Collecting Human-Agent Dialogue Dataset with Frontal Brain Signal toward Capturing Unexpressed Sentiment

Shun Katada, Ryu Takeda, Kazunori Komatani
SANKEN, Osaka University
{katada, rtakeda, komatani}@sanken.osaka-u.ac.jp

Abstract

Multimodal information such as text and audiovisual data has been used for emotion/sentiment estimation during human-agent dialogue; however, user sentiments are not necessarily expressed explicitly during dialogues. Biosignals such as brain signals recorded using an electroencephalogram (EEG) sensor have been the subject of focus in affective computing regions to capture unexpressed emotional changes in a controlled experimental environment. In this study, we collect and analyze multimodal data with an EEG during a human-agent dialogue toward capturing unexpressed sentiment. Our contributions are as follows: (1) a new multimodal human-agent dialogue dataset is created, which includes not only text and audiovisual data but also frontal EEGs and physiological signals during the dialogue. In total, about 500-minute chat dialogues were collected from thirty participants aged 20 to 70. (2) We present a novel method for dealing with eye-blink noise for frontal EEGs denoising. This method applies facial landmark tracking to detect and delete eye-blink noise. (3) An experimental evaluation showed the effectiveness of the frontal EEGs. It improved sentiment estimation performance when used with other modalities by multimodal fusion, although it only has three channels.

Keywords: multimodal dialogue dataset, human-agent dialogue, frontal EEGs

1. Introduction

Capturing the user's sentiment during dialogue is important for an adaptive dialogue system. For example, if the user seems interested in the current topic, the dialogue system should continue as is, whereas if the user seems bored, the system should change the current topic. Since the estimation of user's sentiment during dialogue is a difficult task, developing the adaptive dialogue systems is still challenging (Clavel and Callejas, 2015).

A promising approach to achieve accurate sentiment estimation is multimodal sentiment analysis (Morency et al., 2011; Baltrušaitis et al., 2018). This approach conducts sentiment analysis using not only the user's utterance contents but also audiovisual information. So far, many datasets for multimodal sentiment analysis have been created (Zhu et al., 2023).

There are two types of labels in sentiment analysis, that is, sentiment labels annotated by the users themselves (hereinafter referred to as self-reported sentiment) or by third-party human coders (hereinafter referred to as third-party sentiment). It is known that self-reported sentiment does not necessarily correspond to third-party sentiment (Truong et al., 2012). In the ideal, systems should capture self-reported sentiment during dialogue; however, self-reported sentiment are not always expressed explicitly.

The key technique we focus on is one involving biosignals. Biosignals are often used in the field of affective computing to detect emotional changes. Among the biosignals, brain signals recorded using an electroencephalogram (EEG) sensor have often been used to study emotion recognition (Alarcão and Fonseca, 2019). Many studies have shown the potential of EEGs collected in a controlled experimental environment. Thus, we believe that EEGs are also useful for detecting unexpressed sentiment of the user in a human-agent dialogue; however, several issues need to be addressed.

First, a human-agent dialogue dataset with an EEG is needed to investigate the effectiveness of EEGs. In this study, we used a patch-type EEG sensor with three-channel measurement (Li et al., 2019) for our multimodal data collection. The electrode sheet of the sensor is designed to be attachable to a user's forehead and is suitable for collecting frontal EEGs during dialogue since it causes neither disturbance nor discomfort.

Second, EEGs are susceptible to noise because of their low amplitude (10–100 μV) (Teplan et al., 2002). In particular, the frontal EEG is susceptible to eye-blink noise (eye-blink artifacts) (Urigüen and Garcia-Zapirain, 2015). Thus, EEG denoising is needed.

Third, it is necessary to investigate whether the simple three-channel EEG sensor is effective for self-reported sentiment estimation. Specifically, comparing the effectiveness with that of other multimodal information is needed. It is also necessary to investigate whether the multimodal fusion of frontal EEGs with other modalities contributes to performance improvement in sentiment estimation.

20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

In this study, we tackled these issues. The main contributions of this work are as follows:

- We created a new multimodal human-agent dialogue dataset with a frontal EEG toward capturing unexpressed sentiment. Text, audiovisual, and physiological data are also included. Both self-reported and third-party sentiment labels are also collected at the utterance level (Section 3).

- We proposed an EEG denoising method with facial tracking for eye-blink noise removal. This method improves the sentiment estimation performance of the model on the basis of the frontal EEG (Section 4 and 5.2).

- Finally, preliminary results of a multimodal sentiment analysis are shown by using the newly created dataset. We showed that the fusion of EEGs and other modalities is effective for self-reported sentiment estimation (Section 5.3).

2. Related Work

In this section, we briefly summarize the related research, i.e., multimodal datasets, EEG datasets, and denoising methods, and finally compare our studies with previous ones.

Various datasets for multimodal sentiment analysis have been created since the early 2000s. IEMOCAP (Busso et al., 2008) is a widely used multimodal dataset that includes text and audiovisual data collected from actors during scripted and unscripted spoken communication. Other popular datasets are MOSI (Zadeh et al., 2016) and MOSEI (Zadeh et al., 2018), which are based on the monologues (such as movie reviews) of YouTube speakers. MOSEI is a larger-scale dataset that includes 23,453 sentences with third-party sentiment annotation. Also, other large-scale datasets created are based on a monologue or dyadic dialogue between humans (Poria et al., 2019; Zadeh et al., 2020). Although datasets comprising human-agent dialogue are scarce, a few datasets (McKeown et al., 2011; Komatani and Okada, 2021) that include multimodal data during human-agent dialogue exist.

There are several publicly available datasets that include EEGs under emotional stimuli such as the MAHNOB-HCI (Soleymani et al., 2011), DEAP (Koelstra et al., 2011), and AMIGOS datasets. (Miranda-Correa et al., 2021). AMIGOS includes EEGs collected from 40 participants with a 14-channel headset during the movie-watching task. Publicly available EEG datasets comprising human-human dialogue are scarce except for the K-EmoCon (Park et al., 2020) and PEGCONV (Saffaryazdi et al., 2022) datasets, which have been more recently created as multimodal dialogue datasets including EEGs. K-EmoCon includes audiovisual, accelerometer, frontal single-channel EEG, and physiological data, collected from 32 participants during a debate task. Emotion annotations are based on self-reported, debate partner, and third-party labels. PEGCONV includes audiovisual, 16-channel EEG, and physiological data collected from 23 participants. Each emotion is induced by imagining, recalling, and expressing emotional memories, and annotated by a self-reported questionnaire.

Various denoising methods have been developed since EEGs have a lower amplitude than other biosignals. Blind source separation (BSS) has been widely used in EEG noise removal to separate target EEGs from noise-contaminated raw EEGs (Urigüen and Garcia-Zapirain, 2015). Among BSS methods, independent component analysis (ICA) is a well-known algorithm that measures independence among the source signals (Comon, 1994). Variations of ICA, such as time-domain ICA (TDICA, Lee, 1998), frequency-domain ICA (FDICA, Smaragdis, 1998), independent vector analysis (IVA, Kim et al., 2006) and independent low-rank matrix analysis (ILRMA, Kitamura et al., 2016), have been established (more advanced methods have been reviewed in Urigüen and Garcia-Zapirain, 2015). Furthermore, methods based on wavelet transform (Chavez et al., 2018) and empirical mode decomposition (Patel et al., 2016) have been proposed. Deep-learning based EEG denoising was also proposed in (Roy et al., 2019; Zhang et al., 2021; Li et al., 2023).

Most denoising methods have been proposed under the assumption that EEG data is collected using a multi-channel setup covering the entire scalp in a controlled environment. Also, visual inspection of the processed data by neurologists is often needed. Furthermore, since biosignals are not easy to distinguish from ground truth and noise, simulated noise is also used for exploration.

Differences between previous studies and ours are considered here. Although datasets containing dialogue *or* EEG data have already been created, those including dialogue *and* EEG data are insufficient. In particular, EEGs collected in human-agent dialogue are scarce. Thus, the creation of a dataset that includes EEGs during dialogue is first needed to explore estimation methods for capturing unexpressed sentiment.

Furthermore, in conjunction with dataset scarcity, a denoising method for a simple frontal EEG during dialogue has not yet been investigated in sentiment estimation. We propose a denoising method that takes advantage of the multimodal data without an additional learning model and show the effectiveness of the method in the sentiment estima-

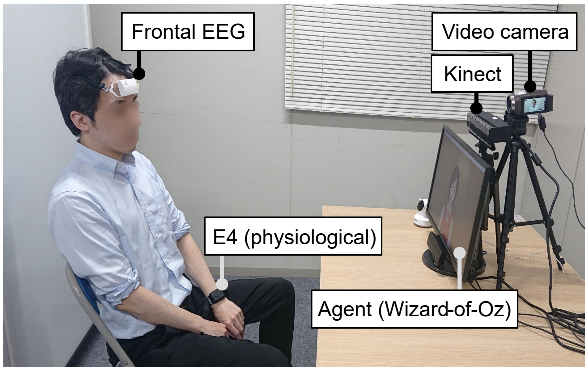

Figure 1: Data collection setup

tion task.

3. Data Collection with EEG

We newly created a multimodal human-agent dialogue dataset with frontal EEGs, Hazumi2306, which will be released as an additional version of the multimodal dialogue corpus Hazumi (Komatani and Okada, 2021). Multimodal data were collected from thirty participants with almost balanced genders (14 male and 16 female) and balanced ages (20 to 70) recruited from the general public from June to July 2023.

A human-agent chit-chat dialogue (dialogue session) is the main core of this data collection. The participants chatted with an agent operated by a human operator. The human operator selected an utterance prepared prior to the study and attempted to make the participant enjoy the conversation. About a dozen topics such as food preference, traveling, and movies were prepared for the dialogue.

The setup, including each device for multimodal data collection with frontal EEGs, is shown in Figure 1, and the overall data collection flow is shown in Figure 2. Multimodal data collection was conducted in three sessions including resting, dialogue, and recall sessions. The dialogue and recall sessions were conducted in Japanese. Self-reported annotation was conducted by the participants themselves after the recall session to collect participants' sentiment labels.

The experimental protocol was reviewed and approved by the research ethics committee of SANKEN, Osaka University, in May 2023. All participants provided written informed consent to participate in the study.

3.1. Setup

The main focus of this study is the collection of brain signals using an EEG sensor with multimodal data in human-agent dialogue. We selected the wearable patch-type EEG sensor HARU-2[1] (PGV Inc., Tokyo, Japan) for collecting frontal EEGs. This EEG sensor was released in 2022 (as the successor to the HARU-1, which can pick up μV brain signals in a manner comparable with traditional EEG equipment (Li et al., 2019)). After skin cleaning, the EEG sensor was attached to the participant's forehead (Figure 1, depicted as "Frontal EEG") with a flexible stretchable electrode sheet and conductive gel to prevent disturbance and discomfort. The three-channel electrodes were positioned in close proximity to the Fpz, Fp1, and Fp2 positions on the forehead, and the reference electrode was attached to the left mastoid (bone behind the left ear). The sampling frequency was 250 Hz, and the room temperature was controlled to be comfortable for the participant.

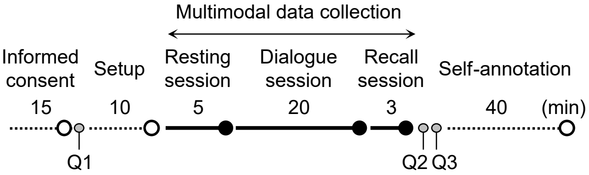

Figure 2: Data collection flow

Other devices were set similarly to a previous multimodal data collection (Komatani and Okada, 2021). In summary, the facial expressions and voice were recorded using a video camera and stored as mp4 files recorded at 30 fps. The sampling rate of the audio recording was 44.1 kHz. The Microsoft Kinect V2 sensor recorded voice, depth information, and upper body posture. For physiological signals, electrodermal activity (EDA, 4 Hz), blood volume pulse (BVP, 64 Hz), and skin temperature (4 Hz) were collected by using an E4 wristband[2] (Empatica Inc., Cambridge, MA, USA) worn around the participant's non-dominant wrist (denoted as "E4 (physiological)" in Figure 1). The system's utterances and timestamps during the dialogue session were also logged.

3.2. Procedure

Multimodal data collection was composed of three sessions (Figure 2). This subsection describes the procedure of each session. The recording summary is shown in Table 1.

3.2.1. Resting Session

An EEG is characterized by low amplitude (μV), and individual differences, and is difficult to prepare the ground truth. Therefore, the resting session was designed to validate frontal EEG data collected from actual participants. We validated

[1] https://www.pgv.co.jp/en/

[2] https://www.empatica.com/research/e4/

Number of participants	30 (M14/F16)
Resting session	
Average duration	5.1 min
Total session duration	151.9 min
Dialogue session	
Average duration	17.2 min
Total session duration	516.6 min
Total number of exchanges	2004
Annotations	
- Self-reported sentiment	per exchange
- Third-party sentiment	per exchange
- Dialogue evaluations	per participant
- Big-Five personality traits	per participant
Recall session	
Average duration	2.3 min
Total session duration	69.9 min
Dataset total	738.4 min

Table 1: Multimodal data collection summary. The dataset includes text, audio, visual (face, depth, posture), physiological signals (EDA, BVP, skin temperature), and frontal EEGs.

the frontal EEG data in both the eyes-open and -closed conditions to determine whether the eyes-open condition induces α-attenuation (Barry et al., 2007) in the frontal forehead region, as observed in a previous study (Li et al., 2019). The participants were instructed to keep their eyes open for 30 seconds while looking at a fixation cross on the display and then to keep their eyes closed for 30 seconds. This cycle was repeated 5 times (5 minutes in total). The collected data was also used to evaluate the effectiveness of our proposed denoising method (mentioned in Section 4).

3.2.2. Dialogue Session

During the dialogue session, the participants chatted with a virtual agent (MMDAgent[3]) shown on a display, operated by a human operator using the Wizard-of-Oz method (i.e., remotely controlling the system from another room). The participants were not informed that the agent was remotely controlled by a human operator until the end of the experiment. In the dialogue session, we defined a small unit as a pair consisting of a system utterance and user utterance, i.e., an exchange (Figure 3, the exchange is denoted as e). Sentiment labels were annotated at the exchange level after the session. More details on the dialogue session procedure were set up similarly to those in previous studies (Komatani and Okada, 2021; Katada et al., 2023).

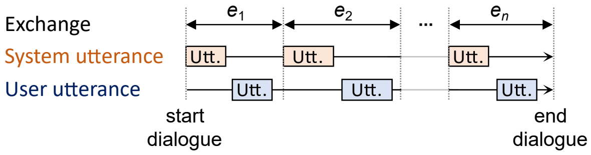

Figure 3: Definition of exchange

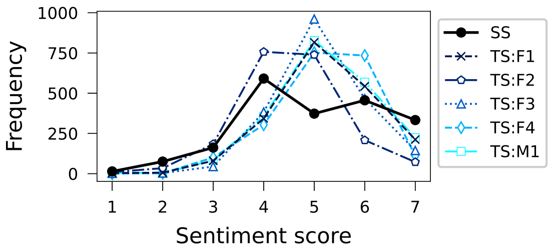

Figure 4: Distribution of the sentiment score. SS (black) indicates self-reported sentiment score, and TS (blue gradient color) indicates third-party sentiment score, respectively. The two characters after TS in legend denote annotator IDs.

3.2.3. Recall Session

We also introduced a recall session, which is the preliminary trial to analyze the relation between positive sentiment and frontal EEGs. Prior to the study, the participants were instructed to prepare a short talk (2 minutes) about something that they had enjoyed recently. After the dialogue session, the system prompted the participants to recall and talk about it in a monologue style. In this way, we explicitly collected frontal EEG data with positive sentiment when participants talked about things enjoyable for them.

3.3. Annotations and Questionnaires

We collected two types of sentiment labels in dialogue sessions at the exchange level: self-reported sentiment and third-party sentiment labels. In particular, the self-reported sentiment label is the main focus of this study. The participants themselves annotated the self-reported sentiment labels after the recall session. They annotated each exchange while watching their videos of the dialogue session. The labels were assigned subjective sentiment scores on a Likert scale ranging from 1 (not enjoying the dialogue) to 7 (enjoying the dialogue).

Third-party sentiment labels were annotated by five annotators while watching the video of the dialogue session ranging from 1 to 7 on whether the participant seemed to enjoy the dialogue. The distributions of the sentiment scores are shown in Figure 4. The agreement between the coder ratings was calculated using Cronbach's alpha, which was 0.88. This high value indicates a high consistency of the annotations.

[3] http://www.mmdagent.jp/

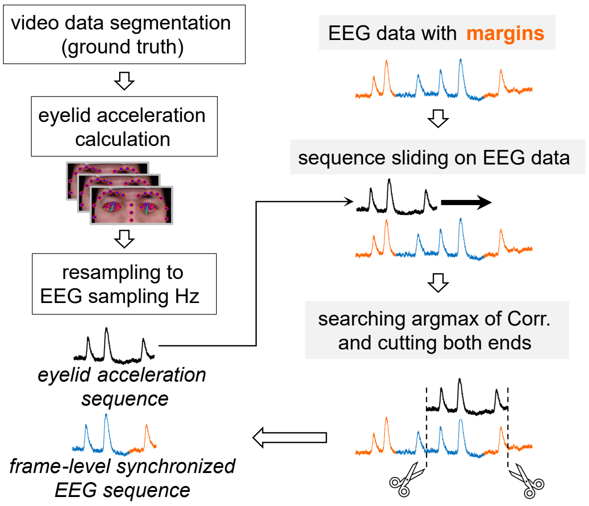
video data segmentation (ground truth)
eyelid acceleration calculation
resampling to EEG sampling Hz
eyelid acceleration sequence
frame-level synchronized EEG sequence
EEG data with margins
sequence sliding on EEG data
searching argmax of Corr. and cutting both ends

Figure 5: Video-EEG synchronization

Three kinds of questionnaires, namely pre-dialogue evaluation (Q1, 18 items), post-dialogue evaluation (Q2, 18 items), and Ten-Item Personality Inventory (Oshio et al., 2012, Q3, 10 items) were also collected. Q1 and Q2 are based on (Kimura et al., 2005), which was derived from (Bernieri et al., 1996). Q3 is what is called Big-Five (Goldberg, 1990). The timing of each questionnaire assessment is depicted in Figure 2. More details of the annotations and questionnaires were set up similarly to those in (Komatani and Okada, 2021).

4. Proposed EEG Denoising

We propose a new eye-blink noise removal algorithm for frontal EEGs denoising that takes advantage of the multimodal information. The proposed method is divided into two steps: video-EEG synchronization (Section 4.1, Figure 5) and EEG denoising with facial tracking (Section 4.2, Figure 6).

4.1. Video-EEG Synchronization

The first step of the proposed denoising method is synchronization between video and EEG data. Although each stand-alone wireless device, such as a video camera and EEG sensor, can be synchronized by the Network Time Protocol before measurement, there is no guarantee that all video frames and EEG data are aligned accurately at the millisecond level during measurement. Thus, a synchronization method for alignment between video and EEG data is needed for accurate analysis. Importantly, this step is essential to proceed to the next step of the proposed method. The flow of video-EEG synchronization is illustrated in Figure 5.

4.1.1. Eyelid Acceleration Calculation

The proposed method starts with video data segmentation to extract the target region (Figure 5, upper left). The video data segmentation in this study was performed at the exchange level on the basis of the dialogue system log. Next, time-series eyelid acceleration is extracted using the OpenFace library (Baltrusaitis et al., 2018), which can track facial landmarks including the eyelid at the frame level. Then, the frame-level eyelid acceleration is resampled to EEG sampling frequency. EEG data with margins of several seconds before and after each exchange are also prepared (Figure 5, upper right, orange).

4.1.2. Extraction of Synchronized EEG

To search for the EEG sequence corresponding to the video data at the frame level, the eyelid acceleration sequence is slid across the prepared EEG data from edge to edge while calculating the correlation coefficient between the two sequences. Following this process, the maximum value of the correlation coefficient is obtained when the eyelid acceleration sequence is located in the corresponding EEG region (Figure 5, lower right). Finally, the corresponding EEG region is extracted by cutting both ends, and a frame-level synchronized EEG sequence is obtained (Figure 5, lower left).

4.2. EEG Denoising with Facial Tracking

Our proposed blink noise removal is achieved by using eyelid acceleration sequence and the BSS technique (Sawada et al., 2019).

One of the issues with applying BSS to EEG denoising is that we cannot determine the permutation of the blink noise signal. Let the observed signal be expressed as $x(t)$, the separation matrix as W, and output source estimates as $y(t) = Wx(t)$. The ICA family calculates W under the assumption that the source estimates are independent of each other. When observed input signals derived from the three channels are $x_1(t), x_2(t)$, and $x_3(t)$ and separated output signals by BSS are $y_1(t), y_2(t)$, and $y_3(t)$, it is impossible to automatically determine which output signal corresponds to the blink noise signal only from the signals themselves.

However, as described in Section 4.1, a synchronized eyelid acceleration sequence is obtained by our proposed video-EEG synchronization. This sequence can be used to determine the blink noise signal automatically. That is, by calculating the correlation coefficient between the eyelid acceleration sequence and $y_1(t), y_2(t)$, and $y_3(t)$, we can identify the blink noise signals whose correlation coefficients with the eyelid acceleration sequence are higher than a threshold[4].

[4] We set it to 0.2 in our experiment.

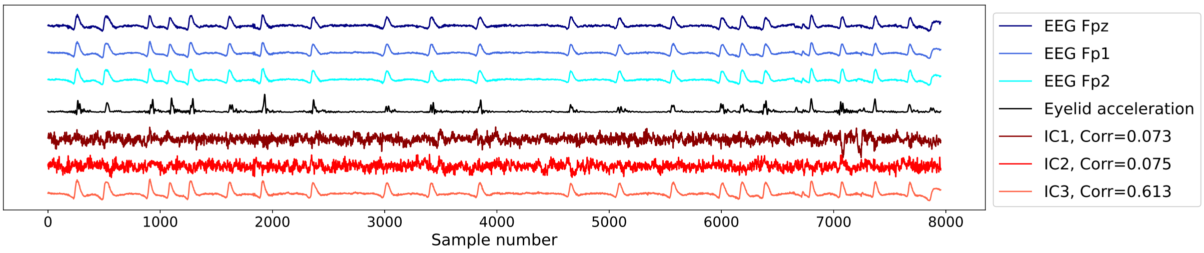

Figure 6: EEG denoising by blink noise identification

An example of the proposed method is shown in Figure 6. Around 30 seconds of frontal EEG data during a resting session from three channels (EEG Fpz, EEG Fp1, and EEG Fp2 corresponding to $x_1(t), x_2(t)$, and $x_3(t)$) are shown in the top three rows. The separated signals by BSS (IC1, IC2, and IC3 corresponding to $y_1(t), y_2(t)$, and $y_3(t)$) are shown in the bottom three rows. The eyelid acceleration sequence is depicted in the central row. The eyelid acceleration sequence is aligned with each sequence at the millisecond level. The correlation coefficient between the eyelid acceleration sequence and IC1, IC2, and IC3 is 0.073, 0.075, and 0.613, respectively. Thus, it is considered that IC3 corresponds to a blink noise signal, and IC1 and IC2 do not. The details of the configuration and evaluation are described in the next section.

5. Evaluation

This section presents the effectiveness of our EEG denoising method and frontal EEG features for sentiment estimation in human-agent dialogue, enabled by our newly created multimodal dataset. Section 5.1 describes the evaluation settings including EEG processing, multimodal processing, and machine learning settings. Section 5.2 shows the effects of our proposed EEG denoising method (described in Section 4) on sentiment estimation. Section 5.3 presents an evaluation of each modality in the unimodal and multimodal sentiment analysis using frontal EEGs.

5.1. Settings

5.1.1. EEG Processing and Configuration

The procedure described here is related to the evaluation of the proposed denoising method (Section 5.2).

The EEG processing consists of linear detrending, standardization (average of 0 and a variance of 1), bandpass filtering (between 1 and 45 Hz), and notch filtering (for 60 Hz hum noise), similar to conventional methods. For the bandpass filter, a zero-phase Butterworth filter was used. Short-time Fourier transform (STFT) or inverse STFT was used to convert to the frequency domain and back to the time domain with a segment length of 256 and a stride length of 128.

For the eyelids sequence, video data of the dialogue session in 30 fps are segmented at the exchange level. The left and right eyelids' acceleration was calculated from the velocity between frames, averaged, and resampled to 250 Hz. The stride length of the eyelid acceleration sequence for searching the maximum values of the correlation coefficient was 5.

To evaluate our proposed method, four types of BSS methods, namely TDICA, FDICA, IVA, and ILRMA are implemented (the details of each BSS are described in Sawada et al., 2019).

After preprocessing in the time or frequency domain, blink noise signals were determined by the method described in Section 4.2.

Then, the remaining source estimates are used for the power calculation of each frequent bin in the range of 1 to 45 Hz. The power value of source estimates is averaged if there are two remaining source estimates. Finally, the summation of the power of each frequent bin is used as frontal EEG features (45-dimensional vector in total).

5.1.2. Multimodal Analysis Settings

The settings described here are related to multimodal analysis (Section 5.3). The newly created dataset was designed for collecting not only frontal EEGs but also text, audio, visual, and physiological information simultaneously (pentamodal in total). In summary, facial expression features as visual features were extracted by OpenFace from segmented video data at the exchange level (66-dimensional vector). Using the wav file derived from the segmented video data, audio features were extracted by openSMILE[5] (384-dimensional vector). The segmented wav file was used for automatic speech recognition (ASR) using Whisper (Radford et al., 2023). The word error rate was 22.7%. Text derived from the ASR is represented by pre-trained[6] BERT (Devlin et al., 2019, 768-

[5] https://www.audeering.com/opensmile/

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

dimensional vector).

Physiological signals (from the E4 wristband) were synchronized by the acceleration log of the hand-clapping, which was performed just before the dialogue session. EDA statistics and a number of the galvanic skin responses in each exchange were extracted as physiological features (14-dimensional vector). More details on feature extraction were set up similarly to those in a previous study (Katada et al., 2023).

5.1.3. Machine Learning Settings

The settings described here are related to all evaluations (Sections 5.2 and 5.3). All evaluations were performed as sentiment estimation using the collected dialogue session data and self-reported sentiment labels at the exchange level.

Linear support vector regression (Cristianini and Shawe-Taylor, 2000) was used for the evaluation. Leave-one-person-out cross-validation (LOPOCV) was performed for 30 participants. That is, the samples (exchanges) from one participant were used as test data, and the remaining samples from 29 participants were used as the training data. For the optimization, a threefold cross-validation scheme was used for the training data set with the penalty parameters set as {0.001, 0.01, 0.1, 1, 10, 100}, the insensitivity parameters set as {0, 0.5, 1}, and the maximum number of iterations set as 1000.

The machine learning settings for the multimodal model evaluations were identical to those of the unimodal ones for fair comparison. Multimodal fusion was conducted by concatenating each feature vector (so-called early fusion (Baltrušaitis et al., 2018)), and inputting the support vector regression with the same configuration as that of the unimodal models. All combinations derived from five modalities were evaluated, and thus $2^5 - 1$ models were evaluated in total.

We reported the average mean absolute error (MAE) and Spearman correlation (Corr) in the LOPOCV. All experiments were performed three times, and the evaluation values were calculated as averages across the three repetitions. The majority baseline for self-reported sentiment estimation has an MAE of 1.166, which is calculated under the assumption that the estimation value is always the average of the sentiment score of training data.

5.2. Effectiveness of Denoising

Table 2 shows the regression performance in self-reported sentiment estimation based only on the frontal EEGs, to which our proposed denoising methods with each BSS were applied. "None" (the second row in Table 2) indicates no blink noise removal. The proposed method with ILRMA achieved the best performance with an MAE of 1.145 and a Corr of 0.200 (the second row from the bottom in Table 2). This performance is better than the majority baseline (MAE of 1.166, the last row in Table 2).

BSS method	MAE	Corr
None	1.187	0.181
TDICA	1.155	0.199
FDICA	1.249	0.195
IVA	1.208	0.182
ILRMA	**1.145**	**0.200**
Majority	1.166	-

Table 2: Frontal EEG-based self-reported sentiment estimation with proposed denoising method

Modality	MAE	Corr
T (Text)	1.096	0.274
A (Audio)	**1.083**	**0.328**
V (Visual)	1.227	0.088
P (Physiological)	1.220	0.149
B (Brain)	1.145	0.200
Majority	1.166	-

Table 3: Self-reported sentiment estimation with unimodal models. Bold indicates the best performance among all unimodal models.

5.3. Sentiment Analysis with EEG

Table 3 presents the estimation performance using unimodal data. Among the unimodal models, the model based on the audio feature achieved the best performance (MAE of 1.083, Corr of 0.328). The second best model is that based on text features, followed by the frontal EEG feature (denoted as "B (Brain)"). The frontal EEG feature, even when used alone, is useful for the estimation.

We evaluated all multimodal models combining features corresponding to each five modality (26 models in total). We reported representative model performance that related to the best performance (Table 4). Among the multimodal models, a combination of the text (T), audio (A), and frontal EEG (B) features achieved the best performance in terms of Corr (0.338). In terms of MAE, a combination of the text (T), visual (V), physiological (P), and frontal EEG (B) features achieved the best result (1.039). The frontal EEG (B) feature improves estimation performance by combining the text (T) or audio (A) features (rows 2 to 5 in Table 4), compared with unimodal models in Table 3. This effect is limited in quad (T, V, P, B, Corr of 0.309) or pentamodal (T, A, V, P, B) models (rows 2 and 4 from the bottom in Table 4).

The strong benchmark in self-reported estimation is the estimation performance based on the human (the last row in Table 4, derived from five human

Modality	MAE	Corr
T, B	1.074	0.302
A, B	1.079	0.333
T, A	1.080	0.333
T, A, B	1.080	**0.338**
T, V, P	1.049	0.321
T, V, P, B	**1.039**	0.309
T, A, V, P	1.045	0.328
T, A, V, P, B	1.044	0.329
Human	1.038	0.387

Table 4: Self-reported sentiment estimation with multimodal models. Bold indicates the best performance among all multimodal models (26 multimodal models were evaluated in total).

annotators). Although there is still a gap in terms of Corr, the best multimodal model has a performance close to human in terms of MAE.

6. Discussion

The created dataset is particularly valuable for exploring the method that enables the dialogue system to consider the user's sentiment, even if it is unexpressed. Further analysis of this dataset by data scientists or EEG researchers can contribute to the development of an adaptive dialogue system.

One of the future issues for the denoising method is that noise other than eye-blink, such as muscle or cardiac noise (Urigüen and Garcia-Zapirain, 2015), is not removed in this study. Additionally, since single-channel-based noise removal methods, such as wavelet transform (Chavez et al., 2018) and empirical mode decomposition (Patel et al., 2016), have been proposed, the comparison and combination of these techniques with our proposed method are needed to evaluate under a multimodal dialogue scenario.

Self-reported sentiment estimation in a strict LOPOCV (user-independent) schema is a difficult task even if frontal EEGs are introduced. Since there are individual differences in brain activity (Greene et al., 2022), one of the alternative ways is user-dependent evaluation, i.e., using data derived from the same user in both the training and test phase. On the basis of the survey (Alarcão and Fonseca, 2019), 43.5 percent of EEG studies use user-dependent data. Although this schema lacks generalizability, clearer effectiveness of the frontal EEG may be observed.

Another point to consider is that combinations of text, audio, and frontal EEGs can be useful for capturing unexpressed sentiment. We have also observed that physiological features contribute to reducing the error rate when fusing text or audio features (data not shown). Since self-reported sentiments may not necessarily be expressed through text and audiovisual modalities, it is suggested that EEGs and physiological signals, which are involuntarily regulated, can complement other modalities. This observation is consistent with a previous study (Katada et al., 2023) that demonstrated the effectiveness of physiological signals in self-reported sentiment estimation. Since third-party sentiment is based on text and audiovisual modalities (but not on biosignals), a comparative evaluation of each modality in self-reported and third-party sentiment estimation is an important area for future research.

Applying the state-of-the-art model of multimodal machine learning is also considered as future work. As shown in Table 4, the performance improvement by fusing frontal EEG features is limited in quad or pentamodal models (rows 2 and 4 from the bottom in Table 4). Because only conventional linear support vector regression was used for evaluation as a preliminary experiment, it seems that there may be no room for complementation if many modalities are fused. Although various multimodal models for third-party sentiment estimation have been proposed recently (Zhu et al., 2023), exploration of effective models for self-reported sentiment estimation with EEGs is still insufficient in terms of learning methods and neural network architectures. Thus, further investigation is needed in parallel with the aforementioned points.

One limitation of our study is that EEG data is collected using only three electrodes in a frontal position. Different emotional stimuli induce different neural patterns in each brain region. For example, neural patterns for negative emotions have higher gamma responses at prefrontal sites, but higher delta responses at parietal and occipital sites (Zheng et al., 2017). Therefore, electrodes covering the entire scalp are used for controlled experiments. On the other hand, our aim in this paper is to investigate whether the simple EEG sensor can effectively capture the user's sentiment, as simple devices are easy to use in a variety of applications, including dialogue. Thus, selecting a device aligned with the research question is crucial.

Another limitation is that it is unknown which of the brain waves such as alpha, beta, and gamma waves are related to the user's sentiment in this study. Since we also collected frontal EEG during resting and recall sessions, a detailed analysis of these data can clarify the relationship between sentiment and frontal EEG.

7. Conclusion

In summary, we presented a new multimodal dialogue dataset that includes frontal brain activity in human-agent dialogue. Together with this, we introduced an eye-blink noise removal technique using video-EEG synchronization to deal with the low amplitude of EEGs. Finally, preliminary results of self-reported sentiment estimation were shown to evaluate the potential of frontal EEGs. We are convinced that this dataset is useful and unique for the exploration of multimodal analysis with frontal EEGs, and further investigation will contribute to the dialogue system development, which can consider the user's unexpressed sentiment.

8. Acknowledgments

This work was partly supported by JSPS KAKENHI Grant Numbers JP23K20002 and JP22H00536.

9. Bibliographical References

Soraia M. Alarcão and Manuel J. Fonseca. 2019. Emotions recognition using EEG signals: A survey. *IEEE Transactions on Affective Computing*, 10(3):374–393.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. *International Conference on Automatic Face & Gesture Recognition (FG)*, pages 59–66.

Robert J Barry, Adam R Clarke, Stuart J Johnstone, Christopher A Magee, and Jacqueline A Rushby. 2007. EEG differences between eyes-closed and eyes-open resting conditions. *Clinical Neurophysiology*, 118(12):2765–2773.

Frank J Bernieri, John S Gillis, Janet M Davis, and Jon E Grahe. 1996. Dyad rapport and the accuracy of its judgment across situations: A lens model analysis. *Journal of Personality and Social Psychology*, 71(1):110.

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 (LREC)*, 42(4):335–359.

Mario Chavez, Fanny Grosselin, Aurore Bussalb, F De Vico Fallani, and Xavier. Navarro-Sune. 2018. Surrogate-based artifact removal from single-channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(3):540–550.

Chloe Clavel and Zoraida Callejas. 2015. Sentiment analysis: From opinion mining to human-agent interaction. *IEEE Transactions on Affective Computing*, 7(1):74–93.

Pierre Comon. 1994. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314.

Nello Cristianini and John Shawe-Taylor. 2000. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 4171–4186.

Lewis R Goldberg. 1990. An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, 59(6):1216–1229.

Abigail S Greene, Xilin Shen, Stephanie Noble, Corey Horien, C Alice Hahn, Jagriti Arora, Fuyuze Tokoglu, Marisa N Spann, Carmen I Carrión, Daniel S Barron, et al. 2022. Brain–phenotype models fail for individuals who defy sample stereotypes. *Nature*, 609(7925):109–118.

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.

Taesu Kim, Torbjørn Eltoft, and Te-Won Lee. 2006. Independent vector analysis: An extension of ICA to multivariate components. *Independent Component Analysis and Blind Signal Separation (ICA 2006)*, pages 165–172.

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

Daichi Kitamura, Nobutaka Ono, Hiroshi Sawada, Hirokazu Kameoka, and Hiroshi Saruwatari. 2016. Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9):1626–1641.

Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2011. DEAP: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31.

Kazunori Komatani and Shogo Okada. 2021. Multimodal human-agent dialogue corpus with annotations at utterance and dialogue levels. *International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8.

Te-Won Lee. 1998. *Independent Component Analysis*, pages 27–66. Springer US, Boston, MA.

Fangzhou Li, Naohiro Egawa, Shusuke Yoshimoto, Haruo Mizutani, Katsuya Kobayashi, Naoko Tachibana, and Ryosuke Takahashi. 2019. Potential clinical applications and future prospect of wireless and mobile electroencephalography on the assessment of cognitive impairment. *Bioelectricity*, 1(2):105–112.

Yuqing Li, Aiping Liu, Jin Yin, Chang Li, and Xun Chen. 2023. A segmentation-denoising network for artifact removal from single-channel EEG. *IEEE Sensors Journal*.

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.

Juan Abdon Miranda-Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. 2021. AMIGOS: A dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing*, 12(2):479–493.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. *International Conference on Multimodal Interfaces (ICMI)*, pages 169–176.

Atsushi Oshio, ABE Shingo, and Pino Cutrone. 2012. Development, reliability, and validity of the Japanese version of Ten Item Personality Inventory (TIPI-J) (in Japanese). *Japanese Journal of Personality*, 21(1).

Cheul Young Park, Narae Cha, Soowon Kang, Auk Kim, Ahsan Habib Khandoker, Leontios Hadjileontiadis, Alice Oh, Yong Jeong, and Uichin Lee. 2020. K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data*, 7(293).

Rajesh Patel, Madhukar PandurangRao Janawadkar, Senthilnathan Sengottuvel, Katholil Gireesan, and Thimmakudy Sambasiva Radhakrishnan. 2016. Suppression of eye-blink associated artifact using single channel EEG data by combining cross-correlation with empirical mode decomposition. *IEEE Sensors Journal*, 16(18):6947–6954.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. *Association for Computational Linguistics (ACL)*, pages 527–536.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. *International Conference on Machine Learning (ICML)*, 202:28492–28518.

Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. 2019. Deep learning-based electroencephalography analysis: A systematic review. *Journal of Neural Engineering*, 16(5):051001.

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. *International Conference on Intelligent User Interfaces (IUI)*, page 229–242.

Hiroshi Sawada, Nobutaka Ono, Hirokazu Kameoka, Daichi Kitamura, and Hiroshi Saruwatari. 2019. A review of blind source separation methods: Two converging routes to ILRMA originating from ICA and NMF. *APSIPA Transactions on Signal and Information Processing*, 8:e12.

Paris Smaragdis. 1998. Blind separation of convolved mixtures in the frequency domain. *Neurocomputing*, 22(1):21–34.

Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. 2011. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55.

Michal Teplan et al. 2002. Fundamentals of EEG measurement. *Measurement Science Review*, 2(2):1–11.

Khiet P Truong, David A Van Leeuwen, and Franciska MG De Jong. 2012. Speech-based recognition of self-reported and observed emotion in a dimensional space. *Speech Communication*, 54(9):1049–1063.

Jose Antonio Urigüen and Begoña Garcia-Zapirain. 2015. EEG artifact removal—state-of-the-art and guidelines. *Journal of Neural Engineering*, 12(3):031001.

Amir Zadeh, Yan Sheng Cao, Simon Hessner, Paul Pu Liang, Soujanya Poria, and Louis-Philippe Morency. 2020. CMU-MOSEAS: A multimodal language dataset for spanish, portuguese, german and french. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020:1801–1812.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. *Association for Computational Linguistics (ACL)*, pages 2236–2246.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88.

Haoming Zhang, Mingqi Zhao, Chen Wei, Dante Mantini, Zherui Li, and Quanying Liu. 2021. EEGdenoiseNet: A benchmark dataset for deep learning solutions of EEG denoising. *Journal of Neural Engineering*, 18(5):056057.

Wei-Long Zheng, Jia-Yi Zhu, and Bao-Liang Lu. 2017. Identifying stable patterns over time for emotion recognition from EEG. *IEEE Transactions on Affective Computing*, 10(3):417–429.

Linan Zhu, Zhechao Zhu, Chenwei Zhang, Yifei Xu, and Xiangjie Kong. 2023. Multimodal sentiment analysis based on fusion methods: A survey. *Information Fusion*, 95:306–325.