Is She Truly Enjoying the Conversation?: Analysis of Physiological Signals toward Adaptive Dialogue Systems

Shun Katada, Shogo Okada, Yuki Hirano
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, Japan
{s2040005,okada_s,s1810154}@jaist.ac.jp

Kazunori Komatani
Osaka University
Ibaraki, Osaka, Japan
komatani@sanken.osaka-u.ac.jp

ABSTRACT

In human–agent interactions, it is necessary for the systems to identify the current emotional state of the user to adapt their dialogue strategies. Nevertheless, this task is challenging because the current emotional states are not always expressed in a natural setting and change dynamically. Recent accumulated evidence has indicated the usefulness of physiological modalities to realize emotion recognition. However, the contribution of the time series physiological signals in human–agent interaction during a dialogue has not been extensively investigated. This paper presents a machine learning model based on physiological signals to estimate a user's sentiment at every exchange during a dialogue. Using a wearable sensing device, the time series physiological data including the electrodermal activity (EDA) and heart rate in addition to acoustic and visual information during a dialogue were collected. The sentiment labels were annotated by the participants themselves and by external human coders for each exchange consisting of a pair of system and participant utterances. The experimental results showed that a multimodal deep neural network (DNN) model combined with the EDA and visual features achieved an accuracy of 63.2%. In general, this task is challenging, as indicated by the accuracy of 63.0% attained by the external coders. The analysis of the sentiment estimation results for each individual indicated that the human coders often wrongly estimated the negative sentiment labels, and in this case, the performance of the DNN model was higher than that of the human coders. These results indicate that physiological signals can help in detecting the implicit aspects of negative sentiments, which are acoustically/visually indistinguishable.

CCS CONCEPTS

• **Computing methodologies** → **Discourse, dialogue and pragmatics**; *Physiological model.*

KEYWORDS

Physiological Signal Processing; Social Signal Processing; Multimodal Interaction

ICMI '20, October 25–29, 2020, Virtual event, Netherlands

ACM ISBN 978-1-4503-7581-8/20/10…$15.00
https://doi.org/10.1145/3382507.3418844

ACM Reference Format:
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 *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3382507.3418844

1 INTRODUCTION

Multimodal behavioral processing technology is a key technique for developing an empathetic dialogue system that can adapt to the behavior of a human user. Although many studies pertaining to human–agent or human–robot interaction settings have focused on verbal information, nonverbal information is also valuable to estimate a user's positive or negative sentiments. In addition, as the user's sentiment states can change dynamically during dialogues, it is necessary to capture the dynamic changes in real time. Explicit behaviors that can be observed as visual information, such as facial expressions and body motion, and acoustic information, such as speaking activity and prosody, are known to be useful in the emotion recognition task [25]. Nevertheless, as people often refrain from expressing their emotions during social interaction, not all emotions are explicitly expressed as linguistic, acoustic, or visual information. Moreover, it is challenging to recognize the changes in the emotional states in a natural situation with no emotional stimuli, by using only observable signals such as visual and acoustic information. Recently, biosignals including electroencephalograms (EEG), electrocardiograms (ECG), and electrodermal activity (EDA) have been used to detect changes in the implicit responses and emotional states of a user. For example, applications utilizing these biosignals have been reported for a movie watching task [6], stress detection [12], and the provision of personalized recommendations [24]. However, the contribution of these biosignals in estimating a user's sentiment during dialogues remains unknown.

Physiological signals can be used to estimate sentiments because these signals are closely related to the states of the autonomic nervous system. The autonomic nervous system consists of the sympathetic and parasympathetic nervous systems, which maintain the homeostasis of organisms by involuntary automatic control of the peripheral organs in the body [15]. For example, the emotions of anger and fear activate the sympathetic nervous system and increase the heart rate (HR) and respiratory rate. In contrast, when relaxing, the parasympathetic nervous system is the dominant part and decreases the HR and respiratory rate. The EDA is another representation of physiological changes and has been widely used in emotion related research [22, 27]. The EDA indicates electrical changes on the skin surface, derived from the activity of

the eccrine sweat glands, and is considered to be an arousal indicator [22]. In addition, a correlation has been reported between the regional cerebral blood flow measured using the positron emission tomography and the HR variability in emotion evoking stimuli [20]. This evidence appears to indicate a strong correlation between the brain and peripheral tissues. Thus, valuable information for emotion recognition can likely be obtained from such physiological signals.

It is difficult to correctly estimate a user's sentiments using only the acoustic and visual information if the user does not explicitly express his/her emotion to the dialogue system. In this regard, biosignals may enhance the performance of user sentiment estimation by supplementing the acoustic and visual information collected simultaneously [4, 14, 19, 29], as long as the wearable sensors do not disturb the dialogue. In this study, we demonstrated that the physiological information collected from participants engaged in dialogues with the agents improved the estimation accuracy of the participants' sentiment labels, which were annotated by the participants themselves for each exchange.

The main contributions of this work can be summarized as follows.

Estimating sentiment labels by using the physiological signals during dialogues: To clarify the effectiveness of the physiological signals in estimating a participant's sentiment label, we evaluated models based on the physiological modality in human–agent interaction settings and compared them with those pertaining to acoustic and visual information. In addition, we verified the effectiveness of combining the physiological signals with acoustic/visual signals on the same task. The experimental results are presented in Section 6.

Comparison between multimodal DNN and human model: We collected a new dialogue corpus, including two types of sentiment labels annotated to each exchange consisting of a system utterance followed by a participant utterance. One is the sentiment labels annotated by the participants themselves and the other is those annotated by multiple human coders. The accuracy of human coders in estimating the participant sentiments was examined to clarify the difference in the two types of sentiment labels. Moreover, the accuracies of estimation by the human coders and models trained with multimodal features helped compare the performances of third party humans and computational models involving physiological signals. The analysis helped demonstrate the challenging nature of the task and the contribution of the automatic multimodal recognition technique in estimating the participants' sentiment states. This analysis is described in Section 7.1.

Example showing relationships between sentiment labels and EDA signals: We investigated the relationship between the participants' sentiment scores and EDA features. The results of the correlation analysis were used to correlate the galvanic skin response (GSR) numbers and an EDA feature with the sentiment scores. We examined the time series sentiment scores and GSR numbers and presented an example of the dynamic changes in these parameters. The analyses are described in Section 7.2.

2 RELATED WORKS

Research on the detection, modeling, and practical application of human emotional behavior is known as affective computing [25]. In the affective computing domain, relationships between the emotional and nonverbal information, such as facial expressions, speech, gestures, and physiological states have been examined [26, 27]. In [6, 33], multimodal data including EEG, ECG, and EDA data were collected while the participants watched a video, and an emotion recognition model was proposed based on these biosignals. For emotion elicitation, these studies used videos that were classified into one of four quadrants of the valence arousal space. Kim et al. [13] investigated the potential of physiological signals for emotion recognition by using biosensors such as electromyogram, ECG, EDA, and respiration sensors. As emotional stimuli, they used music that spontaneously induced real emotional states in the users. Kalimeri et al. [12] presented a multimodal framework to detect the stress of visually impaired people when they were placed in unfamiliar locations. The EEG and EDA data were collected using wearable sensing devices, and a random forest model was used to estimate stressful environmental conditions. With advances in biosignal sensors, many studies have focused on emotion recognition using biosignals [5, 21, 28, 30]. However, only a few studies under nonstressful conditions or without emotional stimuli, especially in human–agent interaction settings, have been conducted. Therefore, in this study, we investigated the effectiveness of physiological signals for sentiment estimation in an interactive chat dialogue.

To implement an adaptive dialogue system, it is important to recognize the user's engagement, interest, and sentiment (e.g., enjoyment during the conversation) based on multimodal behaviors, and many studies have focused on these factors [2, 11, 23]. In [11], a recognition model for user engagement (interest and willingness to continue the dialogue) in human–robot interactions was proposed based on the user's audio–visual information. In [35], to assess the presence of the interest of a user in a time series, they considered an exchange between the system and user as a unit in a chat dialogue. The facial expression, head movement, and prosody of the utterances were used as the multimodal information in this study. Tavabi et al. [34] attempted to generate natural and engaging social interactions in human–agent dialogue systems and estimated the empathy in an uncontrolled environment. They proposed a multimodal DNN to identify opportunities in which the agent should express empathetic responses. In the aforementioned studies, the estimation was based on the user's explicit information, such as the audio/visual information, and the physiological signals were not considered. In our study, we constructed models based on multimodal information, including physiological signals, which can help detect the implicit aspects of a user's sentiment during dialogues.

We used a multimodal dialogue corpus including the user's interest label, user's sentiment label, and topic continuance, which were annotated by human coders at the exchange level [10], to implement an adaptation mechanism of the dialogue strategy in spoken dialogue systems. These three labels were correlated and simultaneously captured the different aspects of the internal state of the user. Considering the relationship among the labels, we applied a multitask learning technique to the binary classification tasks and

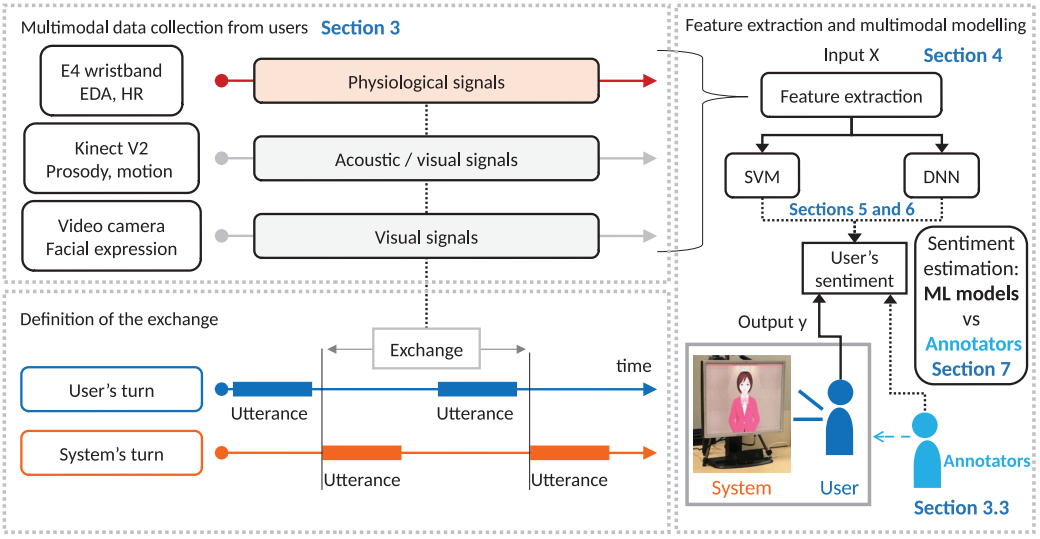

Figure 1: Overview of the estimation of the user's sentiment at the exchange level.

demonstrated that a multitask DNN model trained with multimodal features outperformed a single task DNN. The dialogue corpus we used did not include physiological data. In this study, the newly collected dialogue corpus included not only acoustic/visual features but also physiological features. Moreover, this corpus included the exchange level sentiment labels annotated by both the participants themselves and third party human coders. Thus, the corpus enabled the investigation of the novel aspects of the physiological features in this setting and comparison of the effectiveness corresponding to the physiological and acoustic/visual modalities to estimate the user's sentiment.

Chaminade et al. [3] constructed an experimental setup that provided temporally aligned behaviors along with physiological activity during human–agent interactions. They focused on the communicative behavior in social interactions and showed that the physiological measures were correlated with various communicative behaviors; however, the user sentiment was not annotated. Egorow et al. [8] showed that physiological signals, including electromyograms, skin conductivity, and respiration, could help detect dialogue stages in which the user experienced trouble in interacting with the dialogue system. However, the user's sentiment labels were not annotated by the users and were simply divided into two classes based on the predetermined dialogue situation. In our study, the sentiment score was annotated both by the participants themselves and external coders for every exchange in a natural chat dialogue. Therefore, models that recognize the dynamically changing sentiments of the user can be constructed, and adaptation strategies for multimodal dialogue systems can be implemented.

3 DATA

We used a multimodal dialogue corpus named Hazumi1911, collected from November 2019. The recording setting was almost the same as that of Hazumi1712 [18] and Hazumi1902 [16], except physiological sensors were newly used in Hazumi1911. [1]

3.1 Data Collection

Figure 1 shows an overview of the study flow. Data were collected in the context of a human–agent dialogue as in [18], following which, the participants communicated with a virtual agent known as MMDAgent[2] shown on the display. The agent was operated using the Wizard of Oz method. Specifically, a human operator (Wizard) remotely controlled the system and interacted with participants in another room. The participants were not informed that the agent was remotely controlled by a human operator until the end of the experiment. No specific task was assigned in the dialogue; i.e., the participants simply chatted with the agent.

Basically, the operator selected the utterances of the agents from the pre-defined utterance list by watching the participants' states through a camera. The operator tried to make them enjoy the conversation and want to continue talking. Because the operator was well trained and had time to select the next utterance while the participant was speaking (around 10-second long), there was a small waiting time before the agent started responding. The agent generated random animation of subtle movements as multimodal behavior (head and hand gestures and facial expressions), which is a built-in component of MMDAgent.

The time series physiological signals were collected during the dialogues using a physiological sensor, that is, the Empatica E4 wristband[3]. In general, if the sympathetic nervous system is activated by emotional stimuli, sweat glands are activated, increasing the level of sweating. These changes might not be perceptible by the user; however, the EDA sensor can detect these small changes

[1] Hazumi1712 and 1902 are currently publicity available [17]; Hazumi1911 will be released similarly.
[2] http://www.mmdagent.jp/
[3] https://www.empatica.com/research/e4/

as changes in the skin conductance (SC) by using two electrodes in contact with the skin. Furthermore, as the E4 device is wireless and worn like a wristwatch, it causes neither disturbance nor discomfort during the dialogues. Thus, this device is suitable to investigate a user's sentiment during dialogues. The EDA and HR of the participants were recorded at 4 and 1 Hz, respectively. In addition, a blood volume pulse was obtained, and the HR was computed as the output from this device. In terms of the acoustic signals, the voice of the participant was recorded as a 16 kHz WAV file by using a Microsoft Kinect V2 sensor. In terms of the visual signals, the facial expressions of the participants were recorded using a video camera at 30 frames per second (fps), and motion data were recorded using the Kinect sensor at 30 fps.

3.2 Participants

Thirty participants (aged 20–70 y; male/female, 15/15) were recruited from the general public through a recruitment agency. Data from 26 participants were used for analysis; the data of four participants were disregarded because of missing values after preprocessing. The average duration of the data was 20.5 min per participant. The dialogue data of one participant contained 95 exchanges on average.

3.3 Annotation

Two types of annotations were labeled in this study: (1) self-sentiment annotation and (2) external sentiment annotation, which were annotated by the participants themselves and external coders, respectively. In this study, an exchange was defined as a section that began from the start time of a system utterance and ended at the start time of the next system utterance. Based on this definition, a total of 2468 exchanges obtained from 26 participants were annotated. The annotation procedures were as follows:

(1) Self-sentiment annotation: The participants themselves annotated the labels per exchange while watching their videos after the experiment. The labels were assigned as scores ranging from 1 (not enjoying the dialogue) to 7 (enjoying the dialogue). The positive sentiments included "enjoy talking", "want to continue talking", and "satisfied with the talk", and the negative sentiments included "want to stop talking" and "confused about the system utterances".

(2) External sentiment annotation: Five human coders annotated the labels per exchange as scores ranging from 1 (participants seem to be bored with the dialogue) to 7 (participants seem to enjoy the dialogue) while watching the recorded videos of the dialogues. This assessment was performed considering the acoustic, visual, and linguistic features of the participants. The human coders were instructed not to assign labels considering only a part of the exchange and to assign labels considering the differences among individual participants after watching the entire recording of the target participant.

The agreement between the coder ratings was calculated using Cronbach's alpha. Generally, a Cronbach's alpha of > 0.8 indicates a high consistency between the annotated labels. In this study, Cronbach's alpha was 0.83 in the external sentiment annotation, indicating the reliability of the annotation. A more detailed description of the annotation methods has been presented in [10].

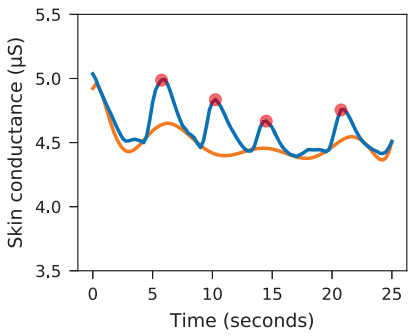

Figure 2: Example of SC analysis, showing the EDA signal (blue curve), tonic component (baseline, orange curve), and GSR (red circles).

4 MULTIMODAL FEATURE EXTRACTION

We focused on the analysis of nonverbal data, especially the analysis of the physiological implicit responses. To compare the effectiveness of the nonverbal features, the physiological, acoustic, and visual information was synchronized using the log data and preprocessed for feature extraction. All the features were extracted from the whole dialogue per exchange, similar to the annotation procedure described in Section 3.3. In this section, we describe the nonverbal features extracted from each modality.

4.1 Physiological Features

The EDA and HR were recorded using the E4 wristband placed on the participants' wrist. The EDA, measured as the SC, reflects the sweat gland activity through the sympathetic nervous system and is widely used to detect the changes in the emotional states at the arousal level [22]. The SC in the time series was decomposed into the SC level (tonic component) and SC response (also known as the GSR). Therefore, the SC level was calculated using polynomial fitting (degree of 10), and the GSR was detected using PeakUtils[4] (amplitude threshold of 0.3). Subsequently, the GSR number per exchange was extracted as an EDA feature (Figure 2). Moreover, we calculated the following statistics for the EDA and HR in each exchange and used them as physiological features: mean, standard deviation, skewness, kurtosis, maximum and minimum values, mean of the first and second differences, range (difference between maximum and minimum values), slope and intercept of the linear approximation, and 25th and 75th percentile values. Overall, 27 features (14 and 13 features from the EDA and HR, respectively) were extracted as the physiological features from each exchange. The data were normalized using the min-max normalization into a range of zero to one.

4.2 Acoustic and Visual Features

Acoustic signals from the participant utterances were used to extract features. The INTERSPEECH 2009 Emotion Challenge feature set (IS09) [31] was extracted using the OpenSMILE[5] software. The

[4]https://pypi.org/project/PeakUtils/
[5]https://www.audeering.com/opensmile/

features were calculated as statistics, and 384 acoustic features were extracted in total from each exchange.

The facial expressions and motion activity in each exchange were extracted as the visual features. Using the OpenFace library[1], the facial landmarks around the eye, mouth, and eyebrow were determined, and the velocity and acceleration were calculated at each point for the facial feature extraction. The estimated categories of the facial action units described in [9] were used as the facial features. The motion data of the hands, shoulders and head, recorded by the Microsoft Kinect sensor were employed, and the calculated velocity and acceleration were used as the motion features. Overall, 87 features were extracted from the facial expressions and motion activity as the visual features. The data were normalized for each participant through the Z score normalization, that is, considering a mean and standard deviation of zero and one, respectively, for all samples pertaining to each participant.

5 EXPERIMENT

The aim of this study was to verify whether physiological features can help estimate a participant's sentiment labels. To this end, we performed binary classification tasks on the sentiment labels by using machine learning models and an external sentiment annotation score (which can be regarded as a "human" model). In the binary classification tasks, the sentiment labels were divided into high and low classes considering a threshold of 4 (neutral state). The number of high/low classes of the sentiment labels was 1119/1349. Similarly, the external sentiment annotation score was processed and divided into high and low classes, and the number of the high/low classes of the external sentiment annotation was 1701/767. In the correlation analysis, the sentiment scores in the range of 1 to 7 were used to calculate the correlation coefficient.

5.1 Machine learning models

5.1.1 Linear Support Vector Machine (SVM). In the binary classification task, linear SVM models [7] based on physiological, acoustic, visual and multimodal features were constructed to compare the estimation accuracy. The SVM models were optimized using a fivefold cross-validation scheme for the training data set with the penalty parameters set as $\{0.001, 0.01, 0.1, 1, 10\}$. The penalty parameter ensures a balance between the loss function and margin maximization. We used the SVM in two ways to fuse the different modalities: early fusion (EF) and late fusion (LF). In EF, the feature vectors from different modalities were concatenated into one feature vector. In the LF, the results of the trained unimodal output were combined to provide a final estimation. In the SVM model, the final estimation was based on the decision function of the unimodal models.

5.1.2 Deep Neural Network (DNN). We used DNN models to verify whether the models improved the performance in the binary classification task. To this end, we used DNNs in two ways to fuse the different modalities, similar to the aforementioned SVM modeling.

To train the unimodal feature set using the EF, the DNN was composed of an input layer, two middle layers with 64 units, two middle layers with 32 units, and an output layer. When using the EF to train the multimodal (bimodal and trimodal) features, the same architecture as that in the unimodal configuration was used, including two middle layers with 128 units for the bimodal features and a layer with 192 units for the trimodal features.

When using the LF to train the multimodal feature set, two layered DNNs were composed. For the lower layer, a neural network with an input layer and two middle layers with 64 units was prepared to extract the unimodal features. For the higher layer, the output units of the unimodal models were concatenated, and the layer with the concatenated units was connected to two hidden layers with 32 units. The concatenated layer had a high dimensional output, and thus, a dropout was implemented after the layer.

In all the DNN models, we set the batch size as 32, total number of epochs as 30, and dropout rate as 0.3. We used the Adam optimizer and set the learning rate as 0.001. For the DNNs, we trained and tested the models three times through random initialization and reported the average accuracy.

5.2 Evaluation procedure

To evaluate the models, the cross-validation method (leave one person out cross-validation, LOPOCV) was performed in the SVM and DNN models. In the LOPOCV, the samples corresponding to each exchange between the participant and dialogue system were used as the test data, and the remaining samples were used as the training data. This procedure ensured that the test data from one participant were completely excluded in the training dataset, thereby avoiding overestimation. We compared the average accuracy of the test data set among the models based on each modality. The majority baseline for the binary classification of the self-sentiment annotation was 54.7%.

6 EXPERIMENTAL RESULT

Table 1 lists the estimation accuracy of the SVM models for the binary classification, and Table 2 lists those of DNN models. We used the following four feature sets to investigate the contribution of physiological signals to estimate the participants' sentiments: P, physiological features; A+P, acoustic + physiological features; P+V, physiological + visual features; A+P+V, fusion of all the features. To analyze the contribution of EDA and HR features, physiological features (P_{EH}) were divided into EDA subset (P_E) and HR subset (P_H), and the estimation accuracy of the models using each feature set was evaluated (rows 4 to 6 and columns 2 to 8 in Table 1 and 2). In addition, acoustic features (A), visual features (V), and acoustic + visual features (A+V) set (columns 9 to 12 in Table 1 and 2) were used for comparison with physiological models.

The EF or LF technique was used to fuse the different modalities, as described in Section 5.1. To investigate the extent to which the human annotators could estimate the participant's positive/negative sentiment labels, the estimation accuracy of the participant's sentiment based on the external sentiment annotation was also evaluated.

Performance of the SVM models: Table 1 lists the estimation accuracy of the SVM models. The unimodal models estimation accuracy are shown in columns 2 (physiological model), 9 (acoustic model) and 10 (visual model) in Table 1. The best unimodal model is the physiological EDA subset (P_E) model (row 5 and column 2 in Table 1) with the accuracy of 61.6%. Comparing the unimodal P_E models to the mutimodal models (columns 3 to 8, 11, and 12 in

Table 1: Binary classification accuracy based on the SVM. The bold value indicates the highest estimation accuracy. The majority baseline was 54.7%. (Uni: unimodal features, Multi: multimodal features, A: acoustic features, P: physiological features, and V: visual features)

Physiological feature set	Uni	Multi						Uni		Multi		Human model
	P	A+P		P+V		A+P+V		A	V	A+V		
		EF	LF	EF	LF	EF	LF			EF	LF	
EDA+HR (P_{EH})	57.7	57.0	60.3	57.5	58.7	56.8	60.2	57.7	58.2	57.1	58.9	63.0
EDA (P_E)	**61.6**	60.4	61.4	60.7	61.2	58.4	61.2					
HR (P_H)	52.5	57.0	55.0	56.7	54.9	56.9	57.1					

Table 2: Binary classification accuracy based on the DNN. The bold value indicates the highest estimation accuracy. The majority baseline was 54.7%. (Uni: unimodal features, Multi: multimodal features, A: acoustic features, P: physiological features, and V: visual features)

Physiological feature set	Uni	Multi						Uni		Multi		Human model
	P	A+P		P+V		A+P+V		A	V	A+V		
		EF	LF	EF	LF	EF	LF			EF	LF	
EDA+HR (P_{EH})	60.1	58.9	58.7	60.5	60.0	59.7	60.1	57.3	57.7	58.4	58.1	63.0
EDA (P_E)	62.2	60.2	59.4	**63.2**	62.9	60.8	61.0					
HR (P_H)	48.6	56.1	55.4	53.7	54.3	55.7	56.9					

Table 1), there is no improvement of estimation accuracy.

Performance of the DNN models: Table 2 presents the accuracy of the binary classification of the DNN models. The unimodal models estimation accuracy are shown in columns 2, 9 and 10 in Table 2 in the same way as Table 1. The best unimodal model is the EDA subset (P_E) model (row 5 and column 2 in Table 2) with the accuracy of 62.2%, which exhibited an improvement of 0.6% compared to the highest SVM models. Comparing the unimodal EDA subset (P_E) models to the mutimodal models (columns 3 to 8, 11 and 12 in Table 2), there is further improvement was observed in the EF of the EDA + visual (P_E+V) model with the estimation accuracy was 63.2% (row 5 and column 5), which meant that this model outperformed the highest performing SVM models by 1.6%.

Accuracy of estimating the participant's sentiment labels by the annotators: Next, we compared the performance of our machine learning models to the "human" model as a benchmark. To this end, the average external sentiment score annotated by five human coders was divided into high and low classes by considering a threshold of 4. We calculated the accuracy of the binary classification of the participant's sentiments based on the external sentiment annotation. The estimation accuracy through human estimation was 63.0% which is higher than that for the highest performing SVM model (P_E model, 61.6%, Table 1). The result (63.2%) of the best DNN model (P_E+V) was equivalent to that of the human annotators. In the next section, we discuss these results in depth.

7 DISCUSSION

As shown in Section 6, the proposed multimodal DNN model achieved an estimation accuracy equivalent to the human performance in the positive/negative sentiment estimation. We further investigated whether the (mis-)classification trend was similar or different between the human model which depends on the explicit information

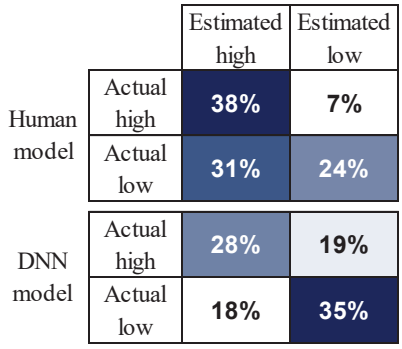

Figure 3: Confusion matrix for binary classification showing the percentage of the total samples (n = 2468). upper: human model, lower: DNN model.

and DNN model which based on implicit biological responses. First, we presented the confusion matrices for the classification results of all the 2468 exchange samples with the "human" model and physiological (P_E) DNN model and compared the results. Second, the classification result of each 26 individuals was considered, and we discussed the differences in the human and machine classification results. Finally, to investigate the physiological features related to the specific outcomes, we performed feature analysis and clarified the physiological factors related to the estimation performance.

7.1 Comparison of Human and Machine

First, to observe the overall classification trend, we evaluated the confusion matrix for the human and physiological (P_E) DNN models (Figure 3). The results showed that there were certain false positives (i.e., misclassified true low into high class) existed in the human estimation (31% of the total sample); however, many positive sentiment labels (true high) were classified as a high class in the human

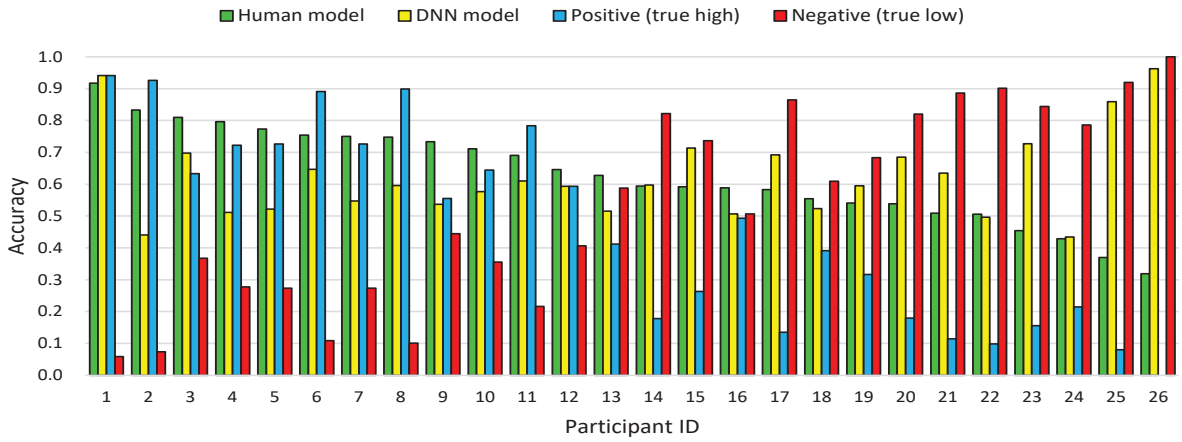

Figure 4: Estimation results for each participant in the LOPOCV (green bar: Human model, yellow bar: DNN model). As a reference, the proportion of the two classes for each participant's sentiment (grand truth) is also shown (blue bar: positive sentiment (true high), red bar: negative sentiment (true low)).

Table 3: Average correlation coefficient r between the EDA feature and sentiment score for all the participants. Bold indicates $r > 0.1$.

Description	r
Standard deviation	**0.157**
Skewness	0.007
Range	**0.161**
Slope of linear approximation	0.075
GSR number	**0.168**

model (38%). In contrast, the DNN model correctly classified many negative sentiment labels (true low) into the low class (35%). This result suggests that the humans could distinguish the participant's positive sentiment labels during the dialogue. To confirm the differences between the human and machine estimation, we evaluated the classification results for each participant. As shown in Figure 4, the humans tended to be more accurate when the participants had a positive sentiment; however, the estimation accuracy was degraded when the participants exhibited a negative sentiment during the dialogue. In contrast, the DNN model classified many negative sentiment labels correctly into the low class, which human models often misclassified. These results suggest that the classification pattern of the human and DNN models is different, even though the total estimation accuracy is comparable. When humans perceive emotions in other people, their perception depends on the explicit acoustic and visual information of the other people, and they cannot detect the physiological implicit state. Thus, it is challenging to estimate the negative or neutral implicit responses of the interactions of the humans. Alternatively, the use of physiological signals or their fusion with other signals could help detect the implicit aspects and estimate the negative sentiment labels for the adaptation of the dialogue systems.

7.2 EDA Feature Analysis

Among the modalities used in this study, the DNN model exhibited that the physiological features are more effective in estimating the participant's sentiments, which change dynamically during dialogues. As the EDA has more effective features compared to those of the HR among the physiological features, we focused on the EDA features and performed an additional analysis. First, to investigate the EDA features that are effective in estimating the participant's sentiment labels, we performed Welch's t-test to verify whether there is a difference between the means of feature of the samples that are classified into high class and the means of those with low class. The results indicated that the standard deviation, skewness, range, and slope of linear approximation of the EDA signals and the GSR number were significantly different for the high and low classes ($p < 10^{-7}$). Subsequently, a correlation analysis was performed between each of the five features and the participant's sentiment score. The average correlation coefficient r between the EDA feature and sentiment score for all the participants was calculated, and it was observed that the GSR number exhibit the highest correlation (Table 3, $r = 0.168$).

Figure 5 presents an example of the time series changes in the sentiment score and the GSR number during the dialogue. It can be noted that the sentiment score is not static but dynamic, and these changes co-occur with the changes in the GSR in this example. This result is reasonable as it is widely recognized that the GSR is related to the human emotional state in the affective computing and psychophysiological domain [13, 22, 27]. This co-occurrence property of the GSR can be applied to estimate the participants' sentiment labels in the DNN model, which exhibits the same performance as that of the human model.

To visualize the relationships between the sentiment score and GSR number, we calculated the quartile of the GSR number, and the samples of the participants' sentiment scores were divided into the quartile group of the GSR number. Figure 6 (upper panel) shows the relative frequency of the sentiment score in each group (Q1: lower quartile, Q4: upper quartile) and indicates the differences in

Figure 5: Example of dynamic changes in the participant's sentiment and GSR number during the dialogue. The sentiment score (blue line, left y axis) and GSR per exchange (orange line, right y axis) are shown. The red stars indicate the timing of the system utterance (Participant ID 11 in Figure 4, dialogue data from the start time to 15th exchange).

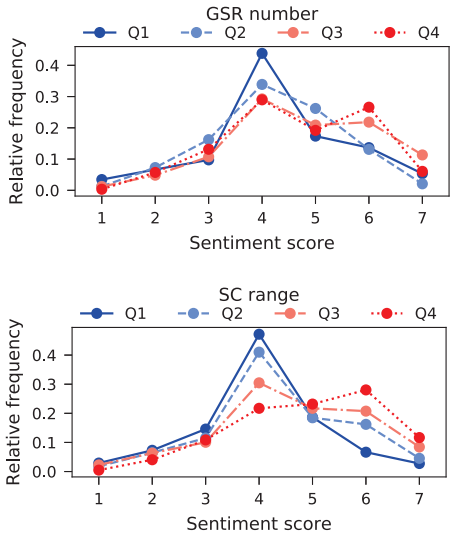

Figure 6: Relationship between the sentiment score of the participant and EDA features in each exchange. The samples of the participants' sentiment score were divided into quartile groups based on the quantile of the GSR number (upper panel) or SC range (lower panel). The relative frequency of the sentiment score in each quartile group is shown.

the sample distribution along with the GSR number. There was a clear difference between quartile groups in the sentiment score of 6. A similar distribution was observed in the quartile group of the SC range (Figure 6, lower panel). Thus, these EDA features were expected to contribute to the sentiment estimation.

7.3 Limitation and Remaining Works

Although the detection of the implicit responses can help develop natural and engaging dialogue systems, it needs real-time feedback to the systems. Therefore, the subsequent objective is to optimize when and how to adapt the systems and to realize automated dialogue adaptation to provide a novel user experience. Weber et al. [36] proposed an autonomous real time adaptation approach that was based on social signals and reinforcement learning in human–robot interaction. A similar feedback approach that can detect the dynamic implicit responses in real time can help realize a more natural and interesting interaction between the user–agent or user–robot. Alternatively, the interbeat interval derived through photoplethysmography is often analyzed in the affective computing or psychophysiological domain. This aspect was not implemented in this work; however, this analysis can provide useful insights regarding autonomic nervous systems. In addition, the presence of individual differences in physiological signals could lead to a performance degradation. A method known as covariate shift adaptation [32], which is based on the density ratio estimation can be used for the domain adaptation in the machine learning domain. Using these methods, the individual differences in physiological signals can be compensated, and the model performance can likely be improved. These aspects will be considered in future work.

8 CONCLUSION

In this study, we collected a new multimodal dialogue corpus Hazumi1911, which included physiological and acoustic/visual signals to investigate the effectiveness of physiological signals in estimating the participant's sentiment at the exchange level. We demonstrated that the SVM model based on physiological signals outperforms the majority baseline and achieves an estimation accuracy of 60.3% when fused with acoustic features. Furthermore, a multimodal DNN model based on the EDA and visual features exhibits an accuracy of 63.2%, which is comparable to the accuracy of sentiment estimation (63.0%) conducted by humans. Although the human and DNN models have similar estimation accuracies, the classification patterns are different. According to the results of the feature analysis, the EDA is correlated with the sentiment score at the exchange level during the dialogue, and thus, detecting these dynamic implicit responses can help in the adaptation of multimodal dialogue systems.

ACKNOWLEDGMENT

This work was partially supported by the Research Program of "Five-star Alliance" in "NJRC Mater & Dev.", and KAKENHI: Grant-in-Aid for Scientific Research (A), Grant No. 19H01120.

REFERENCES

[1] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 59–66.

[2] Dan Bohus and Eric Horvitz. 2009. Learning to Predict Engagement with a Spoken Dialog System in Open-world Settings. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (London, United Kingdom) *(SIGDIAL '09)*. 244–252.

[3] Thierry Chaminade, Léo Biaocchi, Farah H Wolfe, Noël Nguyen, and Laurent Prévot. 2015. Communicative behavior and physiology in social interactions. In *Proceedings of the 1st Workshop on Modeling INTERPERsonal SynchrONy And infLuence*. 25–30.

[4] Chuan-Yu Chang, Jeng-Shiun Tsai, Chi-Jane Wang, and Pau-Choo Chung. 2009. Emotion recognition with consideration of facial expression and physiological signals. In *2009 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*. IEEE, 278–283.

[5] Utkarsh Chauhan, Norbert Reithinger, and John R Mackey. 2018. Real-time stress assessment through PPG sensor for VR biofeedback. In *Proceedings of the 20th International Conference on Multimodal Interaction: Adjunct*. 1–5.

[6] Juan Abdon Miranda Correa, Mojtaba Khomami Abadi, Niculae Sebe, and Ioannis Patras. 2018. Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing* (2018).

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

[8] Olga Egorow and Andreas Wendemuth. 2016. Detection of challenging dialogue stages using acoustic signals and biosignals. (2016).

[9] E Friesen and Paul Ekman. 1978. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto* 3 (1978).

[10] Yuki Hirano, Shogo Okada, Haruto Nishimoto, and Kazunori Komatani. 2019. Multitask Prediction of Exchange-level Annotations for Multimodal Dialogue Systems. In *2019 International Conference on Multimodal Interaction*. 85–94.

[11] Koji Inoue, Divesh Lala, Katsuya Takanashi, and Tatsuya Kawahara. 2019. Latent character model for engagement recognition based on multimodal behaviors. In *9th International Workshop on Spoken Dialogue System Technology*. Springer, 119–130.

[12] Kyriaki Kalimeri and Charalampos Saitis. 2016. Exploring multimodal biosignal features for stress detection during indoor mobility. In *Proceedings of the 18th ACM international conference on multimodal interaction*. 53–60.

[13] Jonghwa Kim and Elisabeth André. 2008. Emotion recognition based on physiological changes in music listening. *IEEE transactions on pattern analysis and machine intelligence* 30, 12 (2008), 2067–2083.

[14] Jonghwa Kim, Elisabeth André, Matthias Rehm, Thurid Vogt, and Johannes Wagner. 2005. Integrating information from speech and physiological signals to achieve emotional sensitivity. In *Ninth European Conference on Speech Communication and Technology*.

[15] R Benjamin Knapp, Jonghwa Kim, and Elisabeth André. 2011. Physiological signals and their use in augmenting emotion recognition for human–machine interaction. In *Emotion-oriented systems*. Springer, 133–159.

[16] Kazunori Komatani and Shogo Okada. 2019. Collection and Analysis of Human-System Multimodal Dialogue Data with Subjective Ratings. *IEICE Technical Report (in Japanese)* 119, 179 (2019), 21–26.

[17] Kazunori Komatani and Shogo Okada. 2020. Osaka University Multimodal Dialogue Corpus (Hazumi). Informatics Research Data Repository, National Institute of informatics. (dataset). https://doi.org/10.32130/rdata.4.1

[18] Kazunori Komatani, Shogo Okada, Haruto Nishimoto, Masahiro Araki, and Mikio Nakano. 2019. Multimodal Dialogue Data Collection and Analysis of Annotation Disagreement. In *Proc. International Workshop on Spoken Dialogue Systems (IWSDS)*.

[19] Jukka Kortelainen, Suvi Tiinanen, Xiaohua Huang, Xiaobai Li, Seppo Laukka, Matti Pietikäinen, and Tapio Seppänen. 2012. Multimodal emotion recognition by combining physiological signals and facial expressions: a preliminary study. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 5238–5241.

[20] Richard D Lane, Kateri McRae, Eric M Reiman, Kewei Chen, Geoffrey L Ahern, and Julian F Thayer. 2009. Neural correlates of heart rate variability during emotion. *Neuroimage* 44, 1 (2009), 213–222.

[21] Iulia Lefter and Siska Fitrianie. 2018. The Multimodal Dataset of Negative Affect and Aggression: A Validation Study. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 376–383.

[22] Dominik Leiner, Andreas Fahr, and Hannah Früh. 2012. EDA positive change: A simple algorithm for electrodermal activity to measure general audience arousal during media exposure. *Communication Methods and Measures* 6, 4 (2012), 237–250.

[23] Yukiko I. Nakano and Ryo Ishii. 2010. Estimating User's Engagement from Eye-gaze Behaviors in Human-agent Conversations. In *Proceedings of the 15th International Conference on Intelligent User Interfaces* (Hong Kong, China) *(IUI '10)*. 139–148.

[24] Phuong Pham and Jingtao Wang. 2018. Adaptive review for mobile mooc learning via multimodal physiological signal sensing-a longitudinal study. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 63–72.

[25] Rosalind W Picard. 2000. *Affective computing*. MIT press.

[26] Rosalind W Picard. 2016. Automating the recognition of stress and emotion: From lab to real-world impact. *IEEE MultiMedia* 23, 3 (2016), 3–7.

[27] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. 2001. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence* 23, 10 (2001), 1175–1191.

[28] Yuning Qiu, Teruhisa Misu, and Carlos Busso. 2019. Driving Anomaly Detection with Conditional Generative Adversarial Network using Physiological and CAN-Bus Data. In *2019 International Conference on Multimodal Interaction*. 164–173.

[29] Hiranmayi Ranganathan, Shayok Chakraborty, and Sethuraman Panchanathan. 2016. Multimodal emotion recognition using deep learning architectures. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–9.

[30] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. 2018. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 400–408.

[31] Björn Schuller, Stefan Steidl, and Anton Batliner. 2009. The interspeech 2009 emotion challenge. In *Tenth Annual Conference of the International Speech Communication Association*.

[32] Hidetoshi Shimodaira. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference* 90, 2 (2000), 227–244.

[33] Ramanathan Subramanian, Julia Wache, Mojtaba Khomami Abadi, Radu L Vieriu, Stefan Winkler, and Nicu Sebe. 2016. ASCERTAIN: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing* 9, 2 (2016), 147–160.

[34] Leili Tavabi, Kalin Stefanov, Setareh Nasihati Gilani, David Traum, and Mohammad Soleymani. 2019. Multimodal Learning for Identifying Opportunities for Empathetic Responses. In *2019 International Conference on Multimodal Interaction*. 95–104.

[35] Sayaka Tomimasu and Masahiro Araki. 2016. Assessment of users' interests in multimodal dialog based on exchange unit. In *Proceedings of the Workshop on Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction*. 33–37.

[36] Klaus Weber, Hannes Ritschel, Ilhan Aslan, Florian Lingenfelser, and Elisabeth André. 2018. How to shape the humor of a robot-social behavior adaptation based on reinforcement learning. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 154–162.