Immediately after each dialogue, interlocutors annotated their subjective evaluations related to emotional valence while watching the recordings of the dialogue. A detailed summary of the collected dataset is provided in Table 1.

Both data collection experiments were conducted in the same sequence: pre-experiment questionnaire administration, sensor placement and attachment, dialogue and annotation conduction, sensor removal, and follow-up questionnaire administration. The dialogue and annotation conduction process was repeated three times for the three topics. For each topic, the following sequence was repeated: dialogue conduction, post-dialogue questionnaire administration, and subjective evaluation. The experiments were approved by the ethics committee of our institution.

In the following subsections, we describe in detail the multimodal data and heterogeneous sensors, the three dialogue topics, the questionnaires, and the subjective evaluation of dialogues.

3.1 Multimodal Data and Heterogeneous Sensors

We used heterogeneous sensors to collect multimodal data including speech, video, physiological signals, gaze information, and motion information. The data collection environment is shown in Fig. 1.

Speech: DPA 4088 uni-directional microphones were worn on the heads of each of the two interlocutors to capture audio recordings containing a single interlocutor's voice. We used Azure Kinect's (hereafter, Kinect) built-in omni-directional microphones to collect audio recordings containing the voices of two interlocutors. For the audio recordings collected by DPA 4088, the close proximity of the interlocutors and the loudness of the other interlocutor resulted in data containing faint sounds from the other interlocutor.

Video: We used Kinect, Logicool C920 Pro HD Webcam (hereafter, Logi webcam), and GoPro Hero 10 (hereafter, GoPro) to record the interlocutors' behavior. In Data2312, two Kinects were placed between the interlocutors to record RGB and depth video of their upper bodies. A Logi webcam was positioned to the side of the interlocutors, capturing their full-body RGB video from a side view. In Data2402, to capture the full-body movements of the interlocutors instead of just the upper body, two Kinects were positioned between them to record separate full-body RGB and depth videos. Additionally, two GoPros were positioned in the same positions as the Kinects to record full-body RGB video. Because of the lack of clarity of the logi webcam, the third GoPro was placed to the side to capture two interlocutors' full-body RGB video from the side. The Kinect and Logi webcam collected AVI files, while the GoPro recorded MP4 files.

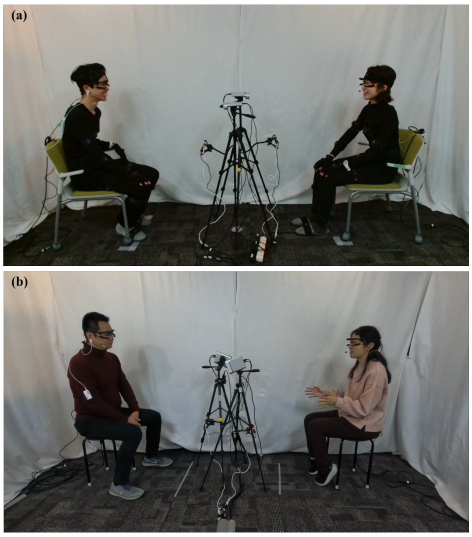

Figure 1: Data collection environments for (a) Data2312 and (b) Data2402. In (a) Data2312, two Kinects and two millimeter-wave sensors were placed between the interlocutors, and each interlocutor wore a set of wearable sensors. In (b) Data2402, two Kinects, two GoPros, and two millimeter-wave sensors were placed between the two interlocutors, and each interlocutor wore a set of wearable sensors.

Physiological Signals: We collected physiological signals during the dialogues by using the EmbracePlus[1] and the Shimmer3 GSR+ [2]. The EmbracePlus is wireless and worn like a wristwatch. We used it to collect BVP, EDA, TEMP, and wrist acceleration (ACC). The Shimmer3 GSR+ can collect EDA using an optical pulse sensing probe attached to the finger and photoplethysmography (PPG) using either an ear clip or an optical pulse probe. Due to the greater stability of data collected through the former, we chose to collect PPG using the ear clip.

Gaze: The Pupil Core eye tracker[3] (hereafter, Pupil Core) was worn by the interlocutor like glasses and was used to collect gaze data, pupil information, and eye video during the dialogue.

Motion: We positioned two IWR1443 BOOST millimeter-wave sensors between the two interlocutors to capture 3D point cloud and motion data. In Data2312, we utilized two motion capture devices called Perception Neuron 3 Body Kit[4] (hereafter, PN3) on the interlocutors' entire body to gather

[1] https://www.empatica.com/en-int/embraceplus/

[2] https://shimmersensing.com/product/shimmer3-gsr-unit/

[3] https://pupil-labs.com/products/core

[4] https://neuronmocap.com/pages/perception-neuron-3

	Device name	Type of sensor	Data
Devices for Data2312 and Data2402	Pupil Core	World camera Eye cameras	First-person video with gaze measurement Eye video, pupil and gaze information
	DPA 4088	Uni-directional mic	Audio containing one interlocutor's voice
	Shimmer3 GSR+	Ear-mounted sensor	Photoplethysmography (PPG) 120 Hz
	EmbracePlus	Wristwatch sensor	Blood volume pulse (BVP) 64 Hz Electrodermal activity (EDA) 4 Hz Skin temperature (TEMP) 1 Hz Wrist acceleration (ACC) 64 Hz
	IWR1443BOOST	Millimeter-wave sensor	3D point cloud and motion data
Devices for Data2312	Azure Kinect	RGB camera Depth camera Omni-directional mic	Front upper body RGB video Front upper body depth video Audio containing two interlocutors' voices
	Logi Webcam	RGB camera	Face-to-face full-body RGB video of two interlocutors
	Perception Neuron 3	IMU sensor	Skeleton hierarchy information and motion data
Devices for Data2402	Azure Kinect	RGB camera Depth camera Omni-directional mic	Front full-body RGB video Front full-body depth video Audio containing two interlocutors' voices
	GoPro Hero 10	RGB camera	Front full-body recording Face-to-face full-body RGB video of two interlocutors

Table 2: Multimodal data collected from devices.

skeleton hierarchy information and motion data. Due to the time-consuming process of wearing and calibrating the PN3, as well as interference from numerous devices affecting the inertial measurement unit (IMU) sensor signals, in Data2402, we decided not to use the PN3 with the intention of extracting the interlocutors' motion information from video recordings with image processing.

Two computers were used to acquire the sensor signals, which were streamed from each device. One computer served as a time server for ensuring synchronization of timestamps. Since EmbracePlus and GoPro do not support real-time streaming, the timestamps were synchronized post-data acquisition. Table 2 lists the devices and the multimodal data collected by them.

3.2 Dialogue Topics

To elicit a variety of mental states and gestures from the interlocutors, the following three topics were utilized. Example dialogues for each topic are provided in Table 3.

Chit-chat: Free dialogue with no restrictions on topics. Serving both as a means to collect dialogue in normal situations and as an icebreaker.

Narrative: The interlocutor's own special episode. Storytelling provides a wealth of gestures (Colletta et al., 2010), and we can also expect that mental states will be expressed when discussing a cherished or distressing memory.

Discussion: Topics with different opinions for or against. We can expect negative mental states to be expressed during exchanges with an interlocutor who holds an opposing view. Conversely, we can also anticipate positive mental states to be experienced when the interlocutors reach an agreement. Fifteen topics were chosen from a site[5] that deals with discussion topics, which we then translated into Japanese. Before the data collection experiment, the pairs of interlocutors were asked about their opinions in favor of or against the 15 topics and the topics that they had different opinions about were selected as discussion topics.

3.3 Questionnaires

Each interlocutor completed a pre-experiment questionnaire before the start of the experiment, a per-dialogue questionnaire immediately after each dialogue, and a follow-up questionnaire after the end of the experiment. The details of the respective questionnaires are as follows.

Pre-experiment Questionnaire: Asking about the interlocutor's demographic information and personality traits. Demographic information included gender, age, educational background, and employment status. For personality traits, we used a 10-item questionnaire from TIPI-J (Oshio et al., 2012) to measure the Big Five traits: openness to ideas/experience, conscientiousness, extraversion,

[5] https://www.procon.org/

“Chit-chat”: Open-domain dialogue with no restrictions on topics.
02F20: What did you have for breakfast?
02M30: I didn’t have breakfast.
02F20: You didn’t eat? Are you the type of person who only eats two meals a day?
02M30: One or two meals a day.
02F20: One meal a day!? Which one do you eat? Breakfast, lunch, or dinner? I’m the type of person who eats three meals a day, because I often get hungry. So I envy those who only need one meal a day.
02M30: But I may put three meals into one.

“Narrative”: Own personal story that you can’t help but want to tell others about.
04M20: I have done something that people often say is unusual.
04F30: I would like to hear about it.
04M20: People learn various sorts of things, don’t they? Like piano, swimming. I’m often told that the thing I learned was unusual.
04F30: What was it?
04M20: I used to study Kabuki.
04F30: Huh? Amazing!
04M20: That was from grade six to about high school.
04F30: You were doing it for quite a long time.

“Discussion”: Is obesity a disease?
08M50: I’d like to start by defining the term “obesity”.
08M20: I agree.
08M50: What counts as obesity?
08M20: I’m sorry if I’m being a bit light-hearted here, but in short, a fat person. I don’t mean exactly how many kilos or more he weighs, but in terms of his appearance, someone who has a bit of a belly.
08M50: Obesity is generally expressed as a certain value, such as BMI, and that value is considered to be equate to obesity. But I don’t think that certain values equal poor health or disease. What do you think about that?
08M20: I can totally understand. To be honest, I’m not sure if it’s correct or not, because it’s hard to connect a value to disease.

Table 3: Dialogue excerpts on “Chit-chat”, “Narrative”, and “Discussion”. Interlocutor IDs are five characters of the form “NNGAA”, where “NN” is the group number, “G” is the gender of the interlocutor (“M” for male, “F” for female), and “AA” is the age of the interlocutor. These excerpts were translated from the original Japanese to English by the authors.

agreeableness, and emotional stability (Goldberg, 1990).

Post-dialogue Questionnaire: Asking about the quality of the dialogue and interlocutors’ impressions on a 7-point scale. It consists of 24 items in total. For the evaluation items relating to the quality of the dialogue, we used the same six items as the questionnaire by Yamashita et al. (2023). For the evaluation items related to the impressions of the dialogue, we used 18 items from the measurement items regarding the interpersonal communication cognition of the interlocutors (Kimura et al., 2005). The items of the post-dialogue questionnaire are shown in Table 4.

Follow-up Questionnaire: A free-form questionnaire asking about the content of the dialogue that left an impression on interlocutors, any issues the interlocutors encountered during the experiment, and their opinions and impressions of the overall experimental process.

3.4 Subjective Evaluations

To obtain the interlocutors’ real-time subjective evaluation for emotional valence, each interlocutor annotated the emotional valences of the dialogue immediately after the end of each dialogue. Continuous values of 0 to 10 were used, where 0 represents very negative, 5 represents neural, and 10 represents very positive.

To reproduce the dialogue scene and to help the interlocutors recall their mental state at the time, we used video recordings of the other interlocutor as the annotation videos, rather than their own video recordings. Specifically, the interlocutor used the annotation software CARMA (Girard, 2014) and assigned a numerical value that was considered appropriate for “their mental state” at each time in the dialogue while watching the video recording. A screenshot of the CARMA interface is shown in Fig. 2. The sampling rate of annotations was 4 Hz. To familiarize the interlocutors with the use of the annotation software, a five-minute annotation exercise was conducted before the start of data collection.

Dialogue Qualities
1. The dialogue partner was approachable.
2. The dialogue partner's speech was informative.
3. The dialogue partner's speech was easy to understand.
4. I was satisfied with the dialogue.
5. I was interested in the topics discussed in this dialogue.
6. I took the initiative to speak.
Dialogue Impressions
1. I was able to coordinate the conversation well.
2. I was bored with the conversation.
3. The conversation proceeded cooperatively.
4. The conversation was harmonious.
5. The conversation was unsatisfactory.
6. The conversation was slow-paced.
7. The conversation went cold.
8. The conversation was awkward.
9. I was absorbed in the conversation.
10. The conversation lacked focus.
11. The partner and I talked with great interest.
12. The conversation was tense.
13. The conversation was friendly.
14. The conversation was lively.
15. The conversation was positive on both sides.
16. The conversation was boring.
17. The conversation was worthwhile.
18. The conversation was drawn out.

Table 4: Items of the Post-dialogue questionnaire, where "Items enquiring about the quality of the dialogue" refers to (Yamashita et al., 2023) and "Items enquiring about the impressions of the dialogue" refers to (Kimura et al., 2005). The questionnaire was translated from the original Japanese to English by the authors

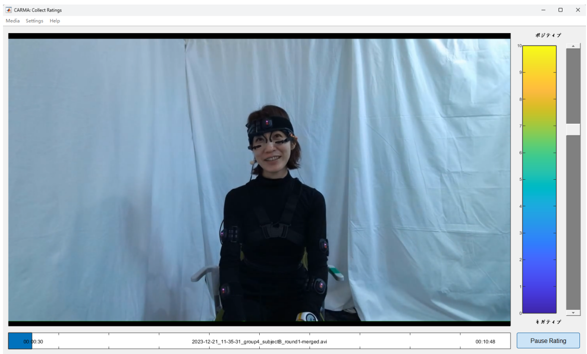

Figure 2: Screenshot of annotation interface. Emotional valence is assigned by manipulating the slide bar on the right of the screen using the controller while the interlocutor watches the other interlocutor's recording.

4 Data Analysis

Human emotional mental states, such as happiness and sadness, are formed through the brain's processing of information from three sources: 1) information from the body (e.g., HR, sweating, and other physiological states), 2) information from the external world (e.g., visual and auditory input, etc.), and 3) memories stored in the brain (Damasio, 1996; Moriguchi and Komaki, 2013). In our collection of multimodal data, the physiological signals obtained from EmbracePlus, Shimmer3 GSR+, and Pupil Core (e.g., EDA, PPG, pupil diameter) captured the interlocutors' physiological states (i.e., information from the body), while subjective evaluations annotated the emotional valence of the interlocutors.

In this study, EDA and BVP (collected from EmbracePlus), PPG (collected from Shimmer3 GSR+), and pupil diameter (collected from Pupil Core) were used as physiological signals. We first performed data preprocessing on these signals for subsequent analysis. We then performed Granger causality analysis to examine the relationship between these physiological signals and subjective evaluations of emotional valence, i.e., whether these physiological signals can be used to predict subjective evaluations. Finally, we analyzed the differences between these physiological signals under different polarities of valence.

4.1 Data Preprocessing

We extracted the physiological signals of the interlocutor during the dialogue on the basis of the start and end times of the dialogue using timestamps.

For EDA, BVP, and PPG, we used the NeuroKit2 toolbox[6] for data preprocessing (denoising, filtering) and feature extraction. We extracted the tonic skin conductance level (SCL) and phasic skin conductance response (SCR) from the EDA. SCL, also known as tonic, measures the overall conductivity of the skin, which reflects the general level of sweat gland activity. SCR, also known as phasic, measures the rapid changes in skin conductivity that occur in response to specific stimuli. The rate (the HR as measured on the basis of PPG/BVP peaks), peak (represents the highest point of PPG/BVP, used as an indicator of the intensity of a heartbeat), and the R-R intervals (RRI, which reflect the changes in time between heartbeats, i.e., HR variability) were calculated from the raw BVP and PPG data. The subjective evaluation annotations and physiological signals of an interlocutor during one minute of dialogue are shown in Fig. 3.

Pupil diameter data was sampled at a rate of 13–26 Hz, collected by Pupil Core, and the actual size of the pupil diameter (unit: mm) was derived by the device's built-in algorithm. Each timestamp has a "confidence" value indicating the quality of the measurement, and data with a confidence > 0.6

[6] https://neuropsychology.github.io/NeuroKit/

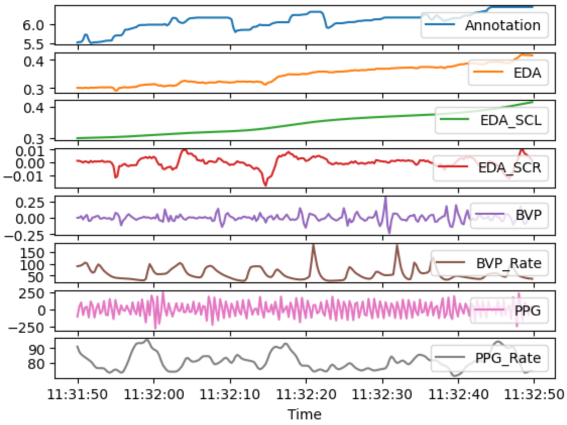

Figure 3: One-minute subjective evaluations and physiological signals of an interlocutor during dialogue. From top to bottom: Subjective evaluations (Annotation), EDA, EDA SCL, EDA SCR, BVP, BVP Rate, PPG, and PPG Rate.

is considered reliable. Since each individual has a different pupil diameter, we normalized all pupil diameter data to 0–1 using the Min-Max normalization.

4.2 Granger Causality Analysis

We computed the Granger causality analysis (Granger, 1969) to identify physiological signals or specific features of these signals that are most indicative of emotional valence changes. This method is used to evaluate the predictive utility of one variable for forecasting another and is also employed to explore the relationship between physiological signals and mental states like emotions (Gao et al., 2020). A time series X is considered to Granger-cause another time series Y if past values of X and Y predict Y significantly better than past values of Y alone (Granger, 1969).

In this study, the null hypothesis is that the physiological signals or specific features of these signals fail to Granger-cause changes in emotional valence.

The two time series used for Granger causality analysis need to be aligned and have the same sampling rate, so as the first step, we resampled all physiological signal features (SCL, SCR for EDA, Rate, Peak, RRI for PPG and BVP, and pupil diameter for left and right eyes) such that they had the same sampling rate as that of the subjective evaluations at 4 Hz, and then aligned all the data in accordance with the timestamps.

In addition, the Granger causality test assumes the series to be stationary and linearly related to make valid results. We therefore conducted the Augmented Dickey Fuller (ADF) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests for stationarity and calculated Pearson correlation for linear relationship assessment. According to the results of ADF and KPSS tests, most emotional valence data were nonstationary. Consequently, we utilized the Toda-Yamamoto causality test [7], which is an adaptation of the Granger causality test suitable for nonstationary data (Toda and Yamamoto, 1995). Regarding the Pearson correlation results, the pair of EDA SCR and emotional valence exhibited almost no linear relationship. In contrast, SCL for EDA, Rate, Peak, RRI for PPG and BVP, and pupil diameter for left and right eyes showed weak linear relationships with emotional valence. Therefore, we conducted subsequent causality tests only on the pairs involving SCL for EDA, Rate, Peak, RRI for PPG and BVP, and pupil diameter for both left and right eyes with emotional valence.

For the Granger causality test, including the Toda-Yamamoto test, the parameter "lag" represents the number of time delays used in predicting future time series data from past time series data. We set the maxlag to 8, corresponding to a maximum time delay of 2 seconds, since all data are sampled at 4 Hz. The analyses were computed for all lags up to maxlag.

4.3 Comparison of Physiological Signal Means under Positive and Negative Valence

To investigate the differences in physiological signal features depending on emotional valence polarities, we performed the Wilcoxon rank-sum test, which is a nonparametric test also known as Mann–Whitney U test, between the means of SCL and SCR of EDA, Rate, Peak, RRI of PPG and BVP, and pupil diameter for both left and right eyes. We conducted the tests with the null hypothesis that two samples of physiological signal features are drawn from the same distribution under the "positive" and "negative" emotional valence. Before conducting the test, we processed the subjective evaluation annotations; we segmented all annotations into 10-second intervals and calculated the average emotional valence value for each segment. The averages greater than 5.5 and less than 4.5 were categorized as "positive" and "negative", respectively. Those emotional values with averages in the 4.5–5.5 range were considered "neutral emo-

[7] `https://github.com/nicolarighetti/Toda-Yamamoto-Causality-Test`

tional states” and were therefore excluded from this analysis.

Given the variability of physiological signals among different interlocutors, we normalized EDA, BVP, PPG, and pupil diameter for each interlocutor before feature extraction. Specifically, we processed EDA, BVP, and PPG with the Z-score normalization, and pupil diameter using the Min-Max normalization. Then, as mentioned in Section 4.1, we extracted SCL and SCR from EDA and extracted Rate, Peak, and RRI from PPG and BVP. Finally, we performed the Wilcoxon rank-sum tests on the means of SCL and SCR of EDA, Rate, Peak, RRI of PPG and BVP, and pupil diameter for both left and right eyes, comparing them between the “positive” and “negative” emotional valence.

5 Results

In this section, we present the results of the Granger causality analysis and the differences in the means of physiological signals between “positive” and “negative” valence.

5.1 Results of Granger Causality Analysis

On the basis of the dialogue topics, we grouped the collected data into four sets: “Chit-chat” (40 dialogues), “Narratives” (40 dialogues), “Discussions” (40 dialogues), and all types of dialogues (120 dialogues). The Toda-Yamamoto Granger causality tests were conducted between the features of EDA, BVP, PPG, pupil diameter, and subjective evaluation annotations (e.g., between the RRI of the PPG and the subjective evaluation annotations) across these four datasets, with maxlag of 8. During the causality testing between EDA SCL and emotional valence, we encountered issues with idiosyncratic ranks, which prevented the construction of the model for the causality test. As a result, we excluded causality analyses between EDA SCL and emotional valence.

The proportion of dialogues featuring Granger causality is shown in Table 5. The results show that the PPG Rate has the highest potential to predict the interlocutor’s emotional valence in all dialogues. We also found that PPG Rate is the most useful feature for predicting emotional valence in all three topics :“Chit-chat”, “Narrative”, and “Discussion”.

Signal	Feature	Chit-chat	Narrative	Discussion	All
BVP	Rate	.28	.23	.08	**.19**
	Peak	.05	.10	.18	.11
	RRI	.10	.18	.08	.12
PPG	Rate	.45	.48	.30	**.41**
	Peak	.10	.13	.18	.13
	RRI	.10	.18	.08	.12
Pupil diameter	Left	.13	.10	.13	.12
	Right	.18	.15	.10	**.15**

Table 5: Results of Toda-Yamamoto causality tests (maxlag = 8). The proportion of dialogues with a significant difference of $p < 0.05$ in “Chit-chat” (40 dialogues), “Narrative” (40 dialogues), “Discussion” (40 dialogues), and “All dialogues” (120 dialogues) is shown. Bold numbers are the highest proportion for the BVP, PPG, and pupil features, and underlined numbers are the highest proportion for each topic of dialogue.

5.2 Results of the Differences between Physiological Signal Means under Different Valence Polarities

We performed the Wilcoxon rank-sum tests on SCL, SCR of EDA, Rate, Peak, RRI of PPG and BVP, and pupil diameter for left and right eyes under “positive” and “negative” emotional valence. Note that the ratio of sample size between “positive” and “negative” is around 6:1.

The mean, standard deviation, and results of the Wilcoxon rank-sum test are shown in Table 6. The results indicate that the means of EDA SCR, BVP Rate, and BVP RRI were significantly different ($p < 0.05$) between the “positive” and “negative” valence. Specifically, our experimental results showed significant differences in EDA SCR under different emotional valences, but not in EDA SCL. This may be because SCR captures instantaneous changes in the skin and is more responsive to short-term emotional responses, whereas SCL reflects slower changes in the skin and is more indicative of longer-term emotional states. Additionally, in the “positive” valence during the dialogue, RRI values (i.e., the interval between heartbeats) are generally higher than in the “negative” valence, and the variability of RRI is also greater. This is probably because positive emotional states such as relaxation and contentment are associated with a slower heart rate, resulting in increased RRI values. Conversely, negative emotional states, such as anxiety and stress, are generally linked to a faster heart rate and consequently shorter RRI values.

6 Conclusion and Future Work

In this study, we collected dialogue data containing comprehensive multimodal data and subjective

		Positive		Negative		
Signal	Feature	Mean	Std	Mean	Std	p-value
EDA	SCL	1.3e-4	0.01	8.0e-3	0.02	0.229
	SCR	1.38	1.70	1.29	1.71	1.3e-7**
BVP	Rate	68.4	13.4	69.0	12.8	0.042**
	Peak	10.2	2.16	10.3	2.07	0.053
	RRI	944	216	933	209	0.045**
PPG	Rate	85.2	13.7	85.2	10.5	0.644
	Peak	13.2	2.13	13.2	1.67	0.765
	RRI	726	118	718	88.3	0.690
Pupil diameter	Left	0.38	0.23	0.41	0.25	0.115
	Right	0.44	0.24	0.42	0.25	0.070

Table 6: Mean, standard deviation (Std), and the p-value of the Wilcoxon rank-sum test for means under positive and negative valence (**$p < 0.05$).

evaluations at each time step during the dialogue using heterogeneous sensors. Through our analysis of the relationship between physiological signals and emotional valence using the Granger causality analysis, we identified several physiological signals that could be useful for predicting real-time emotional valence. We also clarified how physiological signals differ depending on the "positive" or "negative" polarity of the valence.

However, several limitations of our study should be acknowledged. First, the relatively small sample size limits the statistical power of our findings and reduces the generalizability of the results to a larger population. Second, the imbalanced ratio of positive to negative valence samples may potentially lead to biased conclusions about the relationship between physiological signals and emotional valence.

Future research needs to apply methodologies for analyzing imbalanced and small sample size data. Moreover, we plan to expand our analysis to include sensor signals, linguistic information, questionnaires about personality traits, and impressions of the dialogue, in addition to physiological signals. We will also use the information from the sensors to predict emotional valence in real-time. Ultimately, our goal is to achieve a dialogue system capable of estimating and appropriately responding to the user's mental state in real-time.

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.

Jean-Marc Colletta, Catherine Pellenq, and Michèle Guidetti. 2010. Age-related changes in co-speech gesture and narrative: Evidence from french children and adults. *Speech Communication*, 52(6):565–576.

Antonio R Damasio. 1996. The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 351(1346):1413–1420.

Ellen Douglas-Cowie, Cate Cox, Jean-Claude Martin, Laurence Devillers, Roddy Cowie, Ian Sneddon, Margaret McRorie, Catherine Pelachaud, Christopher Peters, Orla Lowry, et al. 2011. The HUMAINE database. *Emotion-oriented Systems: The Humaine Handbook*, pages 243–284.

Yunyuan Gao, Xiangkun Wang, Thomas Potter, Jianhai Zhang, and Yingchun Zhang. 2020. Single-trial EEG emotion recognition using Granger Causality/Transfer Entropy analysis. *Journal of Neuroscience Methods*, 346:108904.

Jeffrey M Girard. 2014. CARMA: Software for continuous affect rating and media annotation. *Journal of Open Research Software*, 2(1):e5.

Lewis R Goldberg. 1990. An Alternative "description of personality": The Big-Five Factor Structure. *Journal of Personality*, 59(6):1216–1229.

Clive WJ Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37:424–438.

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.

Shun Katada, Shogo Okada, and Kazunori Komatani. 2023. Effects of Physiological Signals in Different Types of Multimodal Sentiment Estimation. *IEEE Transactions on Affective Computing*, 14(3):2443–2457.

Masaki Kimura, Makio Yogo, and Ikuo Daibo. 2005. Expressivity halo effect in the conversation about emotional episodes. *Japanese Journal of Research on Emotions*, 12(1):12–23. (in Japanese).

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

Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2011. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17.

Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *Proc. AAAI*, pages 1359–1367.

Yoshiya Moriguchi and Gen Komaki. 2013. Neuroimaging studies of alexithymia: Physical, affective, and social perspectives. *BioPsychoSocial Medicine*, 7(1):1–12.

Catharine Oertel, Fred Cummins, Jens Edlund, Petra Wagner, and Nick Campbell. 2013. D64: A corpus of richly recorded conversational interaction. *Journal on Multimodal User Interfaces*, 7(1-2):19–28.

Atsushi Oshio, Shingo Abe, and Pino Cutrone. 2012. Development, reliability, and validity of the Japanese version of Ten Item Personality Inventory (TIPI-J). *Japanese Journal of Personality/Pasonariti Kenkyu*, 21(1):40–52.

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.

James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145–172.

Nastaran Saffaryazdi, Yenushka Goonesekera, Nafiseh Saffaryazdi, Nebiyou Daniel Hailemariam, Ebasa Girma Temesgen, Suranga Nanayakkara, Elizabeth Broadbent, and Mark Billinghurst. 2022. Emotion Recognition in Conversations Using Brain and Physiological Signals. In *Proc. IUI*, page 229–242.

Hiro Y Toda and Taku Yamamoto. 1995. Statistical inference in vector autoregressions with possibly integrated processes. *Journal of econometrics*, 66(1-2):225–250.

Sanae Yamashita, Koji Inoue, Ao Guo, Shota Mochizuki, Tatsuya Kawahara, and Ryuichiro Higashinaka. 2023. RealPersonaChat: A Realistic Persona Chat Corpus with Interlocutors' Own Personalities. In *Proc. PACLIC*, pages 852–861.

Zhou Yu, Alexandros Papangelis, and Alexander Rudnicky. 2015. TickTock: A Non-Goal-Oriented Multimodal Dialog System with Engagement Awareness. In *Proc. AAAI Spring Symposium on Turn-taking and Coordination in Human-Machine Interaction*, pages 108–111.