JAIST Repository
https://dspace.jaist.ac.jp/

Title	ロバストな生体信号処理に基づくオンライン感性推定とマルチモーダル統合への応用
Author(s)	堅田, 俊
Citation	
Issue Date	2022-12
Type	Thesis or Dissertation
Text version	ETD
URL	http://hdl.handle.net/10119/18187
Rights	
Description	Supervisor:岡田 将吾, 先端科学技術研究科, 博士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Doctoral Dissertation

Robust Physiological Signal Processing for Online Sentiment Estimation and its Application for Multimodal Fusion

Shun Katada

Supervisor: Shogo Okada

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information science)

December 2022

Abstract

For modeling human intelligence, understanding emotional intelligence is an important and challenging issue. In affective computing, it has been reported that not only text, acoustic, and visual signals (observable signals) but also physiological signals (unobservable signals) are useful for estimating emotions and their related states. Physiological signals are expected to provide additional and less biased information compared with observable signals. Thus, coupled with growing interest in the development of emotionally intelligent systems, many studies related to physiological signals have been reported thus far; however, techniques that apply physiological signals for realistic emotion estimation tasks such as online (sequential) recognition for dialogue systems are still in the research phase, and there are unresolved issues in fundamental and applied research. In this thesis, three main research problems that have not previously been explored are addressed.

First, as one of the fundamental unresolved issues, physiological signals have individual differences that cause performance degradation of machine learning models based on physiological signals. Generally, it is assumed that both training and test data for machine learning are derived from the same distribution. Thus, estimation performance can degrade if there are physiological individual differences in unseen individual test data. In this thesis, physiological individual differences are considered a covariate shift to resolve this problem, and the Importance-Weighting (IW) method is introduced, which complements the model and is robust against individual differences for performance improvement of the models trained with physiological data. As a result, Importance-Weighted Support Vector Machine (IW-SVM) models outperform conventional models based on physiological features in emotion and personality estimation. These results indicate that IW in machine learning models can reduce the effects of physiological individual differences in physiological responses and contribute to the proposal of a new model for emotion and personality estimations based on physiological signals.

Second, although fundamental research on physiological signals provides insight into their potential, the effectiveness of physiological signals is often evaluated under emotion-evoked conditions. Thus, few studies have analyzed physiological signal effectiveness in naturalistic conditions. In particular, the evaluation and comparison of physiological signals with other observable signals under naturalistic human-agent interactions are insufficient. In human-agent interactions, it is necessary for the systems to identify the current internal state of the user to adapt their dialogue strategies. Nevertheless,

this task is challenging because the current user's sentiment is not always expressed by observable signals in a natural setting and changes dynamically. However, it is possible that physiological signals provide valuable information for online sentiment estimation since physiological responses cannot be consciously regulated. As applied research, a machine learning model based on physiological signals to estimate a user's sentiment at every exchange during a dialogue is presented in this thesis. Using a wearable sensing device, the physiological data including the Electrodermal Activity (EDA) and Heart Rate (HR) in addition to acoustic and visual information during a dialogue are evaluated. The sentiment labels are annotated by the participants (referred to as Self-repoted Sentiment (SS) label) for each exchange consisting of a pair of system and participant utterances. The experimental results show that a multimodal Deep Neural Network (DNN) model combined with the EDA and visual features achieves an accuracy of 63.2%. The analysis of the SS estimation results for each individual indicate that the human coders often incorrectly estimate negative sentiment labels, and in this case, the performance of the DNN model is 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.

Finally, although the potential of the physiological signals in online SS estimation during dialogue is clarified in the abovementioned task, there is no comprehensive and thorough analysis of physiological signal application for multimodal fusion. Thus, the second task is extended by introducing different types of sentiment labels (annotated by third-party), which further clarify the contributions of physiological signals. Additionally, two state-of-the-art language models and six machine learning models, including recently reported multimodal DNN, are introduced. Furthermore, these analyses enable the creation of a robust multimodal physiological model that combines the proposed physiological signal processing method and the Transformer language model, named Time-series Physiological Transformer (TPTr). This model can capture sentiment changes based on both time-series linguistic and physiological information. In ensemble models, the proposed methods significantly outperform the previous best result ($p < 0.05$). These results provide new insight into machine learning methods that utilize both linguistic information and physiological responses during dialogue exchanges, which has not previously been explored.

In summary, this thesis presents novel robust physiological signal processing for emotion/sentiment estimation and its application to adaptive dialogue systems. This proposal will lead to a new application of physiological signals that are widely applicable in various fields. For example, the educational sys-

tem can capture the concentration level of students by monitoring students' internal states, and the psychological counseling system can be supported by understanding the context behind words. These emotionally intelligent systems will provide significant improvements in our lives in the future.

Keywords: Sentiment Analysis; Physiological Signal Processing; Machine Learning; Multimodal Signal Processing; Dialogue System.

Acknowledgements

First, I would like to sincerely thank Associate Professor Shogo Okada at the Japan Advanced Institute of Science and Technology for his meticulous guidance, valuable and constructive discussions, and encouragement through my doctoral program. His supervision enabled me to maintain strong motivation toward the information science field for five years.

Additionally, I am grateful to Professor Kazunori Komatani at the Institute of Scientific and Industrial Research (SANKEN), Osaka University, for collaborating on this study. His excellent detailed guidance, which included a specific solution to a problem, made our research more sophisticated and led to high performance in the dialogue system field.

I would like to thank Associate Professor Kiyoaki Shirai at the Japan Advanced Institute of Science and Technology for reviewing my thesis and minor research project in the natural language processing field. I would like to thank Professor Shinobu Hasegawa and Associate Professor Naoya Inoue at the Japan Advanced Institute of Science and Technology for reviewing my thesis with thoughtful and rational comments.

Finally, I thank all of my colleagues at Okada's laboratory (Tokyo and Ishikawa), and my family for their support.

Contents

Abstract i
Acknowledgement iv
Contents v
List of Figures viii
List of Tables ix
List of Abbreviations x

1 General Introduction **1**

2 Related Works **7**
2.1 Emotion and Sentiment 7
2.2 Physiological Signals 8
2.3 Multimodal Sentiment Analysis 9

3 Biosignal-based Emotion Recognition with Importance Weighting **12**
3.1 Introduction 12
3.2 Related Works 14
3.2.1 Databases for Emotion and Personality Research . . . 14
3.2.2 Application of Importance Weighting 15
3.2.3 Performance Comparison of Previous Methods 16
3.3 Methods 18
3.3.1 Dataset 18
3.3.2 Preprocessing and Feature Extraction 19
3.3.3 Machine Learning Model 21
3.4 Experiments 24
3.4.1 Experimental Settings for the Classification Task . . . 24
3.4.2 Results 25
3.5 Discussion 28
3.5.1 Analysis of the Classification Performance by Estimating BER 28
3.5.2 Feature Analysis 29

3.5.3 Performance Comparison with a Previous Report . . . 34
3.5.4 Performance Comparison with a Deep Neural Network 35
3.5.5 Importance-weighted Support Vector Machine in Speech Emotion Recognition . . . 35
3.5.6 Computational Complexity and Hyperparameters . . . 36
3.5.7 Limitations and Future Works . . . 36
3.6 Chapter Summary . . . 37

4 Analysis of Physiological Signals toward Adaptive Dialogue Systems 38
4.1 Introduction . . . 38
4.2 Related Works . . . 39
4.3 Data . . . 41
4.3.1 Data Collection . . . 41
4.3.2 Participants . . . 43
4.3.3 Annotation . . . 43
4.4 Multimodal Feature Extraction . . . 44
4.4.1 Physiological Features . . . 44
4.4.2 Acoustic and Visual Features . . . 45
4.5 Experiment . . . 46
4.5.1 Machine Learning Models . . . 46
4.5.2 Evaluation Procedure . . . 47
4.6 Experimental Result . . . 47
4.7 Discussion . . . 51
4.7.1 Comparison of Human and Machine . . . 51
4.7.2 EDA Feature Analysis . . . 52
4.7.3 Limitation and Remaining Works . . . 54
4.8 Chapter Summary . . . 55

5 Different Types of Multimodal Sentiment Estimation 56
5.1 Introduction . . . 56
5.2 Related Works . . . 57
5.2.1 Text-based Sentiment Analysis . . . 57
5.2.2 Physiological Signal-based Sentiment Analysis . . . 58
5.2.3 Multimodal Dialogue Systems . . . 60
5.3 Data . . . 61
5.3.1 Dialogue Settings . . . 61
5.3.2 Sensors . . . 62
5.3.3 Annotation . . . 63
5.4 Features and Representations . . . 64
5.4.1 Linguistic Feature Extraction . . . 64

5.4.2 Physiological Feature Extraction . . . 66
5.4.3 Audio/Visual Feature Extraction . . . 67
5.5 Experimental Settings . . . 67
5.5.1 Machine Learning Algorithms . . . 67
5.5.2 Evaluation Procedure . . . 70
5.6 Results . . . 70
5.6.1 Self-reported Sentiment Estimation . . . 71
5.6.2 Third-party Sentiment Estimation . . . 75
5.6.3 Estimation with Automatic Speech Recognition . . . 79
5.6.4 Analysis of Physiological Features . . . 79
5.6.5 Comparison of Linguistic and Physiological Models . . 82
5.7 Discussion . . . 83
5.8 Chapter Summary . . . 87

6 Multimodal Transformer with Physiological Signals 88
6.1 Introduction . . . 88
6.2 Related Works . . . 90
6.3 Proposed Methods . . . 92
6.3.1 Time-Series Physiological Signal Processing . . . 92
6.3.2 Time-Series Modeling of Physiological Signals . . . 93
6.4 Experimental Settings . . . 96
6.4.1 Baselines and Hyperparameters . . . 96
6.4.2 Evaluation Procedure . . . 98
6.4.3 Dataset . . . 98
6.5 Results and Discussion . . . 99
6.5.1 Performance of Physiological LSTM Models . . . 100
6.5.2 Performance of TPTr . . . 101
6.5.3 TPTr Based on Other Submodalities . . . 102
6.5.4 Analysis of the Attention Weight . . . 104
6.5.5 Analysis of the Exchange-Level Estimation Pattern . . 106
6.5.6 Limitations and Future Works . . . 106
6.6 Chapter Summary . . . 107

7 Conclusion 108
Bibliography . . . 111
Appendix . . . 129
Publication List . . . 134

List of Figures

1.1 Self-repoted Sentiment (SS) and Third-party Sentiment (TS) . 2
1.2 Summary of this study 5

3.1 Overview 18
3.2 An example of the analysis of the ECG data 20
3.3 An example of the analysis of the GSR data 21
3.4 Analysis of the classification performance by estimating BER . 29

4.1 Overview 42
4.2 Example of SC measured during the conversation 45
4.3 Confusion matrix for binary classification 52
4.4 Estimation results for each participant in the LOUOCV 52
4.5 Example of dynamic changes in the sentiment and GSR number 54
4.6 Relationship between the sentiment score and EDA features . 54

5.1 Overview 61
5.2 Distribution of the sentiment scores 64
5.3 Summary of the multimodal model architecture 68
5.4 t-SNE visualization with test samples 82
5.5 Confusion matrix for classification 83
5.6 Differences in the estimation results between the models . . . 84

6.1 An approach for capturing SS changes 89
6.2 Conventional and proposed models architectures 94
6.3 Attention weights extracted from Transformer models 105
6.4 Estimation pattern of conventional and proposed models . . . 107

List of Tables

2.1 Multimodal dataset for sentiment/emotion analysis 10

3.1 Emotion recognition results reported in previous studies . . . 16
3.2 Personality recognition results reported in previous studies . . 16
3.3 List of ECG features 19
3.4 Performance comparison in the emotion estimation 26
3.5 Performance comparison in the personality estimation 27
3.6 Contribution of each feature for emotion estimation 30
3.7 Contribution of each feature evaluated by a stepwise method . 31
3.8 Contribution of each feature for personality estimation 33

4.1 Binary classification accuracy based on the SVM 50
4.2 Binary classification accuracy based on the DNN 50
4.3 Correlation between the EDA feature and sentiment score . . 53

5.1 Dataset summary 62
5.2 SS estimation results for unimodal models 72
5.3 SS estimation results for multimodal models 74
5.4 TS estimation results for unimodal models 76
5.5 TS estimation results for multimodal models 78
5.6 Estimation results for models trained on ASR data 80
5.7 Analysis of physiological features with the stepwise method . . 81

6.1 Sentiment estimation results of physiological LSTM models . . 101
6.2 Performance of the conventional and proposed models 102
6.3 TPTr model performance based on physiological submodalities 103

S1 User utterance distribution 130
S2 SS estimation results with handcrafted features or fastText . . 132
S3 TS estimation results with handcrafted features or fastText . . 133

List of Abbreviations

Acc	Accuracy.
ANS	Autonomic Nervous System.
ASR	Automatic Speech Recognition.
BCI	Brain Computer Interface.
BER	Bayesian Error Rate.
BERT	Bidirectional Encoder Representations from Transformers.
BiLSTM	Bidirectional Long Short-Term Memory.
BVP	Blood Volume Pulse.
CMTr	CrossModal Transformer.
CNN	Convolutional Neural Network.
Corr	Pearson correlation coefficient.
DNN	Deep Neural Network.
ECG	Electrocardiogram.
EDA	Electrodermal Activity.
EEG	Electroencephalogram.
EF	Early Fusion.
EMG	Electromyogram.
EOG	Electrooculogram.
F1	macro F1-score.
FACS	Facial Action Coding System.
FNN	Feedforward Neural Network.
GRU	Gated Recurrent Unit.
GSR	Galvanic Skin Response.
HCI	Human-Computer Interaction.
HiF	High Frequency.
HR	Heart Rate.
HRV	Heart Rate Variability.
IW	Importance-Weighting.
IW-LR	Importance-Weighted Logistic Regression.
IW-SVM	Importance-Weighted Support Vector Machine.
KuLSIF	Kernelized variant of uLSIF.

L-SVM	Linear Support Vector Machine.
LF	Late Fusion.
LoF	Low Frequency.
LOUOCV	Leave-One-User-Out Cross-Validation.
LR	Logistic Regression.
LSTM	Long Short-Term Memory.
MAE	Mean Absolute Error.
MAG	Multimodal Adaptation Gate.
MulT	Multimodal Transformer.
NB	Naive Bayes.
NN50	Total number of successive R-wave interval differences that differ by more than 50 ms.
pNN50	percentage value of NN50.
R-SVM	Support Vector Machine with Radial basis function kernel.
RMSSD	Root Mean Square of Successive R-wave interval Differences.
RNN	Recurrent Neural Network.
RRI	R-wave Interval.
SC	Skin Conductance.
SOTA	State-Of-The-Art.
SS	Self-repoted Sentiment.
SVM	Support Vector Machine.
TEMP	Skin temperature.
TFN	Tensor Fusion Network.
TPTr	Time-series Physiological Transformer.
Tr	Transformer.
TS	Third-party Sentiment.
uLSIF	unconstrained Least-Squares Importance Fitting.
vLoF	very Low Frequency.

Chapter 1

General Introduction

Affective computing is defined as computing that relates to, arises from, or influences emotions [1]. The global affective computing market size gradually increased and was valued at USD 20.23 billion in 2019 [2], suggesting that not only researchers but also many developers and ordinary people focus on this research-related domain. The development of emotion research is expected to provide many valuable applications, such as mental health care and educational support; thus, the demand for emotionally intelligent agents will further increase.

The estimation of a user's sentiment during a dialogue is one of the most fundamental concerns in the affective computing field. Although user sentiment estimation per session (i.e., batch processing) is important, user sentiment estimation per exchange (i.e., online processing) is needed to create an adaptive dialogue system. In fact, as a user's sentiment states can change dynamically during dialogues, it is necessary to capture the dynamic sentiment changes in real time. For example, if the user is interested in the current topic, the dialogue system should continue as is, whereas if the user is bored, the system should change the current topic. In pursuit of the realization of emotionally intelligent agents, this simple but challenging task has been considered by many researchers using a variety of approaches [3].

The textual lexicon-based approach has long been mainstream in sentiment analysis. With the spread of social media platforms, which utilize not only text but also images and videos, the effectiveness of multimodal analysis has been extensively investigated in recent years [4, 5]. The technique of fusing verbal and nonverbal information to differentiate sentiment is called multimodal sentiment analysis [6, 7]. Textual, visual and audio features have different characteristics and complement each other for sentiment analysis, as shown in [6].

However, to realize adaptive dialogue systems, several problems remain

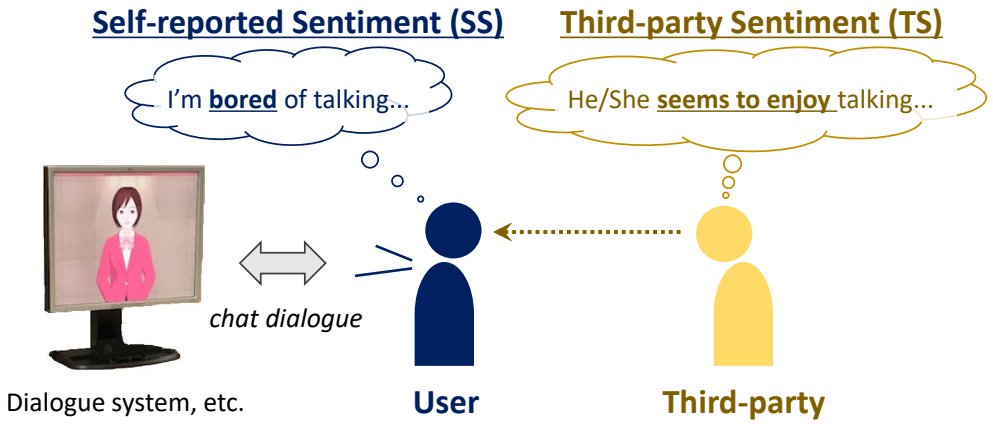

Figure 1.1: Self-repoted Sentiment (SS) and Third-party Sentiment (TS)

to be resolved in multimodal sentiment analysis research.

First, ideally, sentiment labels annotated by users (hereafter referred to as Self-repoted Sentiment (SS) labels) should be used for sentiment analysis, but many previous works use sentiment labels annotated by a third party (hereafter referred to as Third-party Sentiment (TS) labels) [8, 9, 10]. The difference between SS and TS is illustrated in Figure 1.1. In this example, SS is depicted as a thought balloon with his or her negative sentiment (bored). However, a third-party thinks that he or she seems to enjoy talking based on observable user behavior, such as utterances and facial expressions. Of course, third-party cannot access the user's thought balloon in the annotation process; therefore, SS and TS labels are not always consistent. In other words, TS does not completely reflect a user's true sentiment in his or her mind. In fact, Truong *et al.* showed discrepancies between SS and TS (observed emotion ratings) [11].

Second, related to above-mentioned problem, users' emotional states are not always expressed as true sentiments since users can mask or modify their true feelings during a dialogue episode. Textual, audio, and visual information of the user is known to be useful for TS estimation [7]. Models based on this observable information achieve high estimation performance for TS because TS labels are based on observable signals (i.e., text, audio, visual signals) from the user and are correlated with those observable signals. However, there is no guarantee that models based on observable signals achieve similar performance in "SS" estimation.

Furthermore, most works considering physiological signals have explored the ability to capture emotions under induced visual emotional stimuli, and

few studies have investigated whether signals detected in relatively short episodes (approximately 10 seconds) are effective for SS and TS estimation in human-agent interaction settings. Thus, the effects of physiological signals that change quickly under naturalistic conditions on sentiment estimation remain unclear.

Physiological signals can be used to estimate sentiments because these signals are closely related to the states of the Autonomic Nervous System (ANS). The ANS 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 [12]. 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 Electrodermal Activity (EDA) is another representation of physiological changes and has been widely used in emotion-related research [13, 14]. 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 [13]. In addition, a correlation has been reported between regional cerebral blood flow measured using positron emission tomography and HR variability in emotion-evoking stimuli [15]. 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.

However, there are physiological individual differences in applying physiological data to the development of a machine learning model. Generally, when using data from individuals, the test data from one user should be completely excluded from the training dataset. For example, Leave-One-User-Out Cross-Validation (LOUOCV) should be employed to evaluate the machine learning model. Obviously, this user-independent evaluation schema is also important in emotion recognition tasks. On the other hand, estimation performance of the models would decrease if the emotional physiological responses between users are different. This problem, in which the training (source) data are biased and potentially nonrepresentative, is known as a covariate shift. [16]. It is necessary to consider covariate shifts in the LOUOCV schema whenever the dataset includes data, such as physiological signals, that have individual differences. This fundamental problem of physiological signals has not been resolved in previous affective computing studies; thus, there is a need to create appropriate emotion estimation model-based physiological signals that are robust against individual differences.

Sentiment estimation during human-agent interaction is one of many applied studies of affective computing, and realizing an adaptive dialogue sys-

tem is one of the ultimate goals of research and development of dialogue systems. It is possible that physiological signals may resolve the above-mentioned issues of multimodal sentiment analysis, i.e., physiological signals may be useful for SS estimation by capturing subtle physiological changes even in naturalistic human-agent interactions; however, although physiological signals have the potential for emotion estimation to complement models based on textual, visual and audio features, many previous studies have evaluated the effectiveness of physiological signals under emotional stimuli. Thus, evaluation of the physiological signals under naturalistic conditions such as a chat-dialogue (i.e., using the real text, audiovisual, and physiological data collected in naturalistic conditions simultaneously) is clearly lacking. Therefore, physiological signals have unresolved issues from the point of view of applied research.

In this doctoral thesis, above-mentioned fundamental problem in physiological signal processing and problems in physiological signal processing for dialogue as applied research are addressed for the development of adaptive dialogue systems. The problems and their solutions are summarized as follows:

1. **Physiological individual differences**

   Problem: Physiological individual differences, which are the cause of the performance degradation in emotion estimation, exist in applying physiological data to the machine learning model.

   Solution: Using the covariate shift adaptation (also referred to as Importance-Weighting (IW)) technique, a machine learning model that is robust against physiological individual differences is proposed to estimate an individual's emotion and personality (Chapter 3).

2. **Physiological signal processing for dialogue**

   Problem: The effectiveness of the physiological data has been evaluated under the emotion-evoking condition. The potential of physiological signals for online (i.e., exchange-level) multimodal sentiment estimation remains unknown. Additionally, TS has often been used for multimodal sentiment analysis and has not been evaluated SS simultaneously, which is deeply involved in physiological response.

   Solution: The effectiveness of the physiological signals is evaluated using data collected in naturalistic human-agent interaction in real-time and compared with other modalities in SS estimation (Chapter 4).

3. **Comprehensive analysis and model proposal**

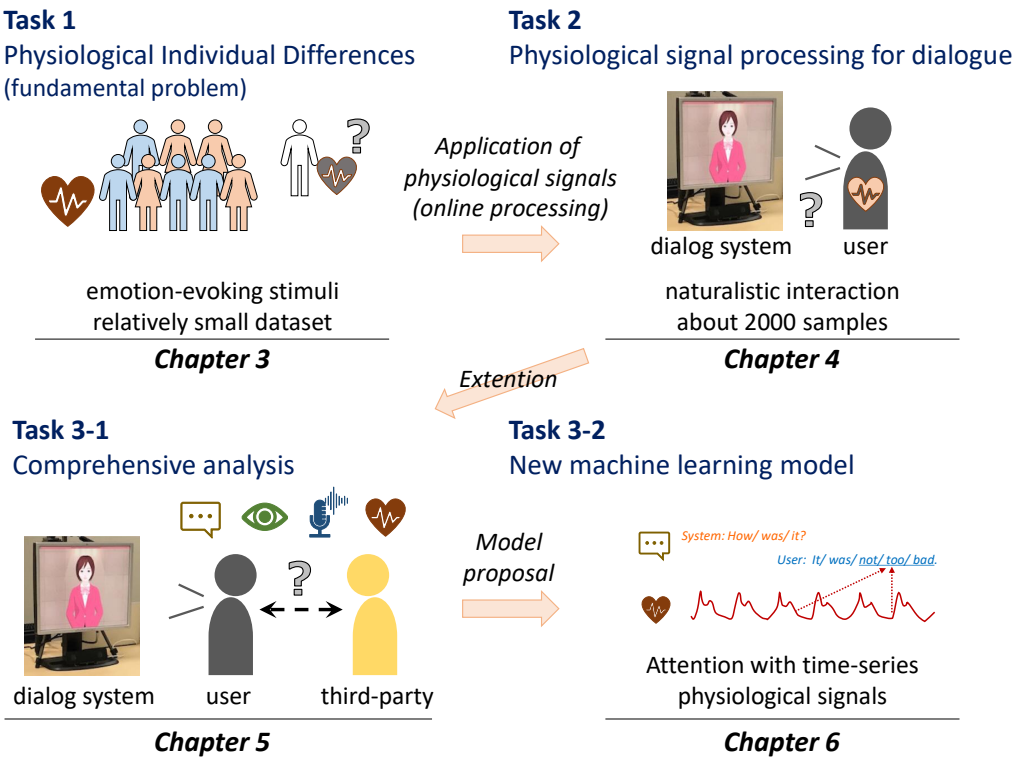

Figure 1.2: Summary of this study.

Problem: Although the potential of physiological signals in online SS estimation is clarified in Chapter 4, there is no comprehensive and thorough analysis of physiological signal application for multimodal fusion, enabling the creation of a robust multimodal physiological model.

Solution: As an extension of Chapter 4, an evaluation of the machine learning model based on the data collected in human-agent interaction settings, including text, acoustic, visual and physiological features represented by State-Of-The-Art (SOTA) representation models, is performed (Chapter 5). Based on this comprehensive and thorough analysis, new time-series physiological signal processing methods are proposed for online SS estimation (Chapter 6).

A summary of this doctoral thesis is depicted in Figure 1.2. Task 1 is positioned in fundamental research (Chapter 3). Task 2 is positioned in applied research (Chapter 4) and further extended by Chapter 5. Finally, new physiological signal processing for online sentiment estimation that is robust against changes in sentiment state is proposed (Chapter 6). Chapter 2 describes an

overview of the physiological signals, Chapters 3 to 6 present each task, and Chapter 7 concludes this study.

Chapter 2

Related Works

This chapter describes a brief explanation of emotion, sentiment, physiological signals and multimodal sentiment analysis. The definitions of "Emotion" and "Sentiment" are described in Section 2.1, since the former is related to Task 1 (Chapter 3) and the latter is related to Tasks 2 and 3 (Chapter 4 to 6). Section 2.2 describes the physiological signals involved in all of the work in this thesis. A basic explanation of the physiological signals and their application to the machine learning reported previously are presented. Finally, recently developed multimodal datasets that are closely related to Tasks 2 and 3 (Chapters 4 to 6) are presented in Section 2.3. An overview of recently developed multimodal datasets clarifies the difference between this thesis and previous studies. More detailed related works are described later in each chapter.

2.1 Emotion and Sentiment

Emotions are a collection of psychological states [17], including subjective experiences, expressive behaviors such as facial expressions and gestures, and physiological signals such as HR and EDA. Six basic emotion categories are proposed by Ekman, i.e., anger, disgust, fear, happiness, sadness, and surprise [18]. In the field of affective computing, emotion detection, modeling and practical applications have been actively investigated [1]. Data analysed in Chapter 3 include the data collected during movie watching as emotion stimuli; thus, the words "emotion" and "emotion estimation" are used for describing this study.

In contrast, "sentiment" refers to an emotional disposition, i.e., a tendency to have a particular type of affective experience (e.g., positive or negative) [19]. Unlike evoked emotions, which typically have external mani-

festations, sentiment is not necessarily expressed explicitly [4]. Moreover, the expression of emotion or sentiment is regulated by emotional intelligence [20], influenced by personality [21], and dependent on context [22]. Hence, estimating the true sentiment in a person's mind is a challenging and interdisciplinary research task that includes elements of psychology and social science. In naturalistic conditions such as chat-dialogue settings, the word "sentiment" is appropriate since there are no explicitly evoked emotions such as fear and sadness. Thus, the words "sentiment" and "sentiment estimation" are used in Chapter 4 to 6.

2.2 Physiological Signals

Linguistic, visual and audio information are essential to building computers that can recognize and express emotions. Additionally, since it is difficult to control physiological responses by oneself, physiological signals have the advantage of extracting implicit emotional responses [23]. For example, when a person is watching a movie or engaged in a video game, it is difficult to measure their interest and immersion level by only their appearance because their emotion might not be shown explicitly. In addition, with the development of physiological signal sensing devices [24], a feedback system that processes and interprets a user's biological responses and returns an appropriate response to the user has been proposed in the Human-Computer Interaction (HCI) [25] and Brain Computer Interface (BCI) [26] domains. Thus, physiological signals are expected to be useful and compensate for other modalities in emotion/sentiment estimation.

The use of physiological signals with machine learning methods has been considered to estimate a user's internal state [14, 27, 28]. Kim and André [27] proposed an emotion recognition approach that involved measured physiological signals, including Electroencephalogram (EEG), Electrocardiogram (ECG), EDA, Electromyogram (EMG) and respiration. In a 2D emotion (valence and arousal) classification task in four quadrants, they achieved an accuracy of 70% for subject-independent classification by exploiting dichotomous categorizations. AlZoubi et al. [29] utilized ECG, EMG and EDA data to detect emotions during interactions between participants and tutoring systems. Participants were instructed to retrospectively report their affective states during 20-second intervals, and these data were used as class labels. Machine learning methods, including Support Vector Machine (SVM), K-nearest neighbor, and Naive Bayes (NB), have been used for classification tasks in previous studies. Recently, deep learning methods have also been applied for emotion recognition tasks with physiological signals [30, 31, 32].

These studies revealed that physiological signals can be used for emotion recognition and implied that physiological signals can effectively complement one another in multimodal approaches.

However, relative to linguistic, visual and audio information, effectiveness of the physiological information in the multimodal approach has not been comprehensively evaluated. In particular, the role of physiological signals in human-agent interaction settings has rarely been investigated. Thus, a thorough and comprehensive analysis of physiological signals is needed to understand their role and appropriate fusing methods for multimodal emotion/sentiment estimation.

2.3 Multimodal Sentiment Analysis

One of the key techniques for extracting a user's sentiment from information other than linguistic information is nonverbal information processing. Humans communicate with each other using not only natural language but also nonverbal behaviors such as facial expressions [33, 34], vocal behavior [35] and gestures [36]. For facial expressions, Ekman and Friesen developed Facial Action Coding System (FACS), which enabled emotion mapping [37] and has been used for numerous affective computing studies. For vocal behavior, acoustic information such as loudness, pitch, and rhythm are expressions of emotion. Similar to facial expression, the relationship between vocal behavior and emotion and their modeling have been extensively investigated [35, 38]. For gestures, although there are a few gesture-based emotion studies relative to facial expression and vocal behavior, they are also related to emotional expression. For example, high-frequency hand clapping expresses joy and satisfaction [39]. These nonverbal behavioral cues are called social signals [38]. Social signal processing is often used to construct automatic user state estimation models for adaptive dialogue systems [40, 41, 42]. For instance, facial expressions, body gestures and prosody are frequently used as nonverbal information for sentiment analysis [3, 4, 5].

Since the 2000s, many multimodal datasets for sentiment/emotion analysis have been created. Table 2.1 lists multimodal datasets that are closely related to this thesis. Busso et al. [10] created a corpus that includes the facial expressions and gestures of actors during scripted and unscripted spoken communication called IEMOCAP. This corpus is widely used for a joint analysis of speech and gestures and contributes to the progress of multimodal analysis. As shown in the 3rd column in Table 2.1, the emotion type of this dataset is "acted". Emotion research based on emotion expressed by the actor has merit in terms of obtaining the ground-truth label. It enables the

Table 2.1: Multimodal dataset for sentiment/emotion analysis. "Type" indicates sentiment/emotion type, i.e., (act)ed, (eli)cited, or (nat)uralistic. L, language; A, audio; V, visual, P, physiological; #Sub, number of subjects; #Samp, number of samples; (con), continuous ratings.

Dataset	Domain	Type	Modality	Label	#Sub	#Samp
IEMOCAP [10]	Scripted & unscripted action	act	L, A, V	TS	10	10037
SEMAINE [43]	Human-agent interaction	eli	A, V	TS	150	(con)
RECOLA [44]	Video conference	nat	A, V, P	TS	46	(con)
MOSI [8]	Movie review (monologue)	nat	L, A, V	TS	98	2199
MOSEI [9]	Open-domain (monologue)	nat	L, A, V	TS	1000	23453
MELD [45]	TV-series (multi-party)	act	L, A, V	TS	407	13708
Hazumi1911 [46]	Human-agent interaction	nat	L, A, V, P	SS, TS	30	2859

correct association of the expressed emotion and label. However, there is a large difference between acted and naturalistic emotions. In fact, performance improvement by the multimodal approach in the acted dataset is three times higher than naturalistic one [47]; therefore, there may be overestimation of the effectiveness of the multimodal model and may not be suitable for research aimed at realizing naturalistic HCI, such as chit-chat dialogue systems. The SEMAINE dataset [43] includes audiovisual information in human-agent interaction settings. The RECOLA dataset [44] includes not only audiovisual information but also physiological information and is often used to investigate the effectiveness of physiological signals (e.g., [48, 30]). Pérez-Rosas et al. [7] created a multimodal corpus called MOUD, which was collected from a product review of YouTube (Spanish). They demonstrated the effectiveness of the joint use of visual, acoustic, and linguistic modalities for utterance-level sentiment analysis (positive, negative, or neutral, as labeled by two annotators) based on video reviews, i.e., naturalistic conditions. In that study, an accuracy of 70.9% was achieved by using only linguistic information, but by fusing linguistic, acoustic and visual information, the accuracy improved to 74.1%. A larger-scale dataset for multimodal sentiment/emotion analysis was obtained in [9], including 23,453 annotated sentences from more than 1,000 online YouTube speakers. The large-scale multimodal multi-party emotional conversational database was also recently created by Poria et al. [45], which contains approximately 13,000 utterances from 1,433 dialogues based on TV-series. In addition, Yao et al. [49] reported multimodal sentiment analysis in real-life settings (consumer interviews) with noisy transcriptions and imbalanced label distributions, i.e., more challenging settings.

All mentioned datasets have helped to drive advances in multimodal sentiment research. However, multimodal sentiment analysis is still in its infancy. As mentioned in Chapter 1, most multimodal analyses are based on

TS labels and ignore SS labels. In particular, multimodal analysis with physiological signals that are closely related to the emotional response has been scarcely investigated. To solve this problem, the recently created dataset Hazumi1911 ([46], the bottom row in Table 2.1), which includes SS labels, TS labels and physiological signals per exchange during dialogue, is used in this thesis (Chapter 4 to 6, more details of the dataset are described in Chapter 4). Furthermore, physiological individual differences have not been considered in previous studies. Thus, this thesis starts by presenting this fundamental problem and its solution (please see Figure 1.2 for the position of the research).

Chapter 3

Biosignal-based Emotion Recognition with Importance Weighting

3.1 Introduction

The advantages of biosignals in emotional inner state estimations are evidenced in theoretical research in physiology. The ANS consists of the sympathetic nervous system and the parasympathetic nervous system, which maintain homeostasis of organisms [50]. Emotional stimuli activate the sympathetic nervous system, and these responses result in an increased HR and EDA. On the other hand, when relaxing, the parasympathetic nervous system becomes dominant and these parameters return to a steady state. Since a correlation between brain activity and Heart Rate Variability (HRV) in emotion-evoking stimuli has also been reported [15], it is likely that valuable information is included in physiological signals.

Emotions are also related to personality [51, 52, 53]. Extraversion is associated with low cortical arousal to external stimuli and a desire for more stimuli that evoke emotions, while neuroticism is associated with confusion and nervousness, even in low-stress situations [51]. It has also been reported that agreeableness is a predictor of efforts to control emotion [52], and conscientiousness predicts lower reactivity to negative emotions [53]. These results suggest that emotion is modulated by personality, so it is necessary to simultaneously understand emotion and personality. As emotions are related to the physiological systems described here, Zuckerman [54] assumed that biological factors such as monoamine neurotransmitters and genetic factors are related to personality. Since personality is also associated with the ANS and

HRV [55], biosignals may also contain valuable information for personality estimation.

Recently, A dataset for Multimodal research of affect, personality traits and mood on Individuals and GrOupS (AMIGOS) dataset [56], which consists of multimodal recordings of biosignals, such as the ECG and Galvanic Skin Response (GSR), was constructed for research on emotion and personality. The performance results of emotion and personality estimations were shown in [56], but most of the results achieved approximately 50-55% mean F1-scores in binary classification tasks. Although biosignals are expected to be useful for estimating emotion and personality, performance improvement is still needed.

One of the concerns in applying nonverbal data, such as biosignal data, to the development of a machine learning model is the existence of individual differences, as shown in [57, 58]. Generally, when using data from individuals, the test data from one user should be completely excluded in the training dataset. For example, Leave-One-User-Out Cross-Validation (LOUOCV) should be employed to evaluate the machine learning model. Obviously, this user-independent evaluation schema is also important in emotion recognition tasks. On the other hand, estimation performance of the models would decrease if the emotional physiological responses between users are different. This problem, in which the training (source) data are biased and potentially nonrepresentative, is known as a covariate shift. [16]. It is necessary to consider covariate shifts in the LOUOCV schema whenever the dataset includes data, such as biosignals, that have individual differences. For this reason, we adapt a machine learning model using the covariate shift adaptation (also referred to as IW) technique to estimate an individual's emotion and personality in this study. The main contribution of our work is summarized as follows:

Emotion and personality estimations with the importance-weighted model: We adapt a machine learning model using the IW technique to estimate an individual's emotion and personality. We constructed Importance-Weighted Logistic Regression (IW-LR) and an Importance-Weighted Support Vector Machine (IW-SVM) and performed emotion and personality estimations with the AMIGOS dataset. The binary classification results show that the IW method significantly improve the emotion and personality estimation performance results in the SVM model (Section 3.4.2).

Comparison of the estimation performance to the theoretical upper bound: Additionally, to examine the validity of our experimental results, we performed Bayesian Error Rate (BER) analysis with the AMIGOS dataset.

The BER represents a lower bound on the error rate and is used to benchmark the classification algorithms [59, 60]. Comparing the accuracy based on the estimated BER with our constructed model performance, we analyzed the upper bound of the estimation performance results with the AMIGOS dataset (Section 3.5.1).

Effective features for emotion and personality estimation: To evaluate the contribution of each biological feature to binary classification tasks in a scalable manner, feature analysis was conducted by an ablation test. We revealed that ECG features were effective features for valence estimation and that almost all GSR features were effective for arousal estimation. GSR features also contribute to personality estimation, especially the 2nd difference of Skin Conductance (SC), which was the most effective feature. These results are discussed in Section 3.5.2.

3.2 Related Works

3.2.1 Databases for Emotion and Personality Research

Emotion recognition during video-watching tasks or human-machine interface interaction has been mainly based on audiovisual data. For example, the SEMAINE database [43] includes the visual, audio, and emotional scores of 150 subjects during interactions with an artificial agent. This database was constructed for a dialogue system to enhance user experience and is utilized in some research on emotion recognition [61, 62]. In addition, McDuff et al. [63] constructed a database to analyze interest in commercial videos; this database contains the facial expression data of 242 users, who were recorded with webcams while they watched videos (Affectiva-MIT Facial Expression Dataset (AM-FED) dataset). In recent years, there have been several reports on the construction of multimodal databases that include not only audiovisual data but also biological signals. The MAHNOB-human computer interface database contains the visual, audio, gaze, ECG, GSR, respiration, Skin temperature (TEMP), EEG, and emotional scores of 27 individuals [64]. Each data point was acquired while the subjects viewed 20 kinds of videos that evoke emotions such as joy and fear. The Database for Emotion Analysis using Physiological Signals (DEAP) used a one-minute music video to evoke emotions [65]. This database contains the EEG, GSR, respiration amplitude, TEMP, ECG, blood volume by plethysmograph, EMG, Electrooculogram (EOG) and emotional scores of 32 subjects. DECAF contains magnetoencephalography (MEG) data, which have higher spatial resolution than EEG

data and minimal physical contact; the data include the emotional responses of 30 subjects to 40 music videos and 36 movie clips annotated in advance by experts with the kind of emotion evoked [66]. In addition to MEG data and emotional responses, EOG, ECG, EMG, and facial expression data based on near-infrared light are included. The databases described here do not contain information about subjects' personalities.

As previously mentioned, many databases for emotion recognition based on biological signals have been reported in recent years. However, because personality significantly influences emotions and is an important factor related to emotional responses to stimuli, it is necessary to investigate the relation between emotion and personality for model implementation and application in the real world. There are several studies on personality trait recognition [67, 68, 69], which are often based on linguistic, acoustic, and visual information. However, to the best of our knowledge, there are only two databases that include biosignals for multimodal research on emotion and personality: ASCERTAIN [70] and AMIGOS [56]; the related studies proposed emotion and personality estimation models in a similar framework. These databases include EEG, ECG, GSR, visual information, emotional scores (valence and arousal) and personality data (Big Five personality traits, [71]).

3.2.2 Application of Importance Weighting

To address the physiological individual differences in emotion and personality estimations, we consider an individual difference as a covariate shift, where the distribution of the training input density and test input density, changes from $p_{\mathrm{tr}}(\boldsymbol{x})$ to $p_{\mathrm{te}}(\boldsymbol{x})$, but the conditional distribution of the outputs given the inputs $p(y|\boldsymbol{x})$ remains unchanged [16]. Covariate shift adaptation, which is also referred to as IW, can be considered domain shift adaptation and categorized into a sample-based approach in domain adaptation methods [72]. In speaker recognition, which is a technology for recognizing individuals from speech, the distribution of audio data (input density) changes depending on the measurement conditions, and these situations can cause performance degradation. To solve this problem, the IW method has been shown to be useful for improving recognition performance [73]. Furthermore, in BCI research area, it has been reported that the IW method improves the performance results of models based on EEG in linear discriminant analysis [74]. To the best of our knowledge, there is no report that evaluates the IW method in the context of biosignal-based emotion and personality estimations. It is possible that the IW method can also improve the performance results of emotion and personality recognition models based on other physio-

Table 3.1: F1-score for the emotion recognition models reported in ASCERTAIN [70] and AMIGOS [56].

Dataset	Classifier	Feature	Valence	Arousal	Mean
ASCERTAIN [70]	NB	ECG	0.600	0.590	0.595
	NB	GSR	0.680	0.660	0.670
	L-SVM	ECG	0.560	0.570	0.565
	L-SVM	GSR	0.640	0.610	0.625
AMIGOS [56]	NB	ECG	0.535	0.550	0.543
	NB	GSR	0.531	0.548	0.540

Table 3.2: F1-score for the personality recognition models reported in ASCERTAIN [70] and AMIGOS [56].

Dataset	Classifier	Feature	Ex	Ag	Co	Ne	Op	Mean
ASCERTAIN [70]	NB	ECG	0.560	0.550	0.600	0.530	0.480	0.544
	NB	GSR	0.450	0.390	0.570	0.490	0.280	0.436
	L-SVM	ECG	0.060	0.450	0.510	0.600	0.350	0.394
	L-SVM	GSR	0.000	0.340	0.350	0.560	0.360	0.322
AMIGOS [56]	L-SVM	ECG	0.621	0.513	0.590	0.140	0.483	0.469
	L-SVM	GSR	0.268	0.510	0.655	0.362	0.699	0.499

logical signals, such as ECG or GSR. With respect to physiological individual differences, we focus on the IW method and verify whether the IW method is effective for improving the performance results of emotion and personality estimations with biosignals.

3.2.3 Performance Comparison of Previous Methods

NB and Linear Support Vector Machine (L-SVM) are employed for binary classification of emotion (valence and arousal) and personality (Big Five) in [70] and [56]. Tables 3.1 and 3.2 show the mean F1-scores for each model based on the different modalities (ECG and GSR) reported in [70] and [56]. As shown, the macroaverages of the F1-scores (average F1-scores over all dimensions) in the binary classification task reported in [70] are 56.5–67.0% and 32.2–54.4% for emotion estimation and personality estimation, respectively. Similarly, in the short video experiment, the macroaveraged F1-scores reported in [56] are approximately 54% and 46.9–49.9% for emotion estimation and personality estimation, respectively. Thus, it is necessary to improve the performance results of the models based on these peripheral biosignals.

ASCERTAIN [70] and AMIGOS [56] are available datasets; with these datasets, some models have been recently proposed. Zhao et al. [75] employed a hypergraph structure to formulate personality correlations among different subjects and physiological correlations among corresponding stimuli and proposed a novel method called vertex-weighted multimodal multitask hypergraph learning (VM2HL). The emotion recognition model constructed by

VM2HL shows the best accuracy of 74.34% for valence and 79.46% for arousal in the binary classification task with the ASCERTAIN dataset. Harper and Southern [76] proposed a framework based on the Bayesian Deep Neural Network (DNN) and achieved an 88% F1-score in the binary classification task for valence with the ECG data of the AMIGOS dataset. Gjoreski et al. [77] constructed a model that is referred to as a meta-learner based on classification probabilities, which are outputs from seven machine learning algorithms (e.g., decision tree, NB, and SVM) that are input into a random forest and then used for arousal estimation in a binary classification task with six datasets, including ASCERTAIN and AMIGOS [77]. Comparing the meta-learner with the conventional SVM model, the binary classification accuracy improved from 60% to 63%. Miranda-Correa and Patras [32] combined Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) and presented a multitask cascaded DNN model that jointly predicted a subject's emotion and personality. Applying this model to the EEG data, it was mentioned that the macroaverage of the F1-score of personality binary classification was improved 2.7% compared to the best baseline results in [70].

In addition to the previously mentioned models, several models are based on the ASCERTAIN dataset [31, 78] and AMIGOS dataset [79, 80, 81, 82, 83, 84, 85]. The highest performing model achieved an F1-score of 88% in [76]; however, this model is limited to only the valence level and has not been tested on the arousal level or the five personality factors. The models presented in [75] and [77] are also limited to emotion estimation and are not models for personality estimation. Alternatively, many studies have focused on developing automatic personality recognition in the personality computing region (review in [86]). Automatic personality recognition is the task of inferring self-assessed personalities. Although many studies on automatic personality recognition apply text-based or nonverbal behavior-based approaches, but the effectiveness of biosignal-based approaches is not fully understood. In this study, we propose a model for estimating not only valence and arousal but also the five factors of personality with the AMIGOS dataset. In [32], the proposed DNN model included personality estimation, but the macroaverage of the F1-score in binary classification was 57.5%. Thus, further investigation and performance improvement are still needed. Assuming the individual differences in physiological responses as covariate shifts and factors of performance degradation, we utilize the IW technique for covariate shift adaptation and show the performance improvement of the adapted models compared to the conventional procedure.

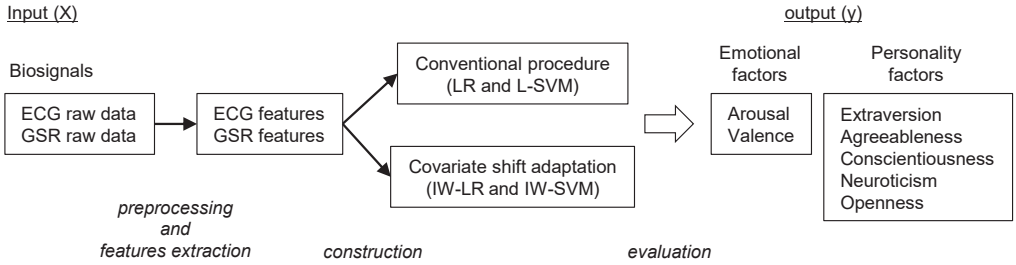

Figure 3.1: Overview of the emotion and personality estimations in this study.

3.3 Methods

Figure 3.1 shows the framework of our research. To investigate the effectiveness of the IW technique in emotion and personality estimation, ECG and GSR raw data from the AMIGOS dataset were preprocessed and feature extraction was conducted. Using these features, conventional models (Logistic Regression (LR) and L-SVM) and IW (IW-LR and IW-SVM) models were constructed for comparison between conventional methods and IW methods for the estimation performance. The performance evaluation was based on LOUOCV and the F1-score of each classification. This section describes the AMIGOS dataset, preprocessing and features extraction methods, and each machine learning model.

3.3.1 Dataset

AMIGOS [56] is a dataset that is employed in research on multimodal emotion and personality estimations. The dataset is publicly available for academic research purposes. To collect the data in AMIGOS, subjects watched videos with emotional stimuli, and EEG, ECG and GSR were recorded by sensor devices during the videos. The subjects self-reported emotional valence and arousal based on Russell's circumplex model [87] and used self-assessment manikins (range 1 to 9) at the end of each trial. Personality traits, that is, extraversion, agreeableness, conscientiousness, neuroticism and openness, proposed by Costa and McCrae [71] were obtained via an online form after the experiment was completed. The short videos for emotion stimuli utilized in the AMIGOS dataset consisted of clips from comedies and horrors, such as Mr. Bean and The Exorcist [66, 64]. The ECG and GSR raw data from the short video experiment in the AMIGOS dataset were utilized

Table 3.3: Nine ECG features extracted in this study.

Five time-domain features	
HR	heart beat per minute
RRI	interval between one R-wave and the next (ms)
RMSSD	root mean square of successive RRI differences
NN50	the number of successive RRI differences >50 ms
pNN50	percentage value of NN50
Four frequency-domain features	
vLoF	very low frequency (0.003 - 0.04 Hz)
LoF	low frequency (0.04 - 0.15 Hz)
HiF	high frequency (0.15 and 0.4 Hz)
LoF/HiF	ratio of LoF to HiF

as input, and emotional and personality scores were employed as output in our study.

3.3.2 Preprocessing and Feature Extraction

ECG Features

An ECG is a recording of the electrical activity of the heart. The heart is composed of the left atrium, left ventricle, right atrium, and right ventricle. The ECG waveform, which shows excitation of the left and right ventricular muscle, is referred to as the QRS wave, and the upward deflection is referred to as the R-wave. HRV analysis based on the R-wave Interval (RRI) is an indicator of autonomic nervous activity and has been suggested to be related to emotions and personality [70, 56, 27]. In the AMIGOS dataset, ECG data were recorded with a Shimmer ECG Unit (Shimmer, Dublin, Ireland)[1] with a sampling frequency of 256 Hz. The analysis referred to the Shimmer platform manual, and feature extraction was performed over the final 50 s of the stimulus presentation in this study, as in a previous report [70]. The PeakUtils library (PeakUtils 1.3.2)[2] was selected for R-wave detection. In the setting of PeakUtils, the amplitude threshold was 0.3, and the interval threshold (minimum distance between each R-wave) was 154, which corresponds to 100 bpm. Since the varying intervals between the R-peaks cause sample-time nonuniformity, the Lomb-Scargle method was utilized to estimate the power spectral density (PSD). The area under the PSD curve was calculated using Simpson's composite rule. Figure 3.2 shows an example of the analysis of the ECG data in the AMIGOS dataset. The five time-domain features and four frequency-domain features, for a total of nine features, as shown in Table 3.3, were extracted with reference to [70, 56, 27].

[1] http://www.shimmersensing.com/
[2] https://pypi.org/project/PeakUtils/

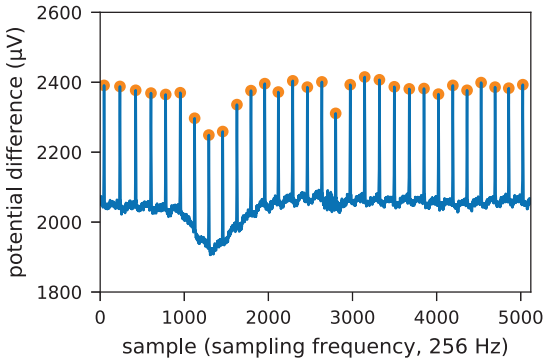

Figure 3.2: An example of the analysis of the ECG data. The horizontal axis shows the number of samples (20 s $\times$ 256 Hz sampling frequency), and the vertical axis shows the potential difference (μV). The blue line corresponds to the heartbeat waveform, and the orange circle corresponds to the peak of the R wave.

GSR Features

GSR reflects sweat gland activity via the sympathetic nervous system, and in the field of psychophysiology, it is widely utilized to detect emotional changes and as an index for arousal level [13]. Because emotional sweating is likely to occur in the fingers, a phenomenon that is associated with emotional (arousal) changes, the SC data in the AMIGOS dataset were measured with electrodes (Shimmer GSR sensors, Shimmer, Dublin, Ireland) attached to fingers. Each measured value was stored as a 16-bit integer in the AMIGOS dataset, and thus, the SC level was calculated from the reciprocal after converting each 16-bit integer into a skin electrical resistance value by referencing the Shimmer manual. Noise was removed by a low-pass filter (sampling frequency of 128 Hz, pass-band edge frequency of 1 Hz, and stop-band edge frequency of 2 Hz). SC changes in time series are decomposed into SC levels (tonic component) and Galvanic Skin Response (GSR)s; the SC level was calculated with polynomial fitting (degree = 10); and the GSR was detected with PeakUtils (amplitude threshold of 0.3, and interval threshold of 3 s). Figure 3.3 shows an example of the analysis of the SC data in the AMIGOS dataset. After preprocessing, five GSR features in total (mean, standard deviation, 1st and 2nd difference of SC, and number of GSRs) were extracted.

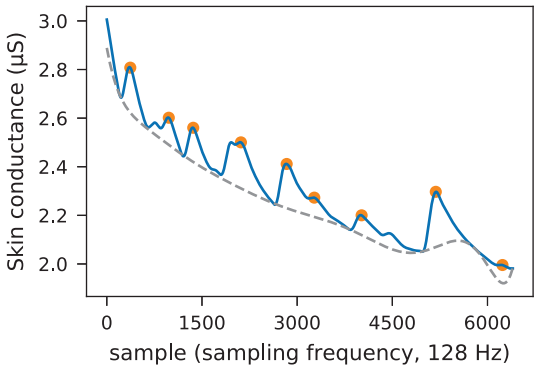

Figure 3.3: An example of the analysis of the GSR data. The horizontal axis represents the number of samples (50 s × 128 Hz sampling frequency), and the vertical axis represents the SC (μSiemens). The gray dashed line corresponds to the SC level, the blue line corresponds to the time series of SC and the orange circle corresponds to the GSR.

3.3.3 Machine Learning Model

Physiological measurements often encounter the problem of individual differences (such as behavior and thought) [57, 58] and nonstationality (such as changes in the experimental condition) [74, 88]. The IW method is shown to be effective in addressing these situations [73, 74]. Thus, to address this problem in biosignal-based emotion and personality estimations, we utilize the IW method in machine learning modeling. Each importance-weighted model is described in this subsection.

Importance-weighted Logistic Regression (IW-LR)

LR is a generalized linear model and popular learning algorithm for binary classification. Assuming that the data consist of an input $\boldsymbol{x}_i$ and output $y_i \in \{+1, -1\}$ and $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)$ drawn from independent and identically distributed (i.i.d.) variables, the logistic loss function is

$$L(\boldsymbol{w}; \boldsymbol{x}, y) := \log(1 + \exp(-yf(\boldsymbol{x}; \boldsymbol{w}))) \tag{3.1}$$

where $\boldsymbol{w}$ is the parameter vector. Standard learning methods do not produce the optimal solution under a covariate shift even when the number of training samples tends to infinity. The influence of a covariate shift could be alleviated by weighting the loss function according to importance (described in Section 3.3.3). An importance-weighted version of LR is referred to as IW-LR [89,

90]), and the optimization problem can be expressed as follows:

$$\underset{w}{\operatorname{argmin}} \left[\frac{1}{2}\|\boldsymbol{w}\|^2 + C \sum_{i=1}^{n_{\mathrm{tr}}} (r_i)^{\gamma} \log(1 + \exp(-y_i f(\boldsymbol{x}_i^{\mathrm{tr}}; \boldsymbol{w}))) \right] \tag{3.2}$$

where C is the penalty parameter that adjusts the trade-off between the loss function and the penalty term $\|\boldsymbol{w}\|^2$. $r(\boldsymbol{x})$ is the importance and γ is the flattening parameter. n_{tr} is the number of the training samples.

Importance-weighted Support Vector Machine (IW-SVM)

SVM is also a popular and conventional learning algorithm for binary classification [91]. Assuming that the data consist of an input $\boldsymbol{x}_i$ and output $y_i \in \{+1, -1\}$, the convex quadratic optimization problem of the soft margin SVM model is given by

$$\underset{w}{\operatorname{argmin}} \left[\frac{1}{2}\|\boldsymbol{w}\|^2 + C \sum_{i=1}^{n_{\mathrm{tr}}} \xi_i \right] \tag{3.3}$$

$$\text{subject to } y_i(\boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_i^{\mathrm{tr}} + b) \geq 1 - \xi_i,\ \xi_i \geq 0,\ 1 \leq i \leq n_{\mathrm{tr}}$$

where $\boldsymbol{w}$ is the parameter vector of the discriminant function, b is the bias parameter, ξ is the slack variable, C is the penalty parameter that adjusts the trade-off between the loss function and the maximization of the margin, and n_{tr} is the number of training samples. An importance-weighted version of L-SVM is referred to as the IW-SVM [90]), and the optimization problem can be expressed as follows:

$$\underset{w}{\operatorname{argmin}} \left[\frac{1}{2}\|\boldsymbol{w}\|^2 + C \sum_{i=1}^{n_{\mathrm{tr}}} (r_i)^{\gamma} \xi_i \right] \tag{3.4}$$

$$\text{subject to } y_i(\boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_i^{\mathrm{tr}} + b) \geq 1 - \xi_i,\ \xi_i \geq 0,\ 1 \leq i \leq n_{\mathrm{tr}},\ \gamma \in [0, 1]$$

where $r(\boldsymbol{x})$ is the importance and γ is the flattening parameter. When training the importance-weighted model, the penalty parameter C is fixed, but the penalty for the samples with greater importance is increased and the penalty for the samples with low importance is decreased by the IW.

Importance Estimation

Importance is defined as the ratio of the test density to the training density and is given by

$$r(\boldsymbol{x}) = \frac{p_{\mathrm{te}}(\boldsymbol{x})}{p_{\mathrm{tr}}(\boldsymbol{x})} \tag{3.5}$$

where $p_{\mathrm{tr}}(\boldsymbol{x})$ is the training input density and $p_{\mathrm{te}}(\boldsymbol{x})$ is the test input density. There are some methods for calculating importance $r(\boldsymbol{x})$, such as Kernel Mean Matching (KMM) [92], the Kullback-Leibler Importance Estimation Procedure (KLIEP) [93] and unconstrained Least-Squares Importance Fitting (uLSIF) [94]. Because the uLSIF method can be computed faster than other methods [94] and uLSIF with a Gaussian kernel (Kernelized variant of uLSIF (KuLSIF) [95]) compares favorably with other approaches, we employed KuLSIF in this study. The squared error of the estimator $r(\boldsymbol{x})$ is expressed as follows:

$$\frac{1}{2}\int \left(\hat{r}(\boldsymbol{x}) - \frac{p_{\mathrm{te}}(\boldsymbol{x})}{p_{\mathrm{tr}}(\boldsymbol{x})}\right)^2 p_{\mathrm{tr}}(\boldsymbol{x})d\boldsymbol{x} \tag{3.6}$$

In this expression, the constant can be safely disregarded, and the expectations are approximated by sample averages. Subsequently, $\hat{r}(\boldsymbol{x})$ is obtained as an optimal solution of

$$\mathrm{loss}(r) = \frac{1}{2n}\sum_{i=1}^{n} \hat{r}(\boldsymbol{x}_i^{\mathrm{tr}})^2 - \frac{1}{m}\sum_{j=1}^{m} \hat{r}(\boldsymbol{x}_j^{\mathrm{te}}) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 \tag{3.7}$$

where $\boldsymbol{\theta}$ is the parameter vector and λ is the regularization parameter. It is assumed that the model for $\hat{r}(\boldsymbol{x})$ is

$$\hat{r}(\boldsymbol{x}) = \sum_d \theta_d \phi_d(\boldsymbol{x}) \quad (\theta_1, \ldots, \theta_D \in \mathbb{R}) \tag{3.8}$$

$$k(\boldsymbol{x}, \boldsymbol{x}') = \phi(\boldsymbol{x})^{\mathsf{T}}\phi(\boldsymbol{x}'), \quad \phi(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}), \ldots, \phi_D(\boldsymbol{x}))^{\mathsf{T}} \tag{3.9}$$

where k is a kernel function. Applying the representer theorem, $\boldsymbol{\theta}$ is given by the following expression:

$$\boldsymbol{\theta} = \sum_{i=1}^{n} \alpha_i k(\boldsymbol{x}, \boldsymbol{x}_i^{\mathrm{tr}}) + \sum_{j=1}^{m} \beta_j k(\boldsymbol{x}, \boldsymbol{x}_j^{\mathrm{te}}) \tag{3.10}$$

Therefore, the optimization problem is reduced to

$$\hat{r}(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i k(\boldsymbol{x}, \boldsymbol{x}_i^{\mathrm{tr}}) + \frac{1}{m\lambda} \sum_{j=1}^{m} k(\boldsymbol{x}, \boldsymbol{x}_j^{\mathrm{te}}) \tag{3.11}$$

Substituting Eqs. (3.10) and (3.11) into the loss function, Eq. (3.7), and given the extremal condition of Eq. (3.7) with respect to parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^{\mathsf{T}}$, the following linear equation is solved for $\boldsymbol{\alpha}$:

$$\left(\frac{1}{n} K_{11} + \lambda I_n\right) \boldsymbol{\alpha} = -\frac{1}{nm\lambda} K_{12} \mathbf{1}_m \tag{3.12}$$

$$(K_{11})_{ii'} = k(\boldsymbol{x}_i^{\mathrm{tr}}, \boldsymbol{x}_{i'}^{\mathrm{tr}}), \quad (K_{12})_{ij} = k(\boldsymbol{x}_i^{\mathrm{tr}}, \boldsymbol{x}_j^{\mathrm{te}}) \tag{3.13}$$

3.4 Experiments

Data from thirty-seven subjects were utilized for the experiment that corresponds to [56]. The feature set was standardized to have an average of 0 and a variance of 1 for each subject. For the binary classification task, the emotion and personality scores were divided into low and high classes via the median value and converted to -1 and $+1$, respectively.

3.4.1 Experimental Settings for the Classification Task

In the case of our experiment, most of the $r(\boldsymbol{x})$ are arranged in 10^{-1} to 10^{-2} order. Thus, we set the penalty parameter $C = 100$ to adjust the order. The flattening parameter γ is the hyperparameter ($0 \leqq \gamma \leqq 1$), which flattens the importance weights [90]. $\gamma = 1$ in this study. We set the total number of epochs to 1000, and the learning rate to 0.01. These hyperparameters for the conventional and IW methods are identical, except for the flattening parameter γ, which is specific for IW.

Binary classification of valence and arousal was performed with the conventional (LR and L-SVM) and importance-weighted (IW-LR and IW-SVM) models (Section 3.3.3). All settings were identical for the conventional and importance-weighted models, except for the presence of IWs. Subjects used a self-assessment to report their emotional valence and arousal scores after watching each of 16 short videos, that is, 16 samples can be applied per subject for model construction. In the fusion model, the ECG and GSR feature vectors were concatenated and utilized for binary classification. The LOUOCV method was performed in the ECG, GSR and fusion models. In this method, the test data from one participant are excluded in the training dataset. This method was repeatedly applied to each user (37 in total), and

the macro F1-score was calculated in each estimation. Each macro F1-score was averaged and applied as the result value. This method is identical to that in [56]. In this way, the F1-scores of the conventional and importance-weighted models were compared. In personality estimation, features obtained from 16 videos were concatenated and used for model construction. The binary classification task for the Big Five traits, namely, extraversion, agreeableness, conscientiousness, neuroticism and openness, was performed in the conventional and importance-weighted models, and similar to the emotion estimation, the F1-scores were obtained by LOUOCV and compared. The difference between the F1-scores of the conventional models and those of the importance-weighted models ($\Delta\%$) was utilized as a performance improvement index for the IW method.

3.4.2 Results

Emotion Estimation Performance

The results of the binary classification task in emotion estimation are summarized in Table 3.4. In Table 3.4, the difference between the F1-score of the conventional and importance-weighted models ($\Delta\%$) was applied as a performance improvement index for the IW method. For the valence estimation task, models trained with ECG features have the best performance results, but performance improvement of the IW methods was not observed (columns 3 and 4). On the other hand, the best models among the conventional SVM models were the fusion models, which reached F1-scores of 60.0% (column 6),. This performance was improved by the IW methods in the arousal estimation task (0.8$\Delta\%$, column 8). For the F1-score macroaverage, all importance-weighted models outperformed the conventional models (shown in column 11), except the SVM model based on GSR features. For LR trained with ECG, GSR, and fusion features, the mean $\Delta\%$ values are 0.4%, 1.7%, and 1.1%, respectively; for SVM, the mean $\Delta\%$ values are 0.1%, 0% and 0.7%, respectively. The maximum performance results of IW-SVM on the emotion classification task reached a 57.6% F1-score macroaverage with the fusion model.

Personality Estimation Performance

The results of the binary classification task in personality estimation are summarized in Table 3.5. For personality estimation, IW improves all the performance results of the SVM models in terms of the macroaveraged F1-scores (average of the F1-scores of the estimated five personality factors, as

Table 3.4: Performance comparison between the conventional and importance-weighted models on the emotion estimation task. Mean F1-scores for emotion estimation using the ECG and GSR models are shown. (W/O, without IW (i.e., the conventional model); IW, the importance-weighted model; fusion, the ECG+GSR features)

Classifier	Feature	Valence			Arousal			Mean		
		W/O	IW	Δ%	W/O	IW	Δ%	W/O	IW	Δ%
LR	ECG	0.564	0.558	-0.6	0.507	0.520	1.4	0.535	0.539	0.4
LR	GSR	0.353	0.396	4.3	0.592	0.583	-0.9	0.472	0.489	1.7
LR	fusion	0.523	0.551	2.8	0.601	0.595	-0.5	0.562	0.573	1.1
SVM	ECG	0.564	0.564	-0.1	0.501	0.502	0.2	0.532	0.533	0.1
SVM	GSR	0.343	0.349	0.6	0.595	0.589	-0.6	0.469	0.469	0.0
SVM	fusion	0.538	0.544	0.6	0.600	0.608	0.8	0.569	**0.576**	0.7
Mean		0.481	0.494	1.3	0.566	0.566	0.1	0.523	0.530	0.7

shown in columns 19 and 20 in Table 3.5). However, there were no performance improvements in the LR models. The best performance result is observed in the IW-SVM model trained with GSR features, which resulted in a 59.4% macroaveraged F1-score. To evaluate the effectiveness of the IW method in total, rather than in each classifier and features, a comparison between the conventional method and the IW method was performed using "Mean" values from Table 3.4 and 3.5 (12 values in total: emotion, 6 mean values; personality, 6 mean values). Comparing this 12 mean values, there is significant difference between conventional and IW method models (paired t-test, $p = 0.032$). Moreover, among the 2 emotion factors and 5 personality factors, the average value at every column of the importance-weighted model is higher than that of the conventional model, except for neuroticism (Table 3.4 row 9 and Table 3.5 row 9). These results indicated that the IW method effectively worked in the binary classification task.

Table 3.5: Performance comparison between conventional and importance-weighted model in the personality estimation. (W/O, without IW; IW, the importance-weighted model; fusion, the ECG+GSR features; Ex, extraversion; Ag, agreeableness; Co, conscientiousness; Ne, neuroticism; Op, openness)

Classifier	Feature	Ex			Ag			Co			Ne			Op			Mean		
		W/O	IW	Δ%	W/O	IW	Δ%	W/O	IW	Δ%	W/O	IW	Δ%	W/O	IW	Δ%	W/O	IW	Δ%
LR	ECG	0.513	0.513	0.0	0.593	0.593	0.0	0.432	0.432	0.0	0.593	0.593	0.0	0.595	0.595	0.0	0.545	0.545	0.0
LR	GSR	0.483	0.473	-1.0	0.649	0.675	2.7	0.510	0.535	2.5	0.590	0.528	-6.2	0.593	0.593	0.0	0.565	0.561	-0.4
LR	fusion	0.432	0.459	2.7	0.621	0.621	0.0	0.456	0.426	-3.0	0.593	0.593	0.0	0.510	0.510	0.0	0.523	0.522	-0.1
SVM	ECG	0.535	0.504	-3.1	0.565	0.608	4.3	0.510	0.560	4.9	0.584	0.590	0.6	0.539	0.535	-0.4	0.547	0.559	1.3
SVM	GSR	0.459	0.539	8.0	0.728	0.702	-2.6	0.560	0.539	-2.0	0.584	0.598	1.4	0.535	0.590	5.5	0.573	**0.594**	2.1
SVM	fusion	0.262	0.528	26.6	0.619	0.590	-2.9	0.473	0.526	5.3	0.504	0.473	-3.2	0.449	0.405	-4.5	0.461	0.504	4.3
Mean		0.447	0.503	5.5	0.629	0.632	0.2	0.490	0.503	1.3	0.575	0.563	-1.2	0.537	0.538	0.1	0.536	0.548	1.2
SVM [56]	ECG	0.621	-	-	0.513	-	-	0.590	-	-	0.140	-	-	0.483	-	-	0.469	-	-
SVM [56]	GSR	0.268	-	-	0.510	-	-	0.655	-	-	0.362	-	-	0.699	-	-	0.499	-	-
DNN [32]	EEG	0.590	-	-	0.754	-	-	0.539	-	-	0.621	-	-	0.371	-	-	0.575	-	-

3.5 Discussion

Traditionally, feature engineering and model selection, such as L-SVM or NB model, are the preferred methods to improve the estimation performance in machine learning. Although these techniques are powerful tools in affective computing and personality estimation even today, performance of a recognition model can be degraded if there are individual differences. Especially, in applying biosignal data, there is the need to take into account the individual differences [57, 58]. Our main focus is the physiological individual differences in the training data, which are considered a covariate shift in machine learning theory. We investigated whether the IW technique compensates for this shift. Our work focuses on the probabilistic distribution of the samples derived from biosignals, and thus, is different from traditional methods, such as feature engineering and the model selection approach for performance improvement. The experimental results showed that the IW method significantly improved estimation performance. These results indicated that the physiological individual differences caused degradation of the estimation performance, which was effectively mitigated by the IW methods in emotion and personality estimations. In this section, we further analyze the estimated upper bound of the classification accuracy for the AMIGOS dataset. In addition, feature analysis was performed to investigate the effective biological features. We discussed the differences between our results and the results of previously reported models.

3.5.1 Analysis of the Classification Performance by Estimating BER

To examine the validity of the binary classification accuracy, we performed BER estimation with the AMIGOS dataset. The BER estimation task was performed with the SmartSVM package[3], which includes the Henze-Penrose estimator of the BER based on the construction of the Euclidean minimal spanning tree [60]. The BER is not the true value but the estimated value obtained with this model. For a comparison of the experimental results of the binary classification accuracy in this study, we considered $1 - BER$ as the upper bound on the accuracy. Figure 3.4a and d shows the comparison of the (IW-)LR model accuracy with the (IW-)SVM model accuracy based on ECG features and $1 - BER$ in emotion and personality estimations. Similarly, figure 3.4b and e shows the case of the models based on GSR features, and figure 3.4c and f shows the case of the fusion models. Most of the results

[3]https://pypi.org/project/smartsvm/

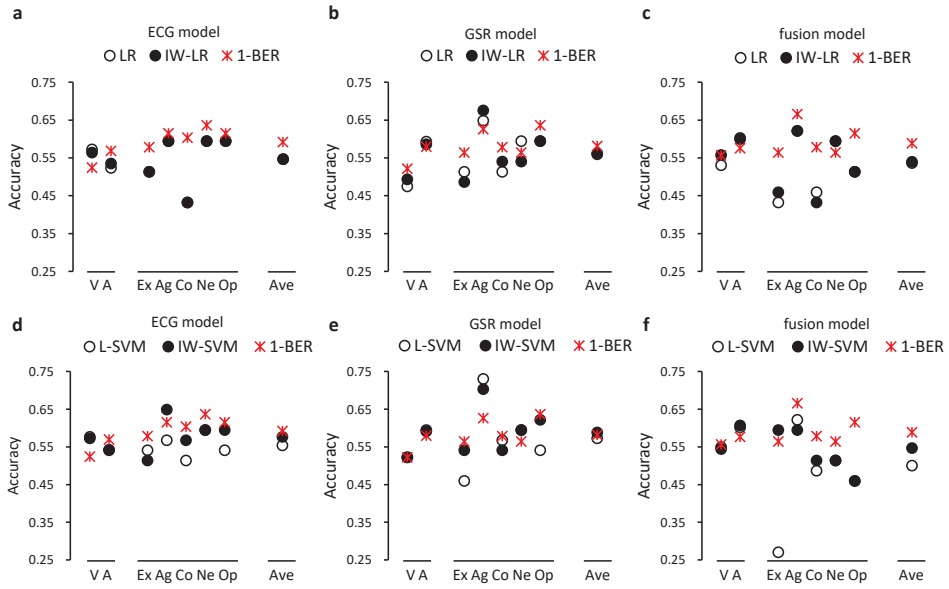

Figure 3.4: Comparison of the BER with the accuracy of the binary classification model. (**a** and **d**) The classification accuracy of the LR (open circle), IW-LR (closed circle), SVM (open circle) and IW-SVM (closed circle) models trained with ECG features is compared to $1 - BER$ (red asterisk). (**b** and **e**) The classification accuracy of the LR, IW-LR, SVM and IW-SVM models trained with GSR features are compared to $1 - BER$, and similarly, (**c** and **f**) shows the fusion models (V, valence; A, arousal, Ex, extraversion; Ag, agreeableness; Co, conscientiousness; Ne, neuroticism; Op, openness; Ave, average accuracy of these seven factors).

based on the conventional method and IW method were less than $1 - BER$, but the models derived from IW methods tended to produce results nearly equal to this estimated upper bound. This result indicated that the experimental results of our constructed model were reasonable, and it is difficult to outperform this estimated benchmark unless improving data collection methods or feature extraction methods.

3.5.2 Feature Analysis

The IW-SVM method based on fusion (ECG and GSR) features was the best model in emotion estimation (Table 3.4), and the IW-SVM model trained with GSR features was the best model in personality estimation (Table 3.5) with respect to the macroaveraged F1-score. Simultaneously, the performance results of the conventional versions of these models have the greatest

Table 3.6: Contribution of each feature for emotion estimation (F1, F1-score; diff, the difference between the original fusion model's F1-score and the F1-score of the feature-removed model).

Features	Emotion					
ECG+GSR features in Table 3.4	Valence		Arousal		Mean	
	0.538		0.600		0.569	
Removed features	F1	diff	F1	diff	F1	diff
HR	0.522	+0.017	0.608	-0.008	0.565	+0.004
RRI	0.537	+0.002	0.597	+0.003	0.567	+0.002
NN50	0.547	-0.008	0.612	-0.012	0.579	-0.010
pNN50	0.547	-0.009	0.617	-0.017	0.582	-0.013
RMSSD	0.533	+0.005	0.606	-0.006	0.570	-0.001
vLoF	0.511	+0.027	0.588	+0.012	0.550	+0.020
LoF	0.561	-0.023	0.590	+0.011	0.575	-0.006
HiF	0.550	-0.012	0.579	+0.022	0.564	+0.005
LoF/HiF	0.556	-0.018	0.585	+0.016	0.570	-0.001
SC mean	0.548	-0.010	0.609	-0.009	0.579	-0.010
SC sd	0.554	-0.016	0.589	+0.011	0.571	-0.002
1st diff	0.550	-0.012	0.558	+0.042	0.554	+0.015
2nd diff	0.550	-0.012	0.600	+0.000	0.575	-0.006
n of GSR	0.541	-0.002	0.594	+0.007	0.567	+0.002

macroaveraged F1-scores among the conventional methods (L-SVM fusion model performance in emotion estimation: 0.569 macroaveraged F1-score, L-SVM GSR model performance in personality estimation: 0.573 macroaveraged F1-score). Although the ECG and GSR features employed in this study are conventional features, the effectiveness of the conventional features for emotion and personality estimation are not simultaneously investigated in detail. Thus, we investigate the contribution of each feature by ablating the features one by one from these models. If the F1-score degraded after feature ablation, the ablated feature was effective for estimation. In contrast, if the F1-score improved, the ablated feature was not effective for estimation. Table 3.6 shows the emotion estimation performance and the change in F1-score caused by ablating each feature from the original fusion model. "diff" indicates the difference between the F1-score of the original fusion model and the feature-ablated model; thus, "diff" values with a positive sign (+) indicate that the ablated feature was effective for estimation. ECG features (HR, RRI, Root Mean Square of Successive R-wave interval Differences (RMSSD), and very Low Frequency (vLoF)) were effective features for valence estimation (columns 2-3 in Table 3.6). On the other hand, RRI, the frequency-domain features of ECG, and GSR features, except for the SC mean, were effective features for arousal estimation (columns 4-5 in Table 3.6). Considering the macroaveraged F1-score, vLoF was the most effective feature (+0.02) among all features.

Based on the mean "diff" values shown in Table 3.6 row 7, four factors

Table 3.7: Contribution of each feature evaluated by a backward-forward stepwise method. The fusion model without pNN50 and 2nd diff are employed for evaluation in the final step. The other remaining features are ablated one by one (diff, the difference between the final step model's F1-score and the F1-score of the feature-removed model).

Features	Emotion					
ECG+GSR features of the final step	Valence		Arousal		Mean	
	0.557		0.609		0.583	
Removed features	F1	diff	F1	diff	F1	diff
HR	0.538	+0.019	0.566	+0.043	0.552	+0.031
RRI	0.543	+0.014	0.573	+0.036	0.558	+0.025
NN50	0.549	+0.008	0.580	+0.029	0.564	+0.019
RMSSD	0.543	+0.014	0.577	+0.032	0.560	+0.023
vLoF	0.532	+0.025	0.562	+0.047	0.547	+0.036
LoF	0.553	+0.004	0.564	+0.045	0.559	+0.024
HiF	0.544	+0.014	0.549	+0.060	0.546	+0.037
LoF/HiF	0.538	+0.019	0.550	+0.059	0.544	+0.039
SC mean	0.547	+0.010	0.563	+0.046	0.555	+0.028
SC sd	0.543	+0.014	0.509	+0.100	0.526	+0.057
1st diff	0.530	+0.027	0.575	+0.034	0.552	+0.031
n of GSR	0.550	+0.007	0.565	+0.044	0.558	+0.025

(Total number of successive R-wave interval differences that differ by more than 50 ms (NN50), percentage value of NN50 (pNN50), SC mean and 2nd diff) seem to have a inhibitory effect to the classification task. To further verify the roles of these features, we additionally evaluated the contribution of features using backward-forward stepwise selection [96]. Following the ablation test shown in Table 3.6 (refers to first step), stepwise selection was performed using the mean "diff" value. The feature-removing criteria is whether the "diff" value is negative and minimum in this step. For example, among the four factors, pNN50 represents the worst case, since the mean "diff" value is -0.013, which is negative and minimum in Table 3.6 row 7. Thus, the feature set "without pNN50" was utilized for the next step. The next evaluation is performed in the same way as in the first step. The re-entering criteria of a feature, which is already removed, is whether the estimation performance improves by re-entering. Based on these criteria, the final results of the stepwise selection is depicted in Table 3.7. There are no features that satisfy the removing/re-entering criteria. Considering these results, the NN50 and SC mean could also contribute to emotion estimation, but we could not clarify the effectivity of pNN50 and 2nd difference of SC (2nd diff) in our experiment.

Similarly, in Table 3.8, the personality estimation performance and the differences between the original GSR model results and the feature-removed results are listed. All GSR features were effective in estimating agreeableness

and neuroticism (columns 4-5 and 8-9 in Table 3.8). The 2nd diff was the most effective feature in total (+0.042, row 8 and column 13). The 1st diff (+0.023, row 7 and column 13) and number of GSRs (+0.018, row 9 and column 13) were also effective features for personality estimation, but it seemed that the SC mean and sd were not very important.

Table 3.8: Contribution of each feature for personality estimation (F1, F1-score; diff, the difference between the original GSR model's F1-score and the F1-score of the feature-removed model).

Features	Personality											
GSR features in Table 3.5	Ex		Ag		Con		Ne		Op		Mean	
	0.459		0.728		0.560		0.584		0.535		0.573	
Removed features	F1	diff	F1	diff	F1	diff	F1	diff	F1	diff	F1	diff
SC mean	0.567	-0.108	0.648	+0.080	0.590	-0.030	0.547	+0.037	0.550	-0.015	0.580	-0.007
SC sd	0.535	-0.075	0.621	+0.107	0.590	-0.030	0.468	+0.116	0.648	-0.113	0.572	+0.001
1st diff	0.539	-0.080	0.593	+0.135	0.461	+0.099	0.540	+0.044	0.619	-0.084	0.550	+0.023
2nd diff	0.553	-0.094	0.581	+0.147	0.461	+0.099	0.552	+0.032	0.510	+0.025	0.531	+0.042
n of GSR	0.458	+0.001	0.608	+0.120	0.480	+0.080	0.535	+0.049	0.695	-0.160	0.555	+0.018

3.5.3 Performance Comparison with a Previous Report

The dataset used here was collected by [56]. Identical subjects and the same biosignal raw data applied in [56] were utilized as input sources in this study. With respect to the emotional scores, Miranda-Correa et al. [56] utilized scores based on external annotations, but we used emotional scores based on self-assessment so there is a difference in the evaluation value (output y). In addition, since Miranda-Correa et al. [56] employed the NB model for emotion estimation, it is difficult to simply compare it with our L-SVM model. Harper and Southern [76] achieved an 88% F1-score in the binary classification task for valence based on a self-assessment. Their model achieved the best performance results to the best of our knowledge, and our model could not outperform it. However, the proposed model in [76] only included the valence factor, but our proposed model comprehensively included two emotional factors and five personality factors based on the self-assessment. Therefore, our work has important significance from the perspective of generalizing the effectiveness of IW methods. Alternatively, for personality estimation, Miranda-Correa et al. [56] used scores based on a self-assessment, and we used the same personality scores and classifier (L-SVM) as [56]. In addition, the experimental settings are aligned with the settings used in [56] to enable a proper comparison of model performance (F1-score). Therefore, to verify the effect of preprocessing and IW methods, we compared the model performance results between [56] and our study (Table 3.5). The macroaverages of the mean F1-score reported in [56] were 46.9% and 49.9% in the ECG model and GSR model, respectively, based on the short video scenario. Therefore, the average performance results of our model were higher than the performance results reported in [56], irrespective of IW. The preprocessing, such as waveform analysis and peak detection, conducted in this study was not substantially different from that employed in [56], and these are conventional methods; however, the feature set utilized for model building may be different. In this study, NN50 and pNN50 (Table 3.3) were extracted from the ECG data, and the standard deviation of SC and the 2nd difference of SC were extracted from the GSR data; however, they were not described in [56]. On the other hand, 60 spectral powers in the bands from [0-6] Hz are extracted as ECG features, and spectral powers in the bands from [0-2.4] Hz are extracted as GSR features in [56] but not in our study. These differences in feature extraction methods may influence the emotion of personality estimation performance.

3.5.4 Performance Comparison with a Deep Neural Network

Miranda-Correa et al. [56] constructed a DNN model with the AMIGOS dataset, and it was reported that the performance results of five personality factors (macroaverage of the mean F1-score) reached a maximum of 57.5% in binary classification (row 11 and column 18 in Table 3.5). In [32], it was mentioned that this model outperformed the best baseline mean F1-score result [70] by 2.7%, on average, when comparing the same five personality factors. In our study, although we did not use the EEG features, the IW-SVM based on GSR features outperformed [32] (57.5% F1-score) by 1.9% in the five personality factor estimations (row 7 and column 19 in Table 3.5). The EEG is relatively complicated measurement compared to ECG or GSR; therefore, applying IW method to the ECG or GSR shown in this chapter may be more practical. In addition, it was reported that the IW method improves the performance result of the model based on EEG in linear discriminant analysis in the BCI research area [74]. Although deep learning-based automatic personality recognition is one of the latest approaches [97], it is not always necessary to use a DNN, which requires greater computational resources for personality estimation, and it seems that feature engineering or the IW method is an option for performance improvement.

3.5.5 Importance-weighted Support Vector Machine in Speech Emotion Recognition

The IW-SVM model based on physiological signals may be effective not only with the AMIGOS dataset but also in other emotional/personality research. In speech emotion recognition (SER), the classification performance of a model can be degraded if the data acquisition environment and conditions are different. Hassan et al. [98] constructed an SER model using one of the two independently acquired datasets as training data and the other dataset as test data. In speech processing, cepstral mean normalization (CMN) and vocal tract length normalization (VTLN) are often utilized to compensate for channel and speaker differences. However, in [98], considering the differences between the datasets as a covariate shift, the IW-SVM algorithm was also applied to compensate for the differences known as domain adaptations. As a result, CMN + VTLN improved the binary classification accuracy by 3.4% compared to the standard L-SVM, but the IW method improved the accuracy by 5.5%. Although we considered physiological individual differences "within" the dataset as covariate shifts, considering the differences "between" the datasets as a covariate shift and compensating for this shift

with the IW method could also be effective.

3.5.6 Computational Complexity and Hyperparameters

We used KuLSIF for importance estimation. The computational cost of KuLSIF is approximately proportional to n^3, where the smaple size is n, and can be improved computationally and made more efficient by numerically minimizing the loss function (n^2), as shown in [95]. Although other methods, which estimate IW exist [92], KuLSIF yields lower computational costs than other estimators [95, 99]. Alternatively, the complexity of the standard SVM is typically n^3 [100] and that of the LR is nm^2 [101], where m is the number of features. Considering the computational complexity, IW estimation does not substantially affect the total computation time. However, IW estimation is needed for every test input, and applying this method to real-time response system is not feasible.

Regarding the hyperparameter, we set the flattening parameter $\gamma = 1$ in the experiment. To evaluate the effect of this hyperparameter, we have conducted additional analysis using the best model (emotion, IW-SVM fusion model; personality, IW-SVM GSR model) with setting $\gamma = \{0.1, 0.5, 1.0\}$. As a result, no improvement nor degradation was observed ($\pm 0.3\Delta\%$) in emotion estimation. On the other hand, the estimation performance degraded ($-2.1\Delta\%$) in personality estimation with the setting $\gamma = 0.1$ probably due to overflattening, which could diminish the effect of IW. Coupled with the evaluation of the other version of the domain adaptation, an effective hyperparameter search could be needed in the future.

3.5.7 Limitations and Future Works

The overall performance is improved by the IW method and this method outperforms the SOTA DNN model in personality estimation. Thus, we considered that IW methods are effective overall. However, the IW method does not always yield better results, as shown in Table 3.4 and 3.5, because the IW method assumes the distribution of the training input density and the test input density changes but the conditional distribution of the outputs given inputs remains unchanged. If these assumptions are violated, there can be no benefit of the IW method.

Also, IW method is not applicable to unseen individual test data since IW method require test input in advance. Thus, compensation by the IW method is started after measurement of physiological signals. These operation and

calibration time cause time lag in real-time systems; therefore, this is one of the limitations of this study.

The IW method is one of the domain adaptation methods and is referred to as a sample-based approach [72], which increases the penalty for the data samples that are more important in the learning process. More recently, other domain adaptation methods, such as deep domain adaptation approaches, have been developed [102]. Thus, other improved versions of the domain adaptation, which yield better results and are more stable, need to be considered.

Although we focused on the individual differences derived from physiological responses, effective data collection and preprocessing methods are also important issues. For example, [103] addressed the problem of the lossy EEG sensor communication pattern. Thus, a comparison and combination of these methods with IW is one of the directions for future works.

3.6 Chapter Summary

We constructed importance-weighted versions of LR and SVM and compared them to conventional models in emotion and personality estimations to evaluate whether the IW method can improve biosignal-based model performance. As a result, in emotion estimation, the IW-SVM outperforms the L-SVM by a 0.7% macroaveraged mean F1-score in the fusion model. Moreover, a performance improvement of 2.1% was achieved by the IW-SVM in the five personality factor estimation models based on the GSR features. These results are reasonable because the model's classification accuracy was similar to the accuracy based on BER estimation. Moreover, our results were comparable to those of a previously reported DNN model, which suggests the usefulness of the IW method in biosignal-based emotion and personality estimations. These results indicate that the IW method can reduce individual differences in peripheral physiological responses and can contribute to the proposal of a new model for emotion and personality estimations based on biosignals.

Chapter 4

Analysis of Physiological Signals toward Adaptive Dialogue Systems

4.1 Introduction

Recently, biosignals including EEG, ECG, and 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 [56], stress detection [104], and the provision of personalized recommendations [105]. However, the contribution of these biosignals in estimating a user's sentiment during dialogues remains unknown.

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 [106, 107, 108, 109], 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 chapter 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 4.6.

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 4.7.1.

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 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 4.7.2.

4.2 Related Works

In the affective computing domain, relationships between the emotional and nonverbal information, such as facial expressions, speech, gestures, and physiological states have been examined [14, 24]. In [56, 70], 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 (details of the is described in Section 3.2, Chapter 3). For emotion elicitation, these studies used videos that were classified into one of four quadrants of the valence arousal space. Kim and André [27] investigated the potential of physiological signals for emotion recognition by using biosensors such as EMG, ECG, EDA, and respiration sensors. As emotional stimuli, they used music that spontaneously induced real emotional states in the users. Kalimeri and Saitis [104] 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 [110, 111, 112, 113]. 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 [114, 115, 40]. In [40], 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 [41], 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. [42] 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 [116], 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 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 chapter, 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. [117] 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 and Wendemuth [118] showed that physiological signals, including EMG, 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.

4.3 Data

We used a multimodal dialogue corpus named Hazumi1911 [46], collected from November 2019. The recording setting was almost the same as that of Hazumi1712 [119] and Hazumi1902 [120], except physiological sensors were newly used in Hazumi1911. [1]

4.3.1 Data Collection

Figure 4.1 shows an overview of the study flow. Data were collected in the context of a human–agent dialogue as in [119], 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.

[1] Hazumi1712, 1902 and 1911 are publicity available [121].
[2] http://www.mmdagent.jp/

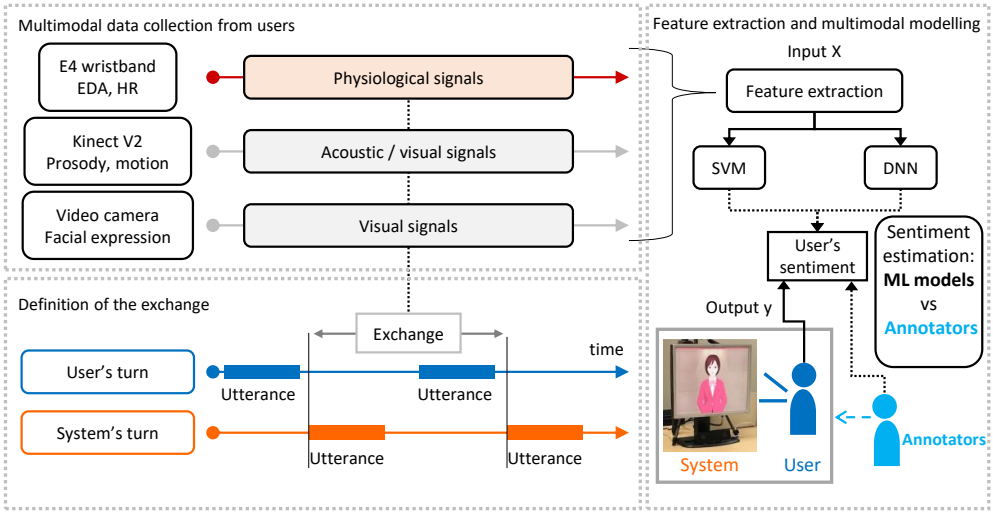

Figure 4.1: Overview of the estimation of the user's sentiment at the exchange level.

Basically, the operator selected the utterances of the agents from the predefined 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 as changes in the 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 (BVP) was obtained, and the HR was computed as the output from this device. In terms of the

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

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.

4.3.2 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.

4.3.3 Annotation

Two types of annotations were labeled in this study: (1) SS 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-reported 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 [116].

4.4 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 4.3.3. In this section, we describe the nonverbal features extracted from each modality.

4.4.1 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 [13]. 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 4.2). 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]https://pypi.org/project/PeakUtils/

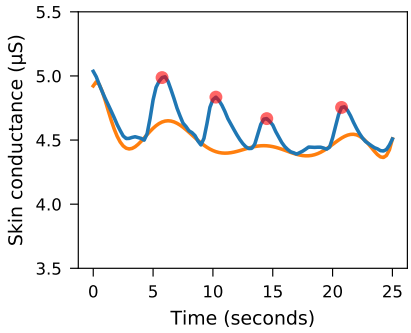

Figure 4.2: Example of SC analysis, showing the EDA signal (blue curve), tonic component (baseline, orange curve), and GSR (red circles).

4.4.2 Acoustic and Visual Features

Acoustic signals from the participant utterances were used to extract features. The INTERSPEECH 2009 Emotion Challenge feature set (IS09) [122] was extracted using the OpenSMILE[5] software. The types of acoustic features are as follows: root mean square frame energy, mel frequency cepstral coefficient 1-12, zero crossing rate from the time signal, voicing probability, and fundamental frequency (F0). To each of these, the delta coefficients are also calculated. Finally, 12 statistics: mean, standard deviation, kurtosis, skewness, minimum and maximum value, position of maximum and minimum values, range between maximum and minimum, slope and offset of a linear approximation with their mean square error are calculated and used as acoustic features. Thus, 384 ($16 \times 2 \times 12$) 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[123], the facial landmarks around the eye (4 points), mouth (4 points), and eyebrow (4 points) were determined, and the velocity (maximum, mean, and standard deviation) and acceleration (maximum) were calculated at each point as the facial tracking features. Also, the estimated categories of the facial action units described in [124] were used as the facial features. OpenFace can detect the presence of 18 types of action units in every frame. We use the proportion of occurrences of these 18 types of action units in each user's utterance as facial action unit features. We combine the facial tracking features and the facial action unit features and use these features as facial feature vectors.

The motion data of the hands, shoulders and head, recorded by the Mi-

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

crosoft 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.

4.5 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.

4.5.1 Machine Learning Models

Linear Support Vector Machine (SVM)

In the binary classification task, L-SVM models [91] 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.

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.

4.5.2 Evaluation Procedure

To evaluate the models, the cross-validation method (LOUOCV) was performed in the SVM and DNN models. In the LOUOCV, 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 SS annotation was 54.7%.

4.6 Experimental Result

Table 4.1 lists the estimation accuracy of the SVM models for the binary classification, and Table 4.2 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 4.1 and 4.2). In addition, acoustic features (A), visual features (V), and acoustic + visual features (A+V) set (columns 9 to 12 in Table 4.1 and 4.2) were used for comparison with physiological models.

The EF or LF technique was used to fuse the different modalities, as described in Section 4.5.1. 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 4.1 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 4.1. The best unimodal model is the physiological EDA subset (P_E) model (row 5 and column 2 in Table 4.1) 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 4.1), there is no improvement of estimation accuracy.

Performance of the DNN models: Table 4.2 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 4.2 in the same way as Table 4.1. The best unimodal model is the EDA subset (P_E) model (row 5 and column 2 in Table 4.2) 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 4.2), 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 4.1). 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.

Table 4.1: 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 4.2: 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					

4.7 Discussion

As shown in Section 4.6, 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 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.

4.7.1 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 4.3). 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 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 4.4, 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

		Estimated high	Estimated low
Human model	Actual high	38%	7%
	Actual low	31%	24%
DNN model	Actual high	28%	19%
	Actual low	18%	35%

Figure 4.3: Confusion matrix for binary classification showing the percentage of the total samples ($n = 2468$). upper: human model, lower: DNN model.

and estimate the negative sentiment labels for the adaptation of the dialogue systems.

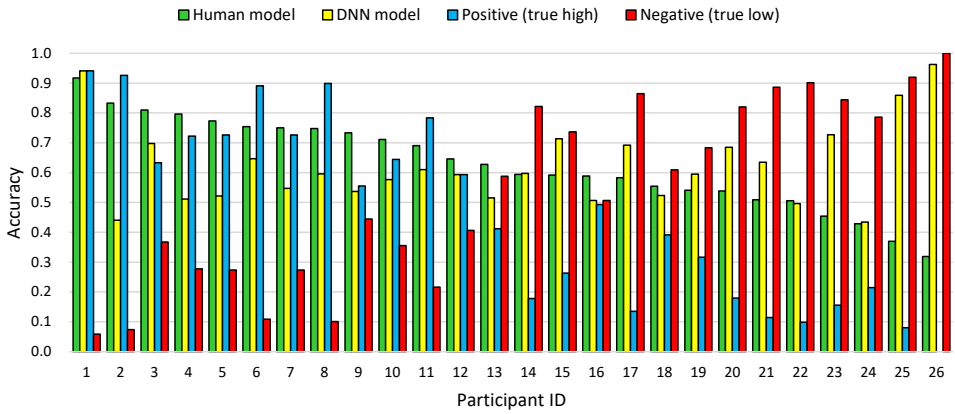

Figure 4.4: Estimation results for each participant in the LOUOCV (green bar: Human model, yellow bar: DNN model). As a reference, the proportion of the two classes for each participant's sentiment (grand truth) is also shown (blue bar: positive sentiment (true high), red bar: negative sentiment (true low)).

4.7.2 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

Table 4.3: 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**

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 4.3, $r = 0.168$).

Figure 4.5 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 [14, 27, 13]. 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 4.6 (left) shows the relative frequency of the sentiment score in each group (Q1: lower quartile, Q4: upper quartile) and indicates the differences in the sample distribution along with the GSR number. There was a clear difference between quartile groups in the sentiment score of 6.

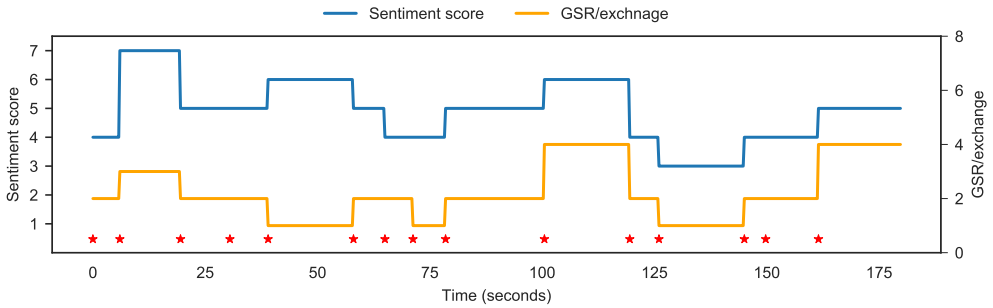

Figure 4.5: Example of dynamic changes in the participant's sentiment and GSR number during the dialogue. The sentiment score (blue line, left y axis) and GSR per exchange (orange line, right y axis) are shown. The red stars indicate the timing of the system utterance (Participant ID 11 in Figure 4.4, dialogue data from the start time to 15th exchange).

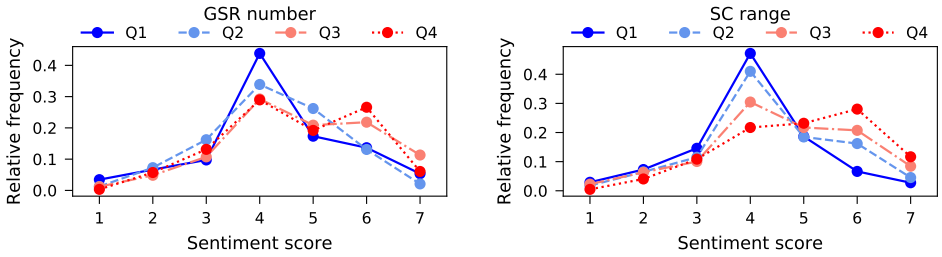

Figure 4.6: Relationship between the sentiment score of the participant and EDA features in each exchange. The samples of the participants' sentiment score were divided into quartile groups based on the quantile of the GSR number (left) or SC range (right). The relative frequency of the sentiment score in each quartile group is shown.

A similar distribution was observed in the quartile group of the SC range (Figure 4.6, right). Thus, these EDA features were expected to contribute to the sentiment estimation.

4.7.3 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. [125] proposed an autonomous real time adap-

tation 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 ANS. In addition, the presence of individual differences in physiological signals could lead to a performance degradation. A method known as covariate shift adaptation [16], 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.

4.8 Chapter Summary

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.

Chapter 5

Different Types of Multimodal Sentiment Estimation

5.1 Introduction

Unobservable signals, i.e., physiological signals as key information for capturing SS that cannot be detected from only observable information (i.e., text, audio, and visual information), should be considered for sentiment estimation, since such physiological signals could potentially include valuable information for estimating SS. In addition, there is a need to clarify the modalities that can effectively estimate SS, including combinations of modalities, and investigate the contributions of observable and unobservable modalities to improving estimation performance. Furthermore, a multimodal dataset that includes different types of sentiment labels, that is, both SS and TS labels, and is annotated at a conversational exchange level in naturalistic human-agent interaction settings is needed. Here, the "exchange" consists of a system utterance followed by a user utterance.

In this chapter, we investigate effects of physiological signals in multimodal sentiment analysis and evaluate all of the fusion models for different types of sentiment estimation on naturalistic human-agent interaction settings. These analyses provide information on the types of modalities and fusion methods that are effective in each task to create an adaptive dialogue system.

For this purpose, we used the multimodal dialogue corpus Hazumi1911, which contains conversational exchanges recorded in human-agent interaction settings, and is identical to the dataset used in Chapter 4. The entries in Hazumi1911 consist of sequential sentiment labels (SS and TS) for each exchange and textual, audio, visual and physiological information. This dataset

is the first to enable the construction and evaluation of multimodal fusion models with physiological signals that can estimate both SS and TS based on detected short-time multimodal signals during dialogue. Thus, this dataset also provides new insights on the roles of multimodality with physiological signals in SS and TS estimation.

Toward creating an adaptive dialogue agent, our study aims to demonstrate the importance of multimodal signals with physiological signals in sentiment analysis and evaluate all of the fusion models in human-agent interaction settings. Our main contributions are as follows:

- We tackle SS estimation problems with physiological signals and extensive baseline in multimodal dialogue settings between humans and agent for the first time (Section 5.6.1).
- We clarify the effects of physiological signals in exchange level SS and TS estimation by comparing other modalities (Section 5.6.1 for SS estimation and Section 5.6.2 for TS estimation).
- We present a comparison of multimodal language models based on the Automatic Speech Recognition (ASR) and manually transcribed data (Section 5.6.3), the effective physiological features for SS estimation (Section 5.6.4), and a comparison of the SS estimation patterns generated by the linguistic model and the physiological model (Section 5.6.5). These analyses provide knowledge for implementing systems and improving performance in multimodal dialogue modeling with physiological signals.

The remainder of this chapter is organized as follows: Section 5.2 reviews research related to our work in the domains of natural language processing (NLP), affective computing and multimodal analysis. Section 5.3 describes the data collection methods used for the multimodal dialogue corpus, and Section 5.4 describes the extraction methods used to obtain the multimodal features and linguistic representations. The experimental settings, including the multimodal model architecture and evaluation schema, are described in Section 5.5. Section 5.6 presents the experimental results, which are further discussed in Section 5.7. Section 5.8 concludes our work.

5.2 Related Works

5.2.1 Text-based Sentiment Analysis

Text-based approaches play a central role in sentiment analysis. Although conventional lexicon- or handcrafted feature-based sentiment analysis meth-

ods and various datasets have been reported, this subsection focuses on recently developed SOTA neural network models in the NLP domain.

Neural network models have recently become a popular approach in sentiment analysis [126]. High performance of CNNs, Long Short-Term Memory (LSTM) networks and their variants has been reported. Kim [127] reported simple CNN models that achieved SOTA performance on 4 of 7 datasets, including the Stanford Sentiment Treebank v2 (SST-2, [128]) dataset (88.1% accuracy). Peters et al. [129] proposed deep contextualized word representations based on a Bidirectional Long Short-Term Memory (BiLSTM) approach called embeddings from language models (ELMo) representations. Using ELMo representations for downstream tasks, new SOTA results were achieved on six NLP tasks, including sentiment analysis. As an alternative approach, Vaswani et al. [130] developed a simple new network architecture, called a transformer, that contains no recurrence or convolutions. Based solely on an attention mechanism, transformers achieved significant improvements in computational efficiency through parallelization and improved performance for machine translation tasks. Following [130], Devlin et al. [131] developed BERT, a language model based on a multilayer bidirectional transformer encoder, which advanced the state of the art on eleven NLP tasks (94.9% accuracy on the SST-2 dataset). BERT has become the de facto standard for NLP tasks, but it seems that such powerful tools and models can contribute to estimation performance during dialogue if the user's sentiment is explicitly expressed as textual information. Since many factors influence sentiment expression and spoken language is noisier and less structured than written language [132], relying solely on linguistic information may not be sufficient, and there is a need to consider a multimodal approach to understanding user sentiment.

5.2.2 Physiological Signal-based Sentiment Analysis

To date, studies on emotion recognition and sentiment analysis have mainly focused on textual, audio and visual modalities; however, physiological signals have also been considered. Previous extensive emotion recognition studies based on physiological signals have provided insights on physiological aspects. Soleymani et al. [64] reported a multimodal database recorded in response to affective stimuli induced by emotional videos, including facial expressions, audio, eye gaze data, EEG signals and peripheral physiological signals. Twenty-seven participants were recruited and self-reported their felt emotions using arousal, valence, dominance, predictability and emotional keywords. More recently, the AMIGOS dataset, which includes audio, visual, depth, EEG, EDA and ECG data for 40 participants, was created [56]

(please see Section 3.2 in Chapter 3). The data were collected during a movie-watching task, and the levels of arousal and valence were annotated by not only the participants themselves but also a third party. The Big Five personality traits [133] and the Positive and Negative Affect Schedule (PANAS) [134] were also measured. Although the data in these studies were collected under emotional stimuli, they show the potential uses of physiological signals in emotion-related research.

The REmote COLlaborative and Affective interactions (RECOLA) corpus [44] was collected under naturalistic (spontaneous) conditions. The participants were recorded in dyads while video conferencing for a collaborative task. The RECOLA corpus includes audio, visual, ECG and EDA data. Annotations of affective (arousal and valence) and social behaviors were provided by the participants (3 time points) and by a third party (first 5 minutes of interaction). The RECOLA corpus has been extensively used for benchmarking emotion recognition and reporting the effectiveness of considering individual modalities, including physiological signals (e.g., [48, 135, 29] evaluated the recognition performance for naturalistic affective states using peripheral physiological signals (ECG, EDA and EMG) in a human-agent interaction setting. The participants interacted with a tutoring system called AutoTutor. AutoTutor's dialogues are organized around difficult questions and evoke naturalistic affective states such as confusion and frustration. The participants annotated their affect during each 20-second interval. The researchers focused on the detailed differences between user-independent and user-dependent models, and classifiers such as support vector machine and NB classifiers were used for evaluation. The results suggested that the user-dependent models based on naturalistic physiological data were feasible, whereas the user-independent models were not accurate, possibly due to individual differences.

The human-computer interface literature has also explored multimodal measures of user engagement [136], which are closely related to user sentiment during dialogue. To obtain more objective data, physiological measures have been incorporated into approaches for measuring user engagement [137].

Although there may be cases in which the use of naturalistic physiological signals in a unimodal model is insufficient for estimating SS, the potential of exploiting complementarity between different modalities, such as linguistic and physiological information, to improve estimation performance has not been considered in previous studies.

5.2.3 Multimodal Dialogue Systems

The multimodal approach has attracted attention as a means of recognizing the internal states of users during human-agent interactions [3]. McKeown et al. [43] created a large database dedicated to emotionally colored conversations called SEMAINE that includes audiovisual data collected during interactions between users and an agent. The recordings represent 150 participants and were annotated with five affective dimensions, including valence, by raters. Tomimasu and Araki [41] considered each exchange between a system and a user as a unit in a chat dialogue and assessed user interest based on audiovisual information. Tavabi et al. [42] proposed a multimodal learning framework for identifying opportunities for an agent to express empathetic responses to achieve engaging social interactions. Textual, audio and visual features were used to construct a multimodal neural network with Gated Recurrent Units GRUs.

In our previous work, we created two multimodal chat-dialogue corpora called Hazumi1712 [119] and Hazumi1902 [120] that include user interest labels, user sentiment labels, and topic continuance annotated at the exchange level. Textual, audio and visual data were recorded in human-agent interaction settings. Considering the relationships among the labels, we applied a multitask learning technique for binary classification tasks and demonstrated that a multitask DNN model trained on multimodal features outperformed a single-task DNN [116]. More recently, we collected another multimodal dialogue corpus, called Hazumi1911, under almost the same recording settings as for Hazumi1712 and Hazumi1902 except that physiological sensors were additionally used for Hazumi1911. Using the Hazumi1911 dataset, the contribution of physiological signals for estimating SS was compared to that of audiovisual data, and the effectiveness of considering such nonverbal information for sentiment analysis was demonstrated as reported in [138] and Chapter 4.

In this study, we extended the work described in Chapter 4) via a multifaceted approach using linguistic, audio, visual and physiological information to obtain a comprehensive understanding of the effects of each modality in sentiment analysis. We investigated the linguistic contribution to multimodal user sentiment analysis by using the SOTA language models fastText [139] and BERT [131]. Then, we used two different types of sentiment labels, that is, sentiment labels annotated by the users themselves (i.e., SS label) or by third parties (i.e., TS label), for the sentiment estimation task. These settings enabled an investigation of the differences in the contributions of "observable" (i.e., linguistic, audio, visual) and "unobservable" (i.e., physiological) signals on both the "self" and "other" axes for the first time.

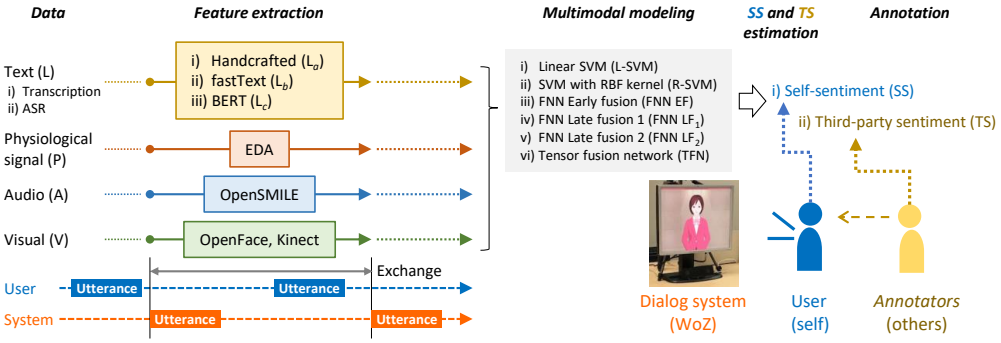

Figure 5.1: Overview of the estimation of SS and TS at the exchange level.

5.3 Data

Our proposed framework is depicted in Figure 5.1. The multimodal dialogue corpus Hazumi1911[1] is also introduced in this chapter. We propose that physiological signals as well as textual, audio and visual data recorded from a participant during a dialogue serve as multimodal data. An "exchange" is defined as a segment that begins at the start time of a system utterance and ends at the start time of the next system utterance. Feature extraction was performed for each exchange, and the extracted features were fed into each model, including the multimodal neural network. Then, binary classification and regression tasks were conducted using SS and TS. This section describes the Hazumi1911 dataset.

5.3.1 Dialogue Settings

Hazumi1911 was collected in the human-agent dialogue context described in [119, 46], in which the participants communicated with a virtual agent known as MMDAgent[2] shown on a display. The agent was operated using the Wizard of Oz (WoZ) method. Specifically, a human operator (the Wizard) remotely controlled the system and interacted with the participants from 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 during the dialogue; i.e., the participants simply chatted with the agent. About a dozen of topics such as food preference, traveling, and movies were prepared for the dialogue, so no particular edu-

[1]Hazumi1911 is publicly available [121].
https://github.com/ouktlab/Hazumi1911/

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

Table 5.1: Dataset summary.

Number of participants	26
Average dialogue duration	20.5 min
Average number of exchanges	95
Total dialogue duration	534.0 min
Total number of exchanges	2468
Sentiment ratings	Discrete values of 1 to 7

cation level was assumed. A well-trained single Wizard-of-Oz operator was engaged in operating the virtual agent. The Wizard selected an utterance prepared prior to the study and tried to make participant enjoy the conversation and want to continue talking. Specifically, the Wizard changed the topic if a participant seemed bored or took the role of the listener if the participant seemed to enjoy talking.

The dataset used in this study is summarized in Table 5.1. Thirty participants (between the ages of 20 and 70; 15 male and 15 female) were recorded, and data from 26 of the participants were used for analysis; the data of four participants were disregarded because of missing values after preprocessing.

This study was reviewed and approved by the Research Ethics Committee of the Institute of Scientific and Industrial Research, Osaka University. All participants provided written informed consent to participate in the study. More details of the dialogue settings are described in [46].

5.3.2 Sensors

EDA, measured as SC, reflects sweat gland activity as part of the sympathetic nervous system and has previously been used to detect changes in emotional states [14, 27, 140]. EDA data were collected as physiological signals during the dialogues using a physiological sensor, namely, the Empatica E4 wristband (Empatica Inc., Cambridge, MA, USA)[3] developed by Empatica Inc., which arose from MIT research [24]. This device can detect changes in SC by means of two electrodes in contact with the skin and has been well validated[4]. Furthermore, because the E4 device is wireless and worn like a wristwatch, it causes neither disturbance nor discomfort, which is a top priority for naturalistic dialogue. Thus, this device is suitable for investigating a participant's sentiment and for comparing the physiological modality

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

[4]https://support.empatica.com/hc/en-us/articles/203005295-Have-you-done-comparative-studiesor-validation-on-Electrodermal-activity-sensor-

with other modalities during naturalistic dialogue. The EDA signals of the participants were recorded at 4 Hz.

Regarding the textual and audio signals, the vocal utterances of each participant were recorded as a 16 kHz waveform audio format (WAV) file using a Microsoft Kinect V2 sensor. The system's utterances, which the Wizard selected during each dialogue, were also logged. Regarding 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. Since the sampling rates of the nonverbal modalities differed, statistics such as the average and maximum of the features were calculated in each exchange and used for feature extraction.

5.3.3 Annotation

One of the definitions of engagement is an attitude that determines the quality of interaction [141]. In contrast, the rating scale that we adopted in this study includes negative sentiments such as "confusion" or "dissatisfaction". We defined sentiment as a particular type of affective experience that is suitable to represent these negative aspects, and the definition was predetermined during the planning of this study. Thus, we considered our labels as positive/negative sentiment labels that represent the emotional aspect of the participant. Two different types of annotation were performed in this study, namely, (1) SS annotation and (2) TS annotation, which were performed by the participants themselves and by third-party coders, respectively. SS and TS were used for the sentiment estimation task described in Section 5.6.1 and Section 5.6.2, respectively, to investigate the contribution of each modality to each task. A total of 2468 exchanges from 26 participants were annotated. We used a 7-point Likert scale for annotation, which is identical to the method used in multimodal sentiment research for datasets such as the MOSI and MOSEI datasets [8, 9]. The annotation procedures were as follows:

(1) Self-reported sentiment annotation: The participants annotated the labels for each of their own exchanges while watching their own videos after the experiment. Only one label was annotated by the SS annotation process in each exchange, and there were not different types of SS labels. The labels were assigned as scores ranging from 1 (no enjoyment of the dialogue) to 7 (enjoyment of the dialogue). The positive sentiments included "enjoy talking", "want to continue talking", and "satisfied with the conversation", and the negative sentiments included "want to stop talking" and "confused about the system utterances".

(2) Third-party sentiment annotation: Five human coders annotated

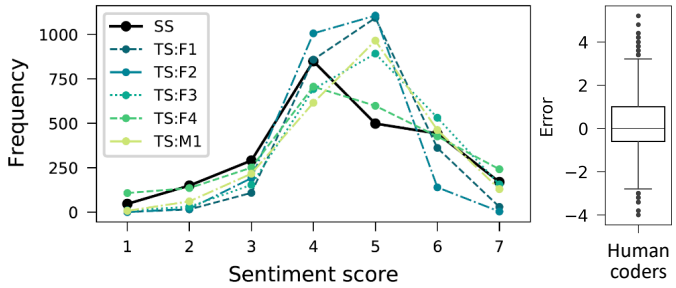

Figure 5.2: (left) Distribution of the sentiment scores among the participants (SS) and five third-party coders (TS). (right) Estimation error of the third-party coders.

each exchange with labels assigned as scores ranging from 1 (participant seems to be bored with the dialogue) to 7 (participant seems to enjoy the dialogue) while watching the recorded videos of the dialogues. The third-party coders were instructed to assign labels by considering the participants' linguistic, audio and visual information. They were also instructed not to assign labels by considering only a part of the exchange and to assign labels after watching the entire recording of the target participant.

The agreement between the third-party coder ratings was calculated using Cronbach's alpha. Generally, a Cronbach's alpha of > 0.8 indicates high consistency between the annotated labels. The Cronbach alpha value was 0.83 for the TS annotation, indicating the reliability of the annotation in this study. Finally, we averaged the values from the five coders for use as the TS label.

Figure 5.2 (left) shows the distribution of the SS and TS labels. The third-party coders tended to annotate the participants' sentiment as positive. Figure 5.2 (right) shows a box plot of the differences in sentiment scores between third-party coders (TS) and participants (SS), i.e. human estimation error. An upward bias is apparent, which indicates that from the viewpoint of the third parties, the participants seemed to enjoy the dialogue, but the participants actually had neutral or negative sentiments.

5.4 Features and Representations

5.4.1 Linguistic Feature Extraction

Three linguistic feature sets were prepared for the construction of multimodal models: a handcrafted feature set based on conventional methods, a set of

word representations from fastText, and a set of sentence representations from BERT. The method used to construct each feature set is presented in this subsection.

Handcrafted Features

The participants' utterances were manually transcribed into text data. The texts were then segmented into words by using the Japanese morphological analysis tool MeCab [142]. The utterance token distribution with respect to parts of speech (PoSs) is summarized in Appendix (Table S1). The linguistic features extracted from each participant's utterances were as follows: the frequencies of words based on bag-of-words (BoW) representation, the frequencies of the PoSs of the words (noun, verb, adjective, adverb, and interjection), the frequency of filler, and disfluency. The polarity of each participant utterance (positive, negative, or neutral) was also evaluated using oseti[5], a sentiment analyzer for Japanese based on sentiment polarity dictionaries [143, 144]. The polarity score [-1, 1] of a participant's utterance and the numbers and proportions of positive and negative words were additionally used as features.

The utterances of the dialogue system could affect a participant's sentiment. Thus, the system's utterances were also extracted from the recorded dialogue log data. The system's utterances were classified into eight dialogue actions (providing information, negative answers, positive answers, other types of answers, starting new topics, wh-questions (six types of wh-questions), yes/no questions, and proposals) and represented as eight dimensional one-hot vectors.

The duration of and the number of words in each utterance of the participants and the system were also extracted. The differences in these factors between the participants' and system's utterances were also calculated as features. The data were normalized per participant by means of Z-score normalization to ensure a mean of zero and a standard deviation of one for all samples from one participant.

FastText word vectors

FastText is a recently developed language model for training word vectors with subword information on Wikipedia and the Common Crawl corpus [139]. This language model is not restricted to English and is available for 157 languages, including Japanese. FastText shows strong performance compared with previous word vector models. Thus, to evaluate other baselines for text

[5]https://github.com/ikegami-yukino/oseti

modality, fastText word vectors were used in this study. The utterance sequences in each exchange were first tokenized by MeCab and subsequently represented by using a pretrained Japanese fastText model[6]. The word vectors in each exchange were averaged, resulting in a single vector with a length of 300.

BERT Representations

BERT is a language representation model that achieves SOTA performance on many NLP tasks [131]. Language model pretraining is important for performance improvement [145, 131], and a pretrained Japanese BERT model[7] has recently been developed at Tohoku University. The pretrained Japanese BERT has shown superior performance to conventional models based on the bag of words in tweet emotion recognition [146]. We used this pretrained Tohoku BERT model in this study. The participant's and system's utterances in each exchange were separated by a special token (`[SEP]`). The utterance sequences were first tokenized by MeCab and split into subwords by the WordPiece algorithm. Then, the sequences were represented by extracting the activations from the second-to-last hidden layer of the BERT model and average pooling, resulting in a single vector with a length of 768 as described in [131]. Finally, this vector was used as the input feature vector for each models. This approach does not require the extraction of complicated handcrafted features and enables easy fusion with other modalities.

5.4.2 Physiological Feature Extraction

Physiological feature extraction method is identical to Chapter 4. In brief, the recorded time series SC data were decomposed into the SC level (the baseline, also known as the tonic component) and the SC response (also known as the GSR). The baseline SC, which reflects the general activity of the perspiratory glands due to the ambient temperature [147, 27], was calculated using polynomial fitting (degree of 10). Then, the GSR was detected using PeakUtils[8] (amplitude threshold of 0.3). Finally, the GSR number per exchange was extracted as a physiological feature. Moreover, the following statistics of the SC in each exchange were calculated and used as physiological features: the 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

[6]https://fasttext.cc/
[7]https://github.com/cl-tohoku/bert-japanese
[8]https://pypi.org/project/PeakUtils/

approximation, and 25th and 75th percentile values. Overall, 14 physiological features were extracted from each exchange. The data were normalized using min-max normalization to a range of zero to one.

5.4.3 Audio/Visual Feature Extraction

The INTERSPEECH 2009 Emotion Challenge feature set (IS09) [122] was extracted using the OpenSMILE[9] software to serve as audio features. The features were calculated as statistics, and 384 acoustic features in total were extracted from each exchange. Using the OpenFace library[123], facial landmarks around the eyes, mouth, and eyebrows were determined, and the velocity and acceleration at each point were calculated for facial feature extraction. The estimated categories of facial action units as proposed by [124] were also used as facial features. Motion data of the hands, shoulders and head recorded by the Microsoft Kinect sensor were additionally considered, and the calculated velocity and acceleration were used as motion features. In total, 86 visual features were extracted from facial expressions and motion activity. The data were normalized per participant by means of Z-score normalization.

5.5 Experimental Settings

In this study, we refer to the estimation task using SS as SS estimation and to the estimation task using TS as TS estimation. These estimation tasks were evaluated as both binary classification and regression tasks. For binary classification, the sentiment labels were divided into positive and negative classes considering a threshold of 4 (neutral state). The numbers of SS labels in the positive and negative classes were 1119 and 1349, respectively. Similarly, the TS labels were divided into positive and negative classes containing 1701 and 767 labels, respectively. For regression, the sentiment scores in the range of 1 to 7 were used for estimation. Each model used for the estimation task is described in this section.

5.5.1 Machine Learning Algorithms

Support Vector Machines

SVM is a popular and conventional learning algorithm for estimation tasks [91]. L-SVM and Support Vector Machine with Radial basis function kernel

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

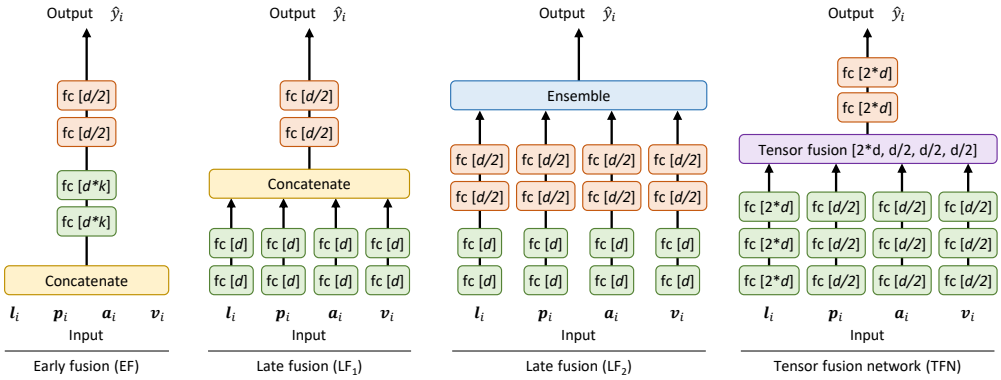

Figure 5.3: Summary of the multimodal model architecture. The case of quadmodal models is shown. $\boldsymbol{l}_i$, $\boldsymbol{p}_i$, $\boldsymbol{a}_i$ and $\boldsymbol{v}_i$ indicate linguistic, physiological, audio and visual feature vectors of the ith exchange, respectively. A fully connected layer is denoted as fc. Brackets indicate the dimension, and we use $d = 64$ in this study. k in early fusion indicates the number of modalities (i.e., $k = 4$ in quadmodal model).

(R-SVM) were used as baseline models in this study.
(1) Linear SVM: The L-SVM models were optimized using a threefold cross-validation scheme for the training data set with the penalty parameters set as {0.01, 0.1, 1, 10} in the classification task and with the insensitivity parameters set as {0, 0.5, 1, 1.5} in the regression task.
(2) SVM with radial basis function kernel: The R-SVM models were optimized using a threefold cross-validation scheme for the training data set with the penalty parameters set as {0.01, 0.1, 1, 10, 100} and with the kernel parameters set as {0.001, 0.0001} in both the classification and regression tasks.

Feedforward Neural Networks

The Feedforward Neural Network (FNN) architecture was used to construct unimodal and multimodal neural networks. Figure 5.3 shows a summary of the multimodal model architecture used in this study.

For training on a unimodal feature set, the FNN was composed of an input layer, two intermediate layers with 64 units, two intermediate layers with 32 units, and an output layer. The rectified linear unit (ReLU) activation function was used. Dropout was applied after each layer to reduce overfitting. To investigate the contributions of each modality to SS and TS estimation, we used EF and two kinds of late fusion (LF_1 and LF_2) [148].

(1) Early fusion: This is the simplest fusion technique. The feature vectors from different modalities are concatenated into a single feature vector, which is then fed to an FNN as the input feature vector. The architecture used in the EF configuration was similar to that used in the unimodal configuration, except that it also included two intermediate layers with 128 units for the bimodal features and a layer with 192 units for the trimodal features.
(2) Late fusion 1: The FNN structure for LF_1 consisted of a lower block and a higher block. For the lower block, a neural network with an input layer and two intermediate layers with 64 units were prepared to extract each unimodal feature type. For the higher block, the output units of each unimodal model were concatenated, and the layer with the concatenated units was then connected to two hidden layers with 32 units, followed by an output layer.
(3) Late fusion 2: In LF_2, fusion is performed after the calculation of an output value derived from each unimodal model (i.e., ensemble). The other hyperparameters were set as follows: a batch size of 32, a dropout rate of 0.3, a learning rate of 0.001 (adaptive moment estimation (Adam) optimizer), and a number of epochs of 30 (classification) or 100 (regression).

Tensor Fusion Networks

Several multimodal neural networks have been proposed for sentiment analysis [6, 149]. These models were proposed for TS estimation (sentiment estimation using opinionated monologue videos) and thus might also work well in estimation of TS derived from human-agent interactions. It is interesting to investigate whether these models are also effective for SS estimation. Thus, we implemented the Tensor Fusion Network (TFN) developed in [149]. Following [149], the lower block had an input layer and three intermediate layers (Figure 5.3). The tensor fusion layer is defined as follows:

$$\boldsymbol{z}_i^m = \begin{bmatrix} \boldsymbol{z}_i^l \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \boldsymbol{z}_i^p \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \boldsymbol{z}_i^a \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \boldsymbol{z}_i^v \\ 1 \end{bmatrix}$$

Here, $\boldsymbol{z}_i^m$ indicates an outer product between linguistic representations $\boldsymbol{z}_i^l$, physiological representations $\boldsymbol{z}_i^p$, audio representations
$textit\mathbf{z}_i^a$, and visual representations $\boldsymbol{z}_i^v$ in the ith exchange. Each representation is derived from the third intermediate layer of the lower block. The extra constant dimension with value 1 generates the unimodal, bimodal and trimodal dynamics. The 4-D tensor $\boldsymbol{z}_i^m$ was flattened and then connected to two hidden layers with 128 units, followed by an output layer.

5.5.2 Evaluation Procedure

To evaluate the models, a cross-validation method (LOUOCV) was applied. In LOUOCV, the samples corresponding to each exchange between one participant and the 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 from the training dataset, thereby avoiding overestimation. The accuracy and macro F1-score were calculated for each evaluation. The average accuracy and average F1-score were reported as the model performance indicators for the binary classification tasks. Similarly, the average of the Mean Absolute Error (MAE) was reported for the regression tasks. All experiments were performed five times with random initialization, and the evaluation values were calculated as averages across the five repetitions. These evaluation values were then compared among the models based on each modality or combination of modalities. The majority baseline for binary classification was 54.7% for the SS labels and 68.9% for the TS labels.

5.6 Results

First, we present the effectiveness of the physiological signals in SS estimation to compare other modalities in this task. This is shown in Section 5.6.1. The results indicate that the unobservable signal, i.e., the physiological (P) modality, is the most effective for SS estimation among unimodal models. It is also shown that our proposed model fusing linguistic representations with physiological features outperforms the best previously reported model.

Second, we show the contribution of each modality in different types of sentiment estimation, i.e., TS estimation, which is contrasted with the SS estimation result. The TS estimation results are presented in Section 5.6.2. In contrast to SS estimation, the linguistic model provides the best result for TS estimation, and the tensor fusion of linguistic representations with audiovisual features, i.e., fusion of the observable signals, is the most effective for TS estimation in the regression task.

Third, we present comparisons of multimodal language models based on ASR and manually transcribed data (Section 5.6.3) to consider the end-to-end model. We also present effective physiological features for SS estimation in Section 5.6.4, and finally, to investigate the effects of physiological signals in the SS estimation task, differences in the SS estimation patterns between the linguistic model and the physiological model are presented in Section 5.6.5.

The following feature sets were used as the basis for unimodal models: L_a, handcrafted linguistic features; L_b, fastText word vectors; L_c, BERT representations; P, physiological features; A, audio features; and V, visual features. The multimodal models were constructed using either an EF, LF or tensor fusion technique. The plus (+) sign indicates each modality fusion, for example, L_c+P indicates BERT representations + physiological features; L_c+A+V indicates BERT representations + audio + visual features; and L_c+P+A+V indicates all fusion. All of the fusion models with L_a or with L_b are listed in Appendix (Tables S2 and S3) to compare the language representation methods.

5.6.1 Self-reported Sentiment Estimation

The effectiveness of the physiological signal in SS estimation is evaluated in this subsection. The results for both classification (Accuracy (Acc) and macro F1-score (F1)) and regression (MAE) are listed in Table 5.2 for unimodal models and Table 5.3 for multimodal models.

(1) Unimodal models: Among the unimodal models, for the classification task, the physiological (P) FNN model achieves the best accuracy of 0.619, and L-SVM trained on the BERT representation (L_c) has the best F1-score of 0.541 (Table 5.2). For the regression task, the physiological (P) FNN model also achieves the best MAE of 1.082. Overall, the physiological (P) modality seems to be the most effective for SS estimation. The unimodal models trained on handcrafted linguistic features (L_a), fastText word vectors (L_b), audio features (A) and visual features (V) exhibit lower performance than the physiological (P) model or the BERT representation (L_c) model for both the classification and regression tasks, except the R-SVM model trained on L_a, which shows performance equal to the physiological (P) FNN model in the regression task.

Table 5.2: SS Estimation results for unimodal models. L_a, handcrafted linguistic features; L_b, fastText word vectors; L_c, BERT representations; P, physiological features; A, audio features; and V, visual features.

Feature	L-SVM						R-SVM						FNN					
	Acc		F1		MAE		Acc		F1		MAE		Acc		F1		MAE	
	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD
L_a	0.571	0.000	0.507	0.000	1.243	0.000	0.404	0.000	0.281	0.000	**1.082**	0.000	0.584	0.007	0.521	0.008	1.127	0.008
L_b	0.580	0.004	0.527	0.003	**1.100**	0.000	0.478	0.000	0.364	0.000	1.103	0.000	0.582	0.010	0.523	0.006	1.120	0.014
L_c	**0.601**	0.005	**0.541**	0.006	1.162	0.000	0.469	0.000	0.424	0.000	1.102	0.000	0.602	0.004	**0.537**	0.005	1.111	0.004
P	0.600	0.000	0.491	0.000	1.136	0.000	0.496	0.000	0.392	0.000	1.088	0.000	**0.619**	0.011	0.499	0.016	**1.082**	0.017
A	0.561	0.004	0.509	0.004	1.322	0.000	**0.531**	0.000	**0.475**	0.000	1.143	0.000	0.567	0.013	0.507	0.010	1.204	0.010
V	0.568	0.001	0.518	0.001	1.342	0.000	0.517	0.000	0.464	0.000	1.103	0.000	0.560	0.002	0.508	0.004	1.175	0.005

(2) Multimodal models: Since the BERT representation (L_c) model shows higher performance than the model trained on handcrafted linguistic features (L_a) and fastText word vectors (L_b), we report the BERT representation (L_c) model for further investigation of multimodal neural network models (please see Table S2 in the Appendix for all of the fusion models with L_a or L_b). Among the multimodal models, the L_c+P FNN model with LF_2 achieves the highest classification accuracy of 0.637, outperforming the best unimodal model by 0.018 (Table 5.3). Furthermore, on the regression task, this fusion model also yields the best MAE of 1.041. On the other hand, the L_c+P+V FNN model and L-SVM trained on all modalities achieves the highest F1-score of 0.557. The TFN model did not outperform FNN with LF_2 model.

In summary, the combination of BERT representations and physiological features is the most effective approach for SS estimation in this experiment.

Table 5.3: SS estimation results for multimodal models.

Feature	L-SVM			R-SVM			FNN EF			FNN LF_1			FNN LF_2			TFN		
	Acc	F1	MAE	Acc	F1	MAE	Acc	F1	MAE	Acc	F1	MAE	Acc	F1	MAE	Acc	F1	MAE
L_c+P	0.608	0.527	**1.151**	0.526	0.462	1.100	**0.615**	**0.548**	**1.088**	0.630	0.542	**1.087**	**0.637**	0.554	**1.041**	**0.615**	**0.543**	1.089
L_c+A	0.583	0.529	1.206	0.553	0.494	**1.094**	0.587	0.528	1.141	0.576	0.519	1.150	0.593	0.532	1.094	0.599	0.538	1.143
L_c+V	0.609	0.551	1.264	0.552	0.490	1.095	0.599	0.540	1.129	0.586	0.531	1.143	0.597	0.539	1.085	0.598	0.534	1.112
P+A	0.594	0.517	1.270	0.543	0.480	1.113	0.575	0.507	1.153	0.593	0.515	1.149	0.599	0.524	1.081	0.588	0.514	1.086
P+V	0.613	0.521	1.278	0.531	0.472	1.100	0.587	0.511	1.106	0.625	0.537	1.108	0.618	0.534	1.077	0.603	0.519	**1.070**
A+V	0.574	0.526	1.228	0.538	0.478	1.106	0.573	0.515	1.162	0.571	0.515	1.169	0.581	0.522	1.131	0.567	0.516	1.115
L_c+P+A	0.607	0.529	1.197	0.519	0.454	1.095	0.580	0.523	1.127	0.611	0.534	1.111	0.627	0.551	1.050	0.602	0.532	1.106
L_c+P+V	0.615	0.541	1.309	0.553	0.489	1.096	0.593	0.532	1.120	**0.632**	**0.550**	1.109	0.633	**0.557**	1.048	0.609	0.539	1.103
L_c+A+V	0.589	0.532	1.221	0.537	0.471	1.131	0.585	0.528	1.120	0.583	0.526	1.145	0.600	0.540	1.082	0.597	0.536	1.139
P+A+V	0.605	0.533	1.235	0.521	0.465	1.102	0.584	0.524	1.145	0.602	0.527	1.138	0.616	0.542	1.075	0.603	0.527	1.095
L_c+P+A+V	**0.628**	**0.557**	1.198	**0.578**	**0.517**	1.121	0.585	0.528	1.116	0.606	0.535	1.116	0.625	0.554	1.052	0.608	0.537	1.106

5.6.2 Third-party Sentiment Estimation

The contributions of each modality are evaluated using TS, which is annotated not by "self" but by "others". The TS estimation results are presented in this subsection in a similar format as SS estimation.

(1) Unimodal models: In a manner similar to the presentation of the SS estimation results, Table 5.4 shows the results for TS estimation. Among the unimodal models, L-SVM trained on the BERT representation (L_c) achieves the best performance in the classification task (accuracy of 0.845, F1-score of 0.801), and FNN trained on the BERT representation (L_c) achieves the best performance in the regression task (MAE of 0.459). The model trained on fastText word vectors (L_b) shows the second best performance in classification task. The nonverbal unimodal models (rows 7 to 9 in Table 5.4) using physiological features (P), audio features (A) and visual features (V) all exhibit lower performance than the verbal models.

Table 5.4: TS estimation results for unimodal models. L_a, handcrafted linguistic features; L_b, fastText word vectors; L_c, BERT representations; P, physiological features; A, audio features; and V, visual features.

Feature	L-SVM						R-SVM						FNN					
	Acc		F1		MAE		Acc		F1		MAE		Acc		F1		MAE	
	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD
L_a	0.777	0.000	0.734	0.000	0.660	0.000	0.687	0.000	0.407	0.000	0.740	0.000	0.795	0.002	0.756	0.003	0.529	0.010
L_b	0.811	0.002	0.758	0.002	0.783	0.000	0.684	0.000	0.519	0.000	0.844	0.000	0.819	0.003	0.778	0.005	0.509	0.009
L_c	**0.845**	0.000	**0.801**	0.001	**0.555**	0.000	**0.742**	0.000	**0.600**	0.000	0.780	0.000	**0.836**	0.005	**0.798**	0.007	**0.459**	0.009
P	0.712	0.001	0.519	0.002	0.716	0.000	0.633	0.000	0.451	0.000	0.822	0.000	0.716	0.005	0.569	0.005	0.640	0.007
A	0.752	0.000	0.686	0.001	0.716	0.000	0.726	0.000	0.597	0.000	**0.730**	0.000	0.725	0.004	0.665	0.006	0.560	0.002
V	0.756	0.000	0.664	0.000	0.727	0.000	0.687	0.000	0.416	0.000	0.976	0.000	0.743	0.008	0.669	0.010	0.595	0.004

(2) Multimodal models: Similar to the SS estimation, the BERT representation (L_c) model shows higher performance than the model trained on handcrafted linguistic features (L_a) and fastText word vectors (L_b) in the TS estimation. Here, we report the BERT representation (L_c) model for further investigation of multimodal models (please see Table S3 in the Appendix for all of the fusion models with L_a or L_b). Among the multimodal models, the L-SVM model trained on L_c+P shows the highest performance in the classification task (accuracy of 0.846, F1-score of 0.803); however, the improvement was limited compared with the best unimodal model (accuracy of 0.845, F1-score of 0.801), as shown in Table 5.5. In contrast to the SS estimation, the TFN model trained on L_c+A+V shows the best performance in the regression task (MAE of 0.407) and outperforms the best unimodal model in TS estimation. TFN with BERT representations tended to have better performance than other architectures. FNN with EF also had better regression performance compared with the best unimodal model.

Table 5.5: TS estimation results for multimodal models.

Feature	L-SVM			R-SVM			FNN EF			FNN LF_1			FNN LF_2			TFN		
	Acc	F1	MAE	Acc	F1	MAE	Acc	F1	MAE	Acc	F1	MAE	Acc	F1	MAE	Acc	F1	MAE
L_c+P	**0.846**	**0.803**	**0.556**	**0.744**	0.588	0.771	**0.841**	**0.801**	0.433	**0.840**	**0.802**	0.466	**0.835**	**0.787**	0.511	**0.836**	**0.796**	0.422
L_c+A	0.818	0.768	0.636	0.726	0.596	0.634	0.815	0.768	0.447	0.806	0.755	0.460	0.801	0.747	**0.477**	0.829	0.788	0.414
L_c+V	0.838	0.793	0.676	0.701	0.573	0.683	**0.841**	**0.801**	0.428	0.832	0.792	0.453	0.828	0.780	0.492	0.832	0.792	0.418
P+A	0.752	0.685	0.719	0.712	0.584	0.734	0.745	0.686	0.545	0.726	0.667	0.559	0.734	0.662	0.571	0.761	0.703	0.525
P+V	0.756	0.666	0.790	0.690	0.422	0.985	0.772	0.691	0.564	0.745	0.673	0.584	0.759	0.665	0.593	0.772	0.682	0.558
A+V	0.756	0.694	0.713	0.727	0.604	0.690	0.760	0.701	0.528	0.737	0.676	0.550	0.755	0.691	0.548	0.773	0.713	0.513
L_c+P+A	0.816	0.768	0.637	0.730	0.593	0.628	0.814	0.767	0.451	0.805	0.757	0.456	0.801	0.738	0.507	0.829	0.788	0.415
L_c+P+V	0.838	0.793	0.618	0.732	0.619	0.736	0.840	0.800	**0.424**	0.828	0.787	**0.451**	0.816	0.755	0.521	0.832	0.791	0.411
L_c+A+V	0.817	0.768	0.630	0.741	**0.629**	**0.602**	0.821	0.774	0.439	0.807	0.758	0.456	0.807	0.753	0.493	0.829	0.788	**0.407**
P+A+V	0.760	0.700	0.723	0.725	0.606	0.687	0.752	0.693	0.526	0.733	0.674	0.546	0.765	0.689	0.561	0.779	0.714	0.514
L_c+P+A+V	0.814	0.765	0.645	0.736	0.622	0.606	0.816	0.766	0.444	0.808	0.761	0.454	0.807	0.747	0.514	0.831	0.790	0.409

5.6.3 Estimation with Automatic Speech Recognition

The results of the BERT representation models presented in Tables 5.2-5.5 based on text data manually transcribed by humans evaluate the potential of physiological signals in error-free text data. However, in an end-to-end automated system, error-free ASR is not realistic. Thus, we also evaluate the BERT representation model based on text data obtained by using a Japanese ASR system. Google speech API was used for ASR. The results are shown in Table 5.6. The models that achieved the best performance in each regression task (FNN LF_2 for SS estimation and TFN for TS estimation, respectively) were used for the experiment. In Table 5.6, "diff" means the difference in MAE between models based on text data from manual transcription and ASR. Larger "diff" values indicate performance degradation of models based on ASR. Unexpectedly, the performance of models based on ASR was similar to that of models using manual transcriptions in SS estimation, as shown in Table 5.6. This might indicate that the tokens that were difficult for the ASR system to recognize had only a little information for SS estimation and that SS estimation might be more dependent on nonverbal information such as physiological signals. In contrast, the performance of the ASR models in TS estimation was clearly lower than that of manual transcription (Table 5.6). This result suggests that the tokens that were difficult for the ASR system to recognize had important information for TS estimation and that TS estimation is more dependent on the linguistic modality than SS estimation. In addition, the multimodal model with physiological signals is also useful in SS estimation, and the multimodal model with audiovisual signals is useful in TS estimation, even under the ASR condition.

5.6.4 Analysis of Physiological Features

We conducted extensive feature analysis to clarify the contribution of each physiological feature. As depicted in Table 5.2, the physiological modality is the best for SS estimation among the four modalities. In addition, the L_c+P model with LF_2 achieves the best estimation performance among the multimodal models. Thus, the L_c+P model with LF_2 was used for feature analysis with backward-forward stepwise selection [96].

Table 5.7 shows the regression performance of the model removing the physiological features one by one in SS estimation. In step 1, the best MAE of 1.041 is derived from the result of SS estimation using the L_c+P model with LF_2 (Table 5.3). If the MAE increases after removing a feature from the model, the removed feature is effective for SS estimation. In contrast, if the MAE improves, the removed feature is not effective for SS estimation. "diff"

Table 5.6: Estimation results for models trained on ASR data.

Feature	SS estimation		TS estimation	
	FNN LF_2		TFN	
	MAE	diff	MAE	diff
L_c	1.111	0.000	0.490	0.032
L_c+P	**1.044**	0.003	0.464	0.042
L_c+A	1.095	0.001	0.449	0.035
L_c+V	1.086	0.001	0.453	0.035
L_c+P+A	1.051	0.001	0.454	0.040
L_c+P+V	1.051	0.002	0.448	0.037
L_c+A+V	1.083	0.001	0.444	0.037
L_c+P+A+V	1.054	0.001	**0.443**	0.033

(column 3 in Table 5.7) indicates the difference between the best MAE of the original model and the feature-removed model; thus, "diff" values with a negative sign (−) indicate that the removed feature is effective for SS estimation. We set the criteria in stepwise selection using "diff" values: If there is a feature that has the maximum and positive "diff" value, such a feature is eliminated and the stepwise selection continues to the next step. If there is no such feature, stepwise selection is finished. As shown in Table 5.7, "Mean of the first difference" had the maximum and positive "diff" value of +0.011 (with an asterisk (∗) in column 3 in Table 5.7); thus, this feature was eliminated, and the stepwise selection was continued. Finally, in step 4, since there is no feature that has a positive "diff" value, and simultaneously, re-entering the eliminated features does not improve performance, so stepwise selection is finished. Thus, the features used in step 4 could be useful for SS estimation. In particular, since "GSR number", "Kurtosis" and "Standard deviation" have negative "diff" values (< -0.01) in all steps, these features could be more effective than other features.

Table 5.7: Analysis of physiological features with the backward-forward stepwise method in SS estimation. The L_c+P model with LF_2 in Table 5.2 is used for evaluation. * indicates the maximum and positive "diff" value in each step.

	Step 1		Step 2		Step 3		Step 4	
Best MAE in one step before	1.041		1.030		1.028		1.019	
Removed feature	MAE	diff	MAE	diff	MAE	diff	MAE	diff
Mean	1.042	-0.001	1.032	-0.002	1.038	-0.010	1.030	-0.010
Standard deviation	1.054	-0.013	1.049	-0.019	1.049	-0.020	1.044	-0.025
Skewness	1.044	-0.003	1.028	+0.002*				
Kurtosis	1.067	-0.026	1.049	-0.018	1.045	-0.017	1.035	-0.016
Maximum value	1.046	-0.005	1.038	-0.007	1.019	+0.009*		
Minimum value	1.032	+0.009	1.030	+0.001	1.040	-0.011	1.035	-0.016
Mean of the first difference	1.030	+0.011*						
Mean of the second difference	1.039	+0.002	1.037	-0.007	1.042	-0.014	1.027	-0.007
GSR number	1.053	-0.012	1.049	-0.019	1.043	-0.015	1.035	-0.016
Range	1.045	-0.004	1.044	-0.013	1.044	-0.015	1.047	-0.028
Slope	1.042	-0.001	1.042	-0.012	1.065	-0.036	1.053	-0.034
Intercept	1.047	-0.006	1.044	-0.013	1.048	-0.020	1.043	-0.024
25th percentile value	1.046	-0.005	1.040	-0.010	1.036	-0.008	1.042	-0.022
75th percentile value	1.037	+0.004	1.043	-0.013	1.037	-0.009	1.027	-0.008

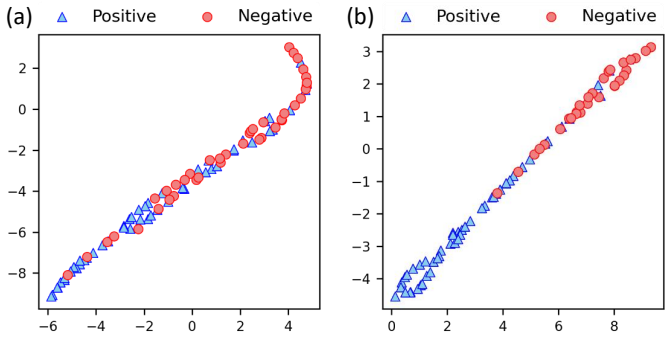

Figure 5.4: t-SNE visualization with test samples. (a) L_c model. (b) physiological model.

5.6.5 Comparison of Linguistic and Physiological Models

Figure 5.4 shows the t-distributed stochastic neighbor embedding (t-SNE) visualization [150] of the example of the learned features (hidden values from FNN with BERT representations and from FNN with physiological features) along with the positive sentiment label and negative sentiment label. This example shows that both models work well in binary classification in this case. However, these models might differ in the classification pattern.

We focus on the differences in the estimation results between the BERT representation (L_c) model and the physiological (P) model to investigate whether each model captures different aspects of sentiment changes. In fact, these two modalities clearly capture different aspects of human behavior (i.e., linguistic content and electrical changes in the skin) in human-agent interactions. In addition, differences between linguistic representations and physiological features in neural network models are rarely investigated in the field of affective computing. Thus, we were interested in comparing the estimation patterns of these models.

Figure 5.5 shows the confusion matrix for classification resulting from the unimodal L_c model and the physiological model. Many samples (65%) are classified as having negative sentiment labels by the L_c model. On the other hand, the samples are almost equally classified as positive and negative by the physiological model. Therefore, the models based on BERT representations and physiological features differ in their classification patterns.

To investigate the differences between these models for SS estimation in more detail, we compared the estimation results for each participant (Figure 5.6). Figure 5.6(a) shows the top five participants with the greatest differences in accuracy between the physiological model and the L_b model

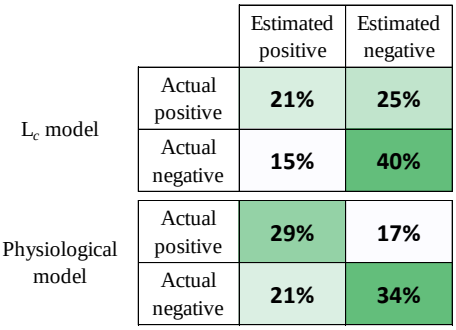

Figure 5.5: Confusion matrix for classification showing percentages of the total samples.

(gray bars in Figure 5.6(a)). The ratio between the actual positive (blue bar) and negative (red bar) sentiment labels is also shown for each participant. If the accuracy difference is greater than 0, then the physiological model classifies the sentiment labels more correctly than the L_c model does for that participant. In contrast, if the accuracy difference is less than 0, then the L_c model classifies the sentiment labels more correctly than the physiological model does, as in the cases shown in Figure 5.6(b). Figure 5.6(c) and (d) show the corresponding plots for the MAE differences. The physiological model exhibits higher estimation performance than the L_c model for participants with positive sentiment, as shown in Figure 5.6(a) and (c). In contrast, the L_c model seems to offer higher estimation performance than the physiological model for participants with negative sentiment, as shown in Figure 5.6(d). These results suggest that there are fundamental differences between SS estimation models based on language and physiology in human-agent interactions. These results are discussed in the following section.

5.7 Discussion

In this study, we present multimodal sentiment analysis considering four modalities, based on linguistic, audio, visual, and physiological features, at the exchange level in naturalistic human-agent interactions. The dataset used in our study, Hazumi1911, includes sentiment labels for each utterance annotated both by the participants themselves and by a third party, thus enabling us to analyze not only one-sided but also two-sided sentiments from the perspectives of SS and TS estimation, respectively, on the same dataset. The experiments reveal that physiological features, which are based on signals that are generally unobservable by humans, are the most effective features for SS

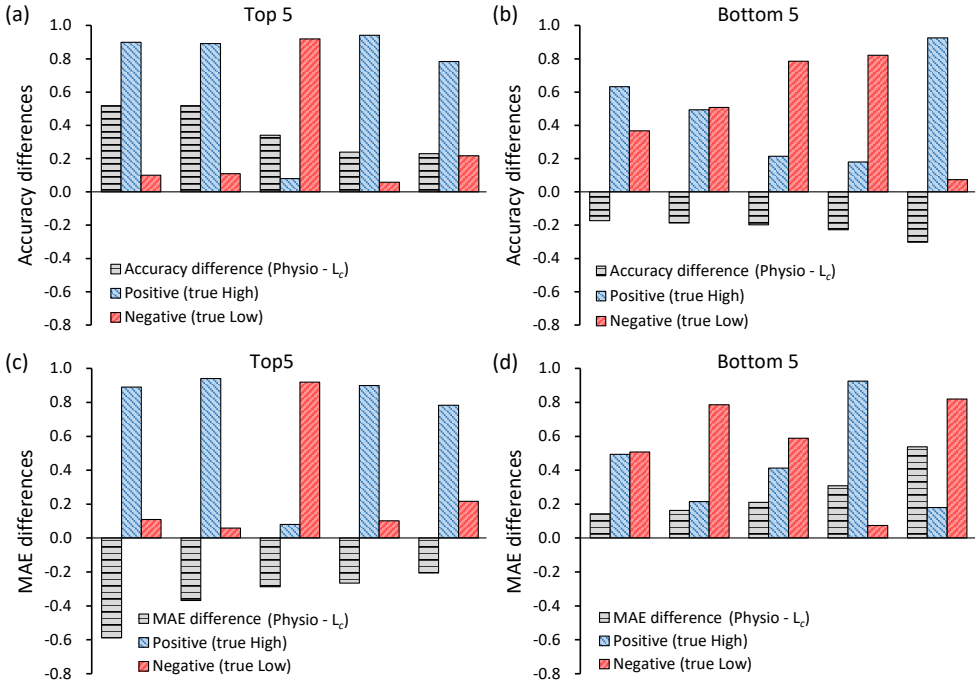

Figure 5.6: The differences in the estimation results between the BERT representation (L_c) model and the physiological (P) model for SS estimation. The top and bottom five participants with the greatest differences in accuracy (a and b) or MAE (c and d) between the models are depicted (gray bars). For each participant, the ratio between the actual positive (blue bar) and negative (red bar) sentiment labels is also shown for comparison.

estimation among the four modalities and that the estimation performance is improved by fusing physiological features and linguistic representations from BERT (Table 5.3). The Wilcoxon signed-rank test was performed to compare each evaluation value (accuracy, F1-score or MAE) between the best unimodal model and the best multimodal model in SS estimation. The best multimodal model outperformed the unimodal one (statistically significant for each evaluation value, $p < 0.05$). The performance of the resulting fusion model (L_c+P) was comparable to human performance (accuracy of 0.630) and superior to the best previously reported model (P+V) [138] (also shown in Chapter 4). In contrast, models based on verbal information showed higher performance than models based on nonverbal information (audio, visual or physiological features) on the TS estimation task (Table 5.4). Unlike SS estimation, a TS estimation model fusing linguistic representations from BERT

and audiovisual features achieved the best estimation performance in the regression task, presumably because these two types of information are easily perceptible (observable) by humans and also complement each other. Similar to the SS estimation, this multimodal model significantly outperformed the best unimodal model (vs L_c unimodal model, $p < 0.05$). We also report the effectiveness of each physiological feature by using stepwise selection (Table 5.7) and differences in the contributions of linguistic and physiological features for SS estimation (Figure 5.5 and 5.6). In this discussion section, a detailed interpretation of our results is presented.

We focus on physiological signals based on the ANS, which regulates the involuntary functions of the human body, to resolve the limitation of multimodal sentiment analysis. Since the physiological features capture emotional arousal, we expected physiological features to be useful for detecting enjoyment, which is related to emotional arousal. The unimodal physiological model achieved the best accuracy and MAE in SS estimation, and the multimodal fusion of linguistic and physiological features improved SS estimation performance, resulting in a model that achieved an accuracy comparable to that of humans (accuracy of 0.630). This implies that physiological responses can capture different aspects of sentiment changes that cannot be captured by BERT representations. Indeed, differences in estimation pattern were observed between the BERT representation model and the physiological model, as shown in Figure 5.5 and 5.6. It seems that physiological signals are superior in capturing positive sentiment that may not be expressed explicitly in textual information. These differences between modalities might complement each other in a corresponding multimodal neural network model, resulting in higher model performance.

Another aspect of our results that needs to be discussed is the differences between SS and TS estimation. It is clear that SS estimation is a more difficult task than TS estimation. Many previous works use TS in multimodal sentiment analysis [8, 9, 10], and SS and TS estimation have rarely been considered simultaneously on the same multimodal dataset. Considering the ultimate goal of sentiment analysis and dialogue system development, the SS estimation task at the exchange level and the development of a model that can capture the true SS adaptively require more extensive work, even if the findings show lower performance than TS estimation.

A previous study D'Mello et al. [151] assessed the contribution of each modality to the agreement between a user and a third party (observer) in naturalistic human-agent interactions. Their results suggested that human perception of others' affective states partially depends on multimodal observables. In line with this, although the dialogue situation in our study is different from that in this previous human study, the machine-based TS

estimation performance was improved by fusing BERT representations and audiovisual features, especially for the regression task (Table 5.5). Taken together with the SS estimation results, these results indicate that the role of each modality in machine learning differs between SS and TS estimation, and these differences appear to depend on whether the corresponding signals are observable by humans. These points need to be considered when designing sentiment studies and interpreting each modality, including physiological signals.

Although the fusion of linguistic and physiological information improves sentiment estimation performance, the benefits of the fusion of linguistic features and audiovisual features are limited (Table 5.3). From this, it can be inferred that there may be many redundancies between these modalities, as BERT is a SOTA language representation model with unparalleled performance on numerous tasks in the NLP domain. Moreover, unlike the BERT model, our audiovisual models do not have the benefit of large-scale pretraining; thus, there is a need to develop another audiovisual modeling method or algorithm as a basis for multimodal fusion.

In addition, although the performance of our proposed fusion model (L_c+P) is comparable to that of humans in SS estimation, further performance improvements will be needed to create emotionally intelligent agents with beyond-human performance to adapt dialogue strategies to provide a high-quality user experience. One possibility is to develop a nonverbal model with large-scale pretraining, similar to BERT [131]. However, compared with verbal information, nonverbal data such as physiological signals are more difficult to collect in a large-scale and noise-free manner, and appropriate labels are also needed. To this end, extensive data collection methods, dataset merging or transfer learning need to be considered.

We normalized each feature over all samples observed from a participant for test data during preprocessing; however, an input sample should be preprocessed and normalized sequentially in the inference (testing) phase in the online continuous recognition. That is, normalizing the validation/test data with the mean and standard deviation which are calculated from training dataset (without using statistics from the test data) is more appropriate, and this is one of the limitations of our study.

The feature extraction method in this study was based on statistics or average pooling at the exchange level. Considering time-dependent changes in each modality, incorporation and representation of temporal context by using multimodal LSTMs, GRUs, or Transformer [130] are intriguing strategies for improving sentiment estimation performance. In addition, the sentiment labels used in this study were assigned and estimated for each exchange independently, i.e., discrete-time levels. Thus, a continuous emotion/sentiment

recognition task with multimodal signals including physiological signals will also be needed.

There is a also need to evaluate the usefulness of other sensors, especially contact-less techniques that cause neither disturbance nor discomfort during dialogue. One candidate is video-based remote measurements of heart and respiratory rates [152]. Combining these techniques with multimodal analysis is an attractive approach to improve performance without physical contact and expand the applicability of such paradigms.

Since the previous datasets, Hazumi1712 and Hazumi1902, include text (L) and audiovisual signals (A and V) during dialogue but not physiological signals (P), and the present study is focused on the physiological signals, direct comparison of sentiment estimation performance is difficult between previous datasets and the present study (Hazumi1911). However, we think physiological signals are effective to complement information of L, A, and V features for estimating SS even in other datasets. Alternatively, a relevant line of work that includes physiological signals during human-agent interaction will enable us to compare between multimodal dialogue corpora with physiological signals directly.

5.8 Chapter Summary

This study demonstrated both SS and TS estimation using multimodal neural networks and revealed the effects of each modality, particularly the physiological modality, in naturalistic human-agent interactions. The experimental results suggest that the combination of BERT representations and physiological features is effective for SS estimation. This result appears to be attributable to the complementarity of the models based on BERT representations and physiological features, since these models capture different aspects of a participant's sentiment. In contrast, combining BERT representations and audiovisual features is effective for TS estimation, as textual and audiovisual features are both signals that are perceptible by humans. These results can advance the understanding of naturalistic sentiment; nevertheless, further research is needed to realize an emotionally intelligent agent with beyond-human capabilities.

Chapter 6

Multimodal Transformer with Physiological Signals

6.1 Introduction

The development of an adaptive dialogue system that can recognize a user's state in real time is necessary to ensure enjoyable conversations in human-agent interactions. During a chat dialogue, the system should behave according to the real-time state of the user. For example, if a user is bored with the current topic, the system should explore other topics, similar to human behavior. However, there are several reasons why this task is challenging. For example, SS cannot necessarily be expressed with the linguistic information obtained from user utterances. Users may mask their sentiment in their mind and not express their true sentiment as an utterance or behavior due to their emotional intelligence [153].

As above-mentioned and presented in Chapter 3 to 5, peripheral physiological signals have been investigated in psychophysiology and affective computing. These signals can potentially reflect emotional changes by capturing physiological changes in the ANS. For example, a faster phasic component in the EDA, which is derived from the activity of the sweat glands, can be used to detect emotional arousal [140, 27]. Since the ANS is involuntary, i.e., it cannot be controlled consciously, physiological changes during dialogue are difficult to mask. Therefore, physiological signals may be suitable for capturing SS changes that cannot be represented by linguistic information in user utterances and can function as complementary information.

However, investigations into the effectiveness of time-series physiological signals for estimating SS during dialogue exchanges have been limited. Most studies on the use of physiological signals to estimate emotion/sentiment

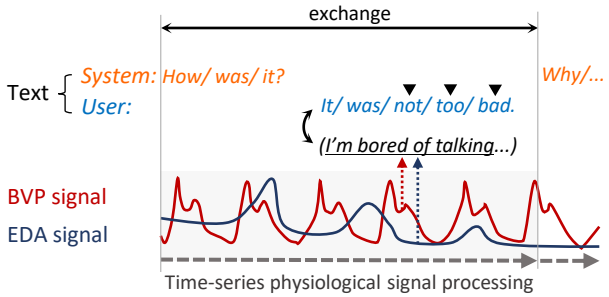

Figure 6.1: Example of capturing SS changes by using linguistic information and physiological signals at the exchange level. The user token sequences "not/too/bad" include both neutral and positive sentiments. However, the true sentiment in his/her mind is "bored" (tentative example). This masked negative sentiment is accompanied by reduced arousal levels and would be captured by time-series physiological signals.

have induced emotional stress with visual stimuli over a relatively long time period (several minutes). Thus, there is a need to investigate whether signals detected in shorter time periods (approximately 10 seconds) are effective for online emotion/sentiment estimation.

In addition, although it is assumed that short-time physiological signals can complement spoken linguistic information in SS detection, there have been no studies that show an effective method for combining time-series physiological signals with token sequences represented by SOTA language models, such as the BERT model [131]. Thus, the exploration of effective methods for fusing physiological signals and token representations is valuable for developing adaptive dialogue systems.

The aforementioned issues and the approach presented in this study are summarized in Figure 6.1. An "exchange" is defined as a segment that begins at the start of a system utterance and ends at the start of the next system utterance. In this case, a model based solely on user token sequences is insufficient for estimating SS. We expect that time-series physiological signals could be used to capture SS changes that are not expressed in linguistic information, and a time-series model that combines physiological signals and language representations can improve the sentiment estimation performance, as these data are complementary.

In this study, we propose an effective method for processing physiological signals and combine this method with a language model. We focus on linguistic information and physiological signals since the models based on BERT representations or physiological signal had dominant performance compared

to audiovisual models (described in Section 6.5.1). The contributions of our work are as follows:

- We propose a time-series physiological signal processing method for exchange-level sentiment estimation. The models based on the time-series data of the EDA phasic component capture short-time sentiment changes during exchanges, showing competitive performance to a linguistic model based on SOTA computational representations, i.e., BERT representations (Section 6.5.1).

- We introduce the Time-series Physiological Transformer (TPTr), which combines time-series physiological signals with BERT representations to capture short-time sentiment changes based on both textual aspects and physiological changes in the user (Section 6.5.2). As a result, our proposed ensemble model outperforms the previously reported best result.

- Our proposed model is extended and validated by using a variety of physiological signals, including the BVP. The performance is further improved with the ensemble method, as shown in Section 6.5.3.

6.2 Related Works

This section specifically focuses on research related to the Transformer language model and multimodal models.

Text-based approaches are critical in sentiment analysis, and neural network models such as LSTM are widely used [126]. However, the Transformer model, which was developed by [130], has become the de facto standard and most commonly used language model. The best Transformer-based model is BERT [131], which achieved numerous successes with sentiment estimation tasks with datasets such as the SST-2 [128]. When BERT is pretrained with a large-scale dataset, representations can be extracted from text data (referred to as BERT representations), and BERT representations can be used as input feature vectors in other architectures. This method allows BERT representations to be easily combined with audiovisual features and is often used in multimodal sentiment analysis.

Although several Transformer-based multimodal models for affective computing and sentiment analysis have recently been proposed [154, 155, 156], a Multimodal Transformer called MulT was the first model proposed in multimodal sentiment analysis research [157]. Language, video and audio modalities, as well as sentiment labels annotated by third parties, were used to

demonstrate the effectiveness of the proposed crossmodal attention model, which latently adapts streams from one modality to another. Although physiological signals were not included in these studies, it has been suggested that Transformer-based models could capture crossmodal attention between text and audiovisual signals. Multimodal Adaptation Gate (MAG) was introduced in [155] and is applied to the Transformer architecture of BERT/XLNet. The CMTr allows to shift the language-only position (representation) of the word to the new position by injecting audio-visual information. The core component of the CMTr is a non-verbal displacement vector derived from the audio and visual vectors with their respective gating vectors. Hazarika et al. [156] proposed modality-invariant and -specific representations, which project language, audio and visual modalities to two distinct subspaces. The respective representations are stacked into a matrix, and Transformer is used to perform a multi-headed self-attention on the matrix.

Compared to linguistic and audiovisual modalities, there are very few publicly available physiological signal datasets for emotion/sentiment research. However, several datasets that include physiological signals have been created while viewing emotional videos [64, 65, 56] or conversations [44, 46]. In [46], a multimodal human-agent dialogue corpus that included linguistic, audiovisual, and physiological information was created. The participants interacted with an agent, and sentiment labels were retrospectively annotated for each exchange by both the participants and a third party. The collected nonverbal signals (audio, visual, and physiological signals) in this dataset were used for sentiment estimation with support vector machine or FNN models, and the results showed that physiological signals, particularly features based on SC signals, were useful for exchange-level sentiment estimation, as presented in Chapter 4. Chapter 4 was extended by Chapter 5), which presented a comprehensive analysis of the effectiveness of physiological signals in multimodal sentiment analysis. Since our proposal in this chapter is an effective hybrid algorithm that combines physiological features and the Transformer language model, this chapter differs considerably from Chapter 4 and 5, which used conventional neural networks.

To the best of our knowledge, there is no publicly available dataset that includes textual and physiological information during dialogue exchanges, except for [46]. As mentioned above, text-based approaches are the most common sentiment analysis methods, and multimodal language models using Transformer and BERT have been proposed. Physiological signals are promising candidates for capturing subtle sentiment changes that cannot be detected in the speaker's explicit information, i.e., text and audiovisual information. Nevertheless, an effective method that combines a SOTA language

model and physiological signals has not yet been developed, most likely because of dataset limitations.

We propose the use of physiological signals with a SOTA language model to estimate sentiment during human-agent interactions. The Hazumi1911 dataset [46], which is the only publicly available dataset that includes time-series textual and physiological information, enables us to evaluate the effectiveness of the combination of physiological signals and text. We propose a time-series physiological signal processing method that effectively combines physiological signals and token sequences of utterances. We show that our proposed method is useful for exchange-level sentiment estimation, and our results are comparable to those of a model based on BERT representations. Then, we show how the time-series physiological signals can be incorporated into a SOTA language model, and proposed model were compared with the previously reported best performing model.

6.3 Proposed Methods

This section presents our proposed methods for incorporating time-series physiological signals at the exchange level. In this study, the physiological signals included the EDA, BVP, HR and TEMP. The EDA is a measure of the electrical activity in human skin and reflects sweat gland activity. The BVP is based on spectral analyses of the skin (blood vessels) and reflects physiological changes in cardiovascular activity. In this study, the raw EDA signal (SC, denoted as $\mathrm{EDA_{SC}}$) was decomposed into a fast phasic component ($\mathrm{EDA_{fast}}$) and a tonic component ($\mathrm{EDA_{tonic}}$) with the same method as in Chapter 4. In Section 6.3.1, we describe a physiological signal processing method for calculating fine-segmented physiological changes. Since each physiological signal has a different sampling rate, a simple segmentation and averaging method was applied. In Section 6.3.2, to evaluate the effectiveness of the processed data, time-series machine learning models are introduced. Specifically, we propose a TPTr model in which the encoder is based on attention weights from the token representations and corresponding physiological signals. We expect this encoder to capture sentiment changes by using both linguistic and physiological information, as sentiment changes cannot be detected with only linguistic information.

6.3.1 Time-Series Physiological Signal Processing

To roughly align physiological signals within the exchanges with the token, a unit of language models, we divide each physiological signal during each

exchange by the number of tokens. Let one exchange duration be s, the sampling rate of the physiological signal in Hz be h, and the number of tokens in one exchange be n. The number of samples per token m is determined by rounding $\frac{sh}{n}$ down to the nearest integer. Then, from the start of the exchange, the raw sampling data per m are averaged in order (i.e., m is the variable window size). Thus, the physiological signal $\boldsymbol{p}$ in the ith exchange is denoted as an n-dimensional vector:

$$\boldsymbol{p}_i^{\alpha} = (p_{i1}^{\alpha}, \ldots, p_{in}^{\alpha})^T \tag{6.1}$$

where α indicates the physiological submodality such as EDA_{fast}, $\text{EDA}_{\text{tonic}}$, EDA_{SC}, BVP, HR, TEMP.

We note that our proposed preprocessing method is not the strict word-level alignment method that has been proposed in prior works [158, 159]. In contrast to acoustic signals, physiological signals do not necessarily have a significant co-occurrence property with the uttered words because the physiological changes may relate to words spoken in the past or future. Thus, physiological signals are not simply weighted with a specific token in this study. Rather, the aim is to extract representations from fine segments of physiological signals with token sequences, which could shift the original representations at the *exchange level*. More details and examples of our experiment are shown in Section 6.5.4.

6.3.2 Time-Series Modeling of Physiological Signals

(1) Physiological LSTMs: The LSTM and BiLSTM models are applied to validate whether our proposed time-series preprocessing method performs comparably to models based on BERT representations, which have deep bidirectionality [131]. An LSTM [160] model based on physiological signals $\boldsymbol{p}_i$ at time t can be represented as

$$\begin{aligned} \begin{pmatrix} \boldsymbol{f}_t \\ \boldsymbol{g}_t \\ \boldsymbol{\iota}_t \\ \boldsymbol{o}_t \end{pmatrix} &= \begin{pmatrix} \sigma \\ \tanh \\ \sigma \\ \sigma \end{pmatrix} W \begin{pmatrix} \boldsymbol{p}_t \\ \boldsymbol{h}_{t-1} \end{pmatrix} \\ \boldsymbol{c}_t &= \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{g}_t \odot \boldsymbol{\iota}_t \\ \boldsymbol{h}_t &= \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t) \end{aligned} \tag{6.2}$$

where $\boldsymbol{f}_t, \boldsymbol{\iota}_t$ and $\boldsymbol{o}_t$ are the forget, input, and output gates, respectively; σ is the sigmoid function; W is the weighting parameter; $\boldsymbol{c}_t$ is the memory cell; $\boldsymbol{h}_t$ is the hidden state; and $\odot$ is the Hadamard product. Note that the time

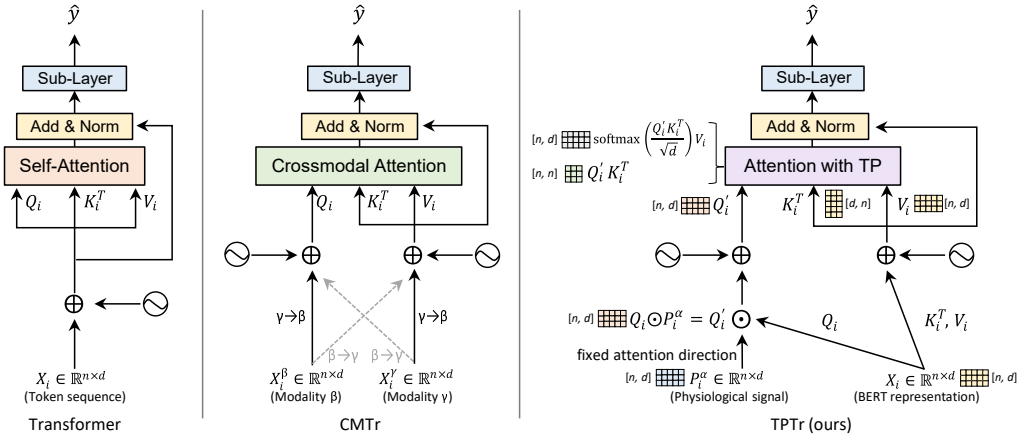

Figure 6.2: Conventional Transformer [130] (left) architecture, CrossModal Transformer [157] architecture based on two modalities, β and γ (CMTr, center), and our proposed Time-series Physiological Transformer (TPTr, right) architecture. In our proposed model (right), exchange-level physiological signals and BERT representations derived from user and system utterances are combined by applying the Transformer encoder. This model allows physiological information to be continuously linked to linguistic information (performing attention with time-series physiological signal processing) and can capture physiological aspects that cannot be detected with linguistic information alone. The number in the bracket indicates the dimension of the corresponding matrix. For a detailed description of the Transformer and CMTr architectures, please see Section 6.4.1.

t corresponds to the number of tokens n, as described in Section 6.3.1. $\boldsymbol{p}_t$ is denoted as a vector in the above equation; however, this variable corresponds to a scalar when the selected physiological submodality is single.

After the preprocessing methods described in Section 6.3.1 were carried out, the raw physiological data of each participant were normalized by Z score normalization. In other words, we normalized each feature over all the samples collected from a participant in the training or testing data during preprocessing. Following this, zero padding was performed since the token length of each exchange differs. Then, the result was fed into the input layer of the LSTM model. The final LSTM block outputs $\boldsymbol{h}_t$ are connected to the final output layer in a mode known as many-to-one, and finally, the estimated values are obtained.

(2) Time-Series Physiological Transformer: After the effectiveness of the physiological LSTM models were confirmed (as described in Sec-

tion 6.5.1), we extended our proposed method to fuse time-series physiological signals with SOTA language representations, i.e., BERT representations, by using the Transformer encoder [130]. A summary of the proposed TPTr architecture is shown in Figure 6.2. Exchange-level BERT representations are extracted with a pretrained BERT model[1], which is represented as a matrix of dimension $\mathbb{R}^{n\times d}$, where d is hidden size of the BERT representations and n is number of tokens. The time-series physiological signal $\boldsymbol{p}_i^{\alpha}$ was turned into $P_i^{\alpha} \in \mathbb{R}^{n\times d}$ by repeating array along the axis to match the dimension of the BERT representations. To incorporate the time-series physiological signals into the linguistic information, a dot-product attention mechanism [130] was applied. The dot-product attention mechanism is composed of a query Q, a key K, and a value V. We consider the BERT representation in the ith exchange as $Q_i = K_i = V_i$, where $Q_i \in \mathbb{R}^{n\times d}$. The dot product between Q_i and K_i^T is computed as the similarity to calculate the attention weight. To combine the time-series physiological signals with the BERT representations, we use the Hadamard product between P_i^{α} and Q_i, denoted as Q_i'. We hypothesize that this modification may shift the attention weight and could provide representations that differ from conventional BERT representations. The output of the dot-product attention operation is:

$$\text{Attention}(P_i^{\alpha}, Q_i, K_i, V_i) = \text{softmax}\left(\frac{Q_i' K_i^T}{\sqrt{d}}\right) V_i \tag{6.3}$$

where the scaling factor $\sqrt{d}$ is used. The PEs are sinusoidal and identical to the modules proposed in [130]:

$$PE_{(pos,2j)} = sin(pos/10000^{2j/d}) \tag{6.4}$$

$$PE_{(pos,2j+1)} = cos(pos/10000^{2j/d}) \tag{6.5}$$

where pos is the position of the token and j is the dimension of the hidden layer. The PEs are added to Q_i', K_i^T and V_i to carry information about the position of the tokens. The weighting parameters $W^{Q'} \in \mathbb{R}^{n\times d}, W^K \in \mathbb{R}^{n\times d}$, and $W^V \in \mathbb{R}^{n\times d}$ are also implemented.

Similar to [130], the Transformer encoder is composed of two sublayers. The first sublayer is the aforementioned dot-product attention mechanism, and the second sublayer is a fully connected FNN. Each sublayer has a skipping connection [161] and layer normalization [162], denoted as "Add" and "Norm" in Figure 6.2, respectively.

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

In summary, our proposed preprocessing method converts data, allowing the model to combine physiological signals with BERT representations, which are both represented as matrices during each exchange. These representations are fed into the Transformer model, with the token positions providing the attention weights, thus allowing physiological changes to be considered during exchanges.

6.4 Experimental Settings

This section describes the experimental settings for the evaluation of our proposed model. One of the strengths of our proposed method is that our method applies short-time episodes (approximately 10 seconds), which enables dialogue systems to adaptively respond to sentiment changes in the user in a timely manner. Only one publicly available dataset includes both the time-series physiological signals and linguistic information of the user at the exchange level: the Hazumi1911 dataset [46]. We use this dataset to evaluate our proposed methods, and Section 6.4.3 summarizes the dataset. Section 6.4.1 describes the models used as baselines for comparison, and the evaluation procedure is described in Section 6.4.2.

6.4.1 Baselines and Hyperparameters

As described in Section 6.3, the proposed time-series physiological signal processing method was evaluated by using the LSTM, BiLSTM, or TPTr models as inputs. This subsection describes the baseline models that were used for comparisons with our proposed method.

(1) Feedforward Neural Network (FNN): The FNN architecture was used as one of our baselines. The FNN was composed of an input layer, four fully connected layers with dropout in each layer, and an output layer. The FNN has two lower intermediate layers with 64 units and two higher intermediate layers with 32 units. The dropout rate was set to 0.3. The ReLU function was used as the activation function.

(2) Long Short-Term Memory Models (LSTMs): In the LSTM model, the number of LSTM blocks was set to 3, with 64 hidden units (in the BiLSTM model, the number of hidden units was set to 128 in total). No dropout was applied. The activation functions (sigmoid and hyperbolic tangent) are described in Section 6.3.2.

(3) Transformer (Tr): A conventional Transformer encoder [130] was used as a baseline. This model used only linguistic information (i.e., BERT representations) for sentiment estimation. As shown in Figure 6.2 (left), the

Transformer encoder was composed of two sublayers. The first sublayer was a self-attention mechanism, and the second sublayer was an FNN. Each sublayer had a skipping connection and layer normalization. The number of Transformer encoder blocks and attention heads is 1. The dimensionality of the input and output is 768, corresponding to the BERT model. The number of units in the pointwise FNN is 128. The dropout rate was set to 0.3. The size of the trainable parameters is 2.6M. The Tr×3 model has three identical parallelized Transformer blocks. $\text{FNN}_{\text{L+P}}$ (described in Section 6.4.3) was used to combine with the Transformer models. CrossModal Transformer (CMTr) and our proposed TPTr(×3) are described below.

(4) CrossModal Transformer (CMTr): The BERT is a core component of the MulT and was proposed in [157]. The MulT model captures multimodal signals according to crossmodal attention and achieves SOTA results in multimodal sentiment estimation. The CMTr model applied crossmodal attention with linguistic, audio, or video modalities, as reported in [157]. The two modalities β and γ, as denoted in Figure 6.2 (center), correspond to linguistic, audio, or video modalities. The transfer of information from modality γ to modality β is denoted as "$\gamma \rightarrow \beta$" in Figure 6.2 (center). The CMTr model also includes reverse attention, which is denoted as "$\beta \rightarrow \gamma$", in which information is assigned to another Transformer block, allowing modality γ to receive information from modality β. Thus, the attention direction is variable. On the other hand, our proposed TPTr model applies attention with linguistic and physiological modalities and has a fixed attention direction. Therefore, the CMTr and TPTr models use different modalities, and the attention mechanism also differs. For a fair comparison, we fuse BERT representations and physiological signals when using the CMTr architecture in this study. The CMTr model has two Transformer encoder blocks that pass information as $\gamma \rightarrow \beta$ and $\beta \rightarrow \gamma$. The output of each CMTr block was concatenated (64 units in total) and connected to the final output layer. The other parameter settings of the CMTr and Transformer models are identical.

(5) Time-Series Physiological Transformer (TPTr): The TPTr and Transformer models have the same parameter settings. The TPTr×3 model has three extended parallelized Transformer blocks. The output of each TPTr×3 block was concatenated (96 units in total) and connected to the final output layer. Other than these settings, we use the same parameter settings in the Tr(×3), CMTr, and TPTr(×3) models to facilitate a fair comparison.

For late fusion models, each higher intermediate layer in the model is concatenated and connected to the output layer. For ensemble models, the output values of each model were averaged and used as the final estimated value. Late fusion and ensemble methods are both widely used in multimodal

machine learning [148]. In consideration of the computational cost, the maximum token length was set as 64 in this study. The other hyperparameters were set as follows: a learning rate of 0.001 with the Adam optimizer and a batch size of 32. The FNN model and models other than the FNN model were trained with 30 and 3 epochs, respectively. Mean squared error was used as a loss function in all experiments. All models were implemented in Keras with TensorFlow backend on NVIDIA GeForce RTX 2060.

6.4.2 Evaluation Procedure

A LOUOCV method was used in our evaluation. In the LOUOCV method, the samples corresponding to each exchange between a participant and the dialogue system were used as the test data, and the remaining samples of the other twenty-five participants were used as the training data. This procedure ensured that the test data of one participant were completely excluded from the training dataset, thereby preventing leakage and overestimation. The MAE and Pearson correlation coefficient (Corr) were calculated for each evaluation. The average MAE and Corr values with the LOUOCV method are reported. All experiments were performed three times with random initializations, and the evaluation values were calculated as the average value across the three repetitions. These evaluation values were then compared among the models.

6.4.3 Dataset

The Hazumi1911 dataset [46], a multimodal human-agent dialogue corpus, is also used in this chapter. In brief, the data were collected while participants chatted with an agent that operated using the Wizard of Oz method. Data from 26 of the participants and 2468 total exchanges were used in our experiment, and the data are denoted in the same manner as in Chapter 4. The participants annotated the labels for each exchange while watching videos of themselves after the experiment. The labels were assigned as sentiment scores ranging from 1 (no enjoyment of the dialogue) to 7 (enjoyment of the dialogue) and used in regression tasks.

In the Hazumi1911 dataset, the participants' utterances were manually transcribed into text data. The language representations were extracted by BERT, as described in Section 6.3. In addition, physiological signals were recorded using an Empatica E4 wristband (Empatica Inc., Cambridge, MA, USA) developed by Empatica Inc. The E4 device is worn like a wristwatch; it causes neither disturbance nor discomfort during dialogue and has been widely used in affective computing research, such as in [163, 164, 165]. Thus,

this device is suitable for the evaluation of our proposed methods. The EDA, BVP, HR and TEMP data were recorded at 4, 64, 1 and 4 Hz, respectively. Each time-series physiological signal was preprocessed as described in Section 6.3.1. Following Chapter 4, statistics such as the mean, standard deviation and maximum values of the physiological signals were used for comparisons with baseline models.

Acoustic and visual features were also extracted in the same manner as described in Chapter 5. In brief, the INTERSPEECH 2009 Emotion Challenge feature set (IS09) [122] was extracted from participant's utterances as acoustic features using OpenSMILE software[2]. A total of 384 acoustic features were extracted. Based on the video data, facial landmarks near the eyes, mouth, and eyebrows were identified with the OpenFace library [123], and the velocity and acceleration at each point were calculated to use as facial features. Based on motion data of the hands, shoulders and head recorded with Microsoft Kinect sensors, the velocity and acceleration were calculated to use as motion features. In total, 86 visual features were extracted from the facial expressions and motion activity. These acoustic and visual features were used for model comparisons based on each modality. Models based on each feature are as follows:
(1) $\mathbf{FNN_L}$: FNN model based on BERT representations
(2) $\mathbf{FNN_P}$: FNN model based on $\mathrm{EDA_{fast}}$ statistics
(3) $\mathbf{FNN_A}$: FNN model based on acoustic features
(4) $\mathbf{FNN_V}$: FNN model based on visual features
(5) $\mathbf{FNN_{L+P}}$: FNN model based on BERT representations and $\mathrm{EDA_{fast}}$ statistics
(6) (Bi)$\mathbf{LSTM_P}$: (Bi)LSTM model based on time-series $\mathrm{EDA_{fast}}$ signals

6.5 Results and Discussion

First, we show the effectiveness of the models based on our proposed time-series physiological signal processing method. The physiological LSTM and BiLSTM models perform better than the conventional FNN model based on the statistics. Furthermore, ensembles with linguistic and physiological modalities further improve the estimation performance (Section 6.5.1). Second, a SOTA language model, namely, the Transformer model, was used to combine the time-series data derived from the physiological and linguistic information. This novel approach captures representations that depend on both token sequences and time-series physiological changes, resulting in further performance improvement with the ensemble model (Section 6.5.2).

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

Third, to explore other effective time-series physiological signals, the TPTr model based on various physiological signals was evaluated in our proposed framework, and its usefulness was demonstrated (Section 6.5.3). This analysis reveals that the time-series BVP signal is another useful physiological signal for sentiment estimation. Fourth, to clarify the effect of the physiological signals, the differences in the attention weights between the conventional Transformer and TPTr models was shown (Section 6.5.4). Finally, a qualitative example of the estimation pattern is shown in Section 6.5.5 to visualize sequential dynamic sentiment changes and the behavior of each model.

6.5.1 Performance of Physiological LSTM Models

Table 6.1 shows the regression performance of the unimodal FNN models (FNNs trained with BERT representations, EDA_{fast} statistics, acoustic features and visual features are depicted as FNN_L, FNN_P, FNN_A and FNN_V, respectively) using the model described in Chapter 5 as baseline (rows 2 to 5 in Table 6.1). Our proposed model, that is, the LSTM models trained on time-series physiological signals (LSTM_P and BiLSTM_P), are shown in rows 6 and 7 in Table 6.1. In the single model results, our proposed physiological LSTM models have higher Corr values than the conventional FNN_P (rows 3, 6 and 7 in Table 6.1). Noted that, FNN_P has a relatively low performance regarding Corr (0.091) similar to regarding F1 (0.499) shown in Table 5.2 in the binary classification task (the row 7 in Table 5.2 in Chapter 5). This problem is improved by our proposed LSTM_P (Corr changed from 0.091 to 0.179). Although the FNN_L model has the best Corr value of 0.254, the physiological models (FNN_P, LSTM_P and BiLSTM_P) have lower MAEs than FNN_L (1.086). The FNNs based on conventional acoustic and visual features (FNN_A and FNN_V) do not outperform FNN_L, LSTM_P or BiLSTM_P.

In terms of the MAE, further performance improvement was observed by combining the linguistic and physiological models (late fusion and ensemble models). The ensemble model $\text{FNN}_{\text{L+P}}$+LSTM_P achieved an MAE of 1.041 and a Corr of 0.250.

These results suggest that our proposed physiological signal processing method is effective for exchange-level sentiment estimation, even if linguistic modalities are not included (LSTM_P and BiLSTM_P). Compared to the experimental condition, which uses emotional stimuli, the estimation of SS in natural dialogue is a difficult task. Nevertheless, our proposed method achieved competitive performance with an FNN trained on BERT representations (FNN_L). Furthermore, our proposed multimodal models based on linguistic and physiological information efficiently complement each modal-

Table 6.1: Sentiment estimation results of physiological LSTM models based on EDA_{fast}. The experimental results based on the model reported in Chapter 5 [166] and the results of our proposed models (ours) are also depicted.

Model		MAE	Corr
Single model	FNN_L [166]	1.086	**0.254**
	FNN_P [166]	1.069	0.091
	FNN_A [166]	1.196	0.145
	FNN_V [166]	1.166	0.145
	$LSTM_P$ (ours)	1.067	0.179
	$BiLSTM_P$ (ours)	1.069	0.176
Late fusion model	FNN_{L+P} [166]	1.079	0.178
	FNN_{L+P}+$LSTM_P$ (ours)	1.062	0.184
	FNN_{L+P}+$BiLSTM_P$ (ours)	1.047	0.191
Ensemble model	FNN_{L+P} [166]	1.047	0.238
	FNN_{L+P}+$LSTM_P$ (ours)	**1.041**	0.250
	FNN_{L+P}+$BiLSTM_P$ (ours)	**1.041**	0.249

ity. These results indicate that our proposed physiological signal processing method can potentially capture sentiment changes that cannot be represented by BERT representations alone.

6.5.2 Performance of TPTr

Table 6.2 shows the regression performance of the conventional Transformer model, the CMTr model proposed in [157], and our proposed TPTr model. The single models and late fusion models did not outperform the abovementioned ensemble model FNN_{L+P}+$LSTM_P$ (Table 6.1). However, all of the ensemble models showed higher performance than the single models. In particular, ensemble model FNN_{L+P}+TPTr×3 achieved the best results, with an MAE of 1.033 and a Corr of 0.262. In a previous study that used the same dataset and machine learning task as we presented here, it was shown that the ensemble model FNN_{L+P} achieved a better performance than other multimodal models [166]. We show here that our proposed ensemble model (FNN_{L+P}+TPTr×3) significantly outperforms the previously reported best model (FNN_{L+P}, $p < 0.05$, Wilcoxon signed-rank test), suggesting the effectiveness of our proposed method. In addition, we observed significant performance improvement for the TPTr×3 model compared to the Tr×3 model

Table 6.2: Sentiment estimation results for the Transformer model and its variant. Tr, Transformer; CMTr, CrossModal Transformer [157]; TPTr, our proposed Time-series Physiological Transformer. "×3" means triplicated Transformer blocks.

Model	Single model		Late fusion with $\text{FNN}_{\text{L+P}}$		Ensemble with $\text{FNN}_{\text{L+P}}$	
	MAE	Corr	MAE	Corr	MAE	Corr
Tr	1.082	0.227	1.057	0.221	1.042	0.259
Tr×3	1.109	0.219	1.069	0.230	1.053	0.257
CMTr [157]	1.083	0.190	1.138	0.198	1.040	0.254
TPTr (ours)	1.114	0.228	1.099	0.223	1.051	0.261
TPTr×3 (ours)	**1.068**	**0.232**	**1.045**	**0.240**	**1.033**	**0.262**

by further experimental repetitions ($p < 0.05$, Wilcoxon signed-rank test).

These results indicate that incorporation of time-series physiological changes into the Transformer language model, which was achieved with our proposed TPTr model, can capture different representations that cannot be captured by using only FNN_{L} or FNN_{P} or the ensemble model $\text{FNN}_{\text{L+P}}$. As shown in Section 6.3, only the dot product of the query and key differs between the conventional Transformer model and our proposed TPTr model, and this difference can affect the TPTr estimation result. The details of the attention weight are analyzed and discussed in Section 6.5.4.

6.5.3 TPTr Based on Other Submodalities

We investigated whether the TPTr model based on other physiological signals and its ensembles were effective for sentiment estimation. We evaluate the following models:

(1) Single model: This model is our proposed TPTr×3 model, which was trained on each preprocessed signal from the physiological submodality α, as shown in Section 6.3. A total of five single models were constructed.

(2) Ensemble of 3 models: The ensemble was constructed using the $\text{FNN}_{\text{L+P}}$ (i.e., FNN_{L} and FNN_{P}), and TPTr×3 models trained on physiological submodalities.

(3) Ensemble of 4 models: The ensemble was constructed using $\text{FNN}_{\text{L+P}}$, and two models selected from Tr×3 or TPTr×3 trained on physiological submodalities. To compare the conventional Tr×3 model with our proposed TPTr×3 models, two sets of ensembles were evaluated: $\text{FNN}_{\text{L+P}}$, TPTr×3

Table 6.3: Sentiment estimation results of the TPTr model based on physiological submodality to explore other effective submodalities. EDA_{tonic}, tonic component of EDA; EDA_{SC}, skin conductance; BVP, blood volume pulse; HR, heart rate; TEMP, skin temperature.

Model	TPTr submodality	MAE	Corr
Single model	EDA_{tonic}	1.113	0.225
	EDA_{SC}	1.115	0.237
	BVP	1.080	0.258
	HR	1.112	0.232
	TEMP	1.100	0.221
Ensemble model (3 models)	EDA_{tonic}	1.052	0.261
	EDA_{SC}	1.052	0.264
	BVP	1.041	0.269
	HR	1.050	0.264
	TEMP	1.053	0.259
Ensemble model (4 models)	EDA_{fast}	1.041	0.268
	EDA_{fast} and BVP	**1.033**	**0.276**
Human	-	1.008	0.406

trained on EDA_{fast}, Tr×3; FNN_{L+P}, TPTr×3 trained on EDA_{fast}, TPTr×3 trained on BVP.

Table 6.3 presents the estimation results of the abovementioned models. Among the five single models based on each submodality, the TPTr model based on the BVP signal has the best result (row 4 in Table 6.3). The TPTr model based on the BVP signal also had the best result for the ensemble of 3 models, with an MAE of 1.041 and a Corr of 0.269 (row 9 in Table 6.3). Finally, we evaluated the ensemble of 4 models: FNN_{L+P}, TPTr×3 based on the EDA_{fast}, and TPTr×3 based on the BVP signal (the second row from the bottom in Table 6.3). This ensemble model achieves the best result in this study, with an MAE of 1.033 and a Corr of 0.276. The ensemble of 4 models including Tr×3 has a worse performance in terms of the MAE (1.041) than the model without Tr×3 (MAE of 1.033, the last row in Table 6.2).

EDA_{fast} is known to be related to emotional arousal, and we have presented its effectiveness (Tables 6.1 and 6.2); however, the BVP signal could also be useful for sentiment estimation with our proposed framework. Both the EDA and BVP signals are related to the ANS; however, the EDA signal reflects changes in sweat gland activity, while the BVP signal reflects

physiological changes in the cardiovascular system. Further improvement was achieved with the ensemble of 4 models by using the TPTr model based on the BVP signal; thus, different physiological submodalities may reflect different aspects of sentiment changes that cannot be explicitly represented by using linguistic information alone, resulting in the ensemble of 4 models achieving further performance improvement. On the other hand, other submodalities appeared to have little effect on the estimation performance. Thus, other time-series processing or feature extraction methods should be considered for these submodalities to determine whether they contribute to the sentiment estimation performance.

The last row in Table 6.3 depicts the sentiment estimation performance by five human annotators (the Cronbach alpha value was 0.83 for the TS annotation, indicating the reliability of the TS annotation). Our best MAE of 1.033 is close to the human performance, which had an MAE of 1.008, although there is still a gap between the correlation coefficients (0.276 vs 0.406). Thus, the preprocessing method and neural network architecture could be improved. We focused on physiological signals in this study since physiological signals can capture sentiment changes that cannot be expressed by textual, acoustic and visual features. The combination of our proposed method and other nonverbal subnetworks for audiovisual modalities, such as those proposed in [159], may further improve the sentiment estimation performance; thus, additional investigations are needed.

6.5.4 Analysis of the Attention Weight

It is assumed that the incorporation of physiological signals into the Transformer architecture leads to changes in the attention weights since time-series physiological signals shift the query from Q to Q' in our proposed module (Figure 6.2). Thus, we compared the attention weights between the Transformer and TPTr models. Test samples were used to extract attention weights from the learned model. Figure 6.3 shows examples of attention weights with negative sentiment (Figure 6.3(a)) and positive sentiment (Figure 6.3(b)) derived from the Transformer (left) and TPTr (center) models, as well as their difference (right). The example shown in Figure 6.3(a) has a true SS score of 3.00 (i.e., a negative example), and the estimated scores of the Transformer and TPTr models are 4.01 and 3.69, respectively. In this example, the segmented Japanese tokens of the system are "SO/NA/N/DESU/NE/," (number of tokens $n = 6$), which means "I got it" in English, and the Japanese token of the user is "HAI/," ($n = 2$), which means "Yes" or "Well", which generally functions as a filler and has a neutral or positive meaning. This ambiguous user utterance makes it difficult

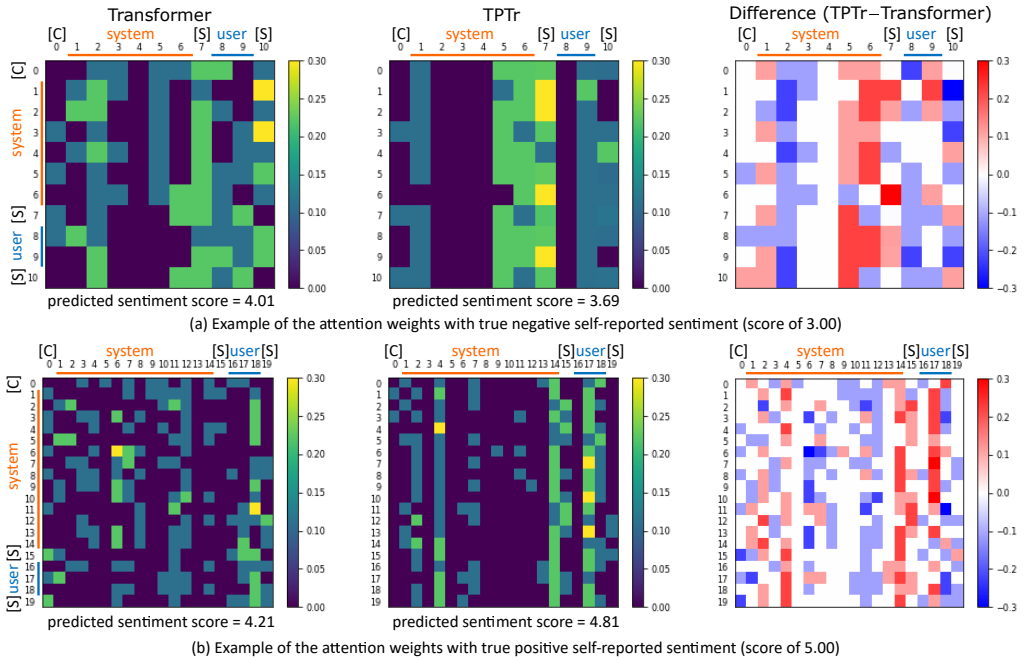

Figure 6.3: Example of the attention weights extracted from Transformer (left) and TPTr (center), and the difference between the two (right). Each square matrix is the attention weight computed from the $Q_i K_i^T$ (left) or $Q_i' K_i^T$ (center, please see equation 6.3). The dimension is equal to the total number of tokens including special tokens in one exchange. (a) Example of attention weights with true negative SS. (b) Example of attention weights with true positive SS. [C] and [S] indicate special tokens of BERT `[CLS]` and `[SEP]`, respectively.

to estimate negative sentiment using only linguistic information; however, the TPTr model gives less attention to this neutral/positive token and estimates a value of 3.69, which is close to the true negative sentiment score. Conversely, the TPTr model pays more attention to user utterances in other cases, as shown in Figure 6.3(b). This example has the system utterance "Which do you like better, sweet or spicy?" and the user utterance "I like both" in English. This example has a true SS score of 5.00 (i.e., a positive example), and the estimated scores of the Transformer and TPTr models are 4.21 and 4.81, respectively. Thus, the TPTr model may change the attention weight more flexibly than the Transformer model, which may improve the ensemble model performance.

Taken together, our proposed TPTr architecture intuitively allows for shifting BERT representations to the physiology-related subspace, resulting

in better estimation performance in the ensemble models. Our proposed models allow physiological information to be continuously linked to linguistic information and has a fixed attention direction, which is different from the prior works [154, 157, 155]. In the preliminary experiment, other architectural designs of the TPTr, such as another attention direction, degraded (or at least did not improve) the estimation performance. Thus, the time-series physiological signals play a supporting role to the Transformer based on the BERT representations (denoted as Q'_i in Section 6.3.2) by capturing SS changes that cannot be represented by linguistic information, although a further thorough investigation is needed.

6.5.5 Analysis of the Exchange-Level Estimation Pattern

To visualize exchange-level SS changes and differences in the estimation patterns among the models, an example of the estimation results during a dialogue session is shown in Figure 6.4. As shown by the black lines, the participant's SS changes dynamically during the dialogue. Thus, SS estimation is a difficult task, and dialogue systems should recognize and adapt to these sentiment changes at the exchange level. In this example, the conventional $\mathrm{FNN_L}$ (blue line in Figure 6.4, MAE of 0.954) and $\mathrm{FNN_P}$ (green dashed line, MAE of 1.077) models cannot dynamically estimate the participant's sentiment and perform conservatively (estimated scores are almost neural scores of 4). In addition, the conventional Transformer model (purple dotted line, MAE of 0.715) is insufficient for estimating positive sentiment, although some performance improvement is observed. On the other hand, the TPTr model (red dot-dashed line, MAE of 0.576) is effective in detecting subtle sentiment changes, particularly positive sentiment changes, which cannot be achieved by any of the other models presented in this example. Thus, the TPTr model could represent different aspects of sentiment changes that cannot be captured by using BERT representations or conventional Transformer.

6.5.6 Limitations and Future Works

There is no publicly available dataset that includes exchange-level SS labels and linguistic and physiological information except for the Hazumi dataset used in this study. Thus, we cannot evaluate our proposed model with another dataset, which will be considered in future work. Although our proposed method could contribute toward capturing short-time sentiment changes during individual exchanges (i.e., intraframe), our methods do not con-

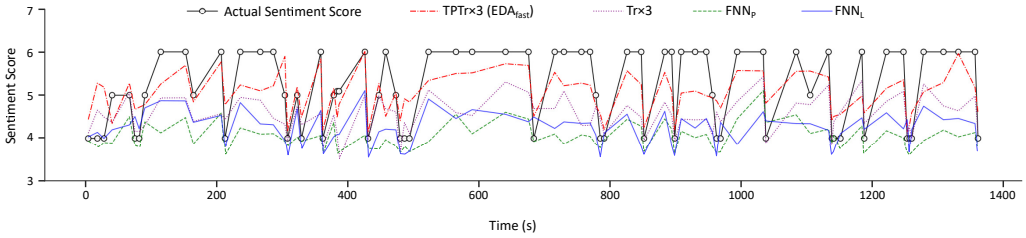

Figure 6.4: Qualitative example of the estimation pattern of each model. The black line with open circles indicates the actual sentiment score of a participant during each exchange in a dialogue session. Each colored line indicates the estimated score based on each model during each exchange.

sider time-series changes in the overall dialogue data (i.e., inter-utterance). Thus, the effectiveness of representations based on exchange sequences and attention mechanisms that capture more context and physiological changes merit further investigation. Also, there is a need to investigate effective methods for adding time-series audiovisual signals into TPTr (i.e., four modalities in total), and comparison with other SOTA language models such as RoBERTa [167] is also needed.

6.6 Chapter Summary

We showed that the model based on our proposed time-series physiological signal processing method has a comparable performance to linguistic-based models. Furthermore, the TPTr model, which introduced time-series physiological signals into a SOTA language model, significantly outperforms the previously reported best result. Furthermore, we presented that adding the BVP signal into the TPTr model based on the EDA_{fast} signal resulted in further estimation performance improvement. It seemed that attention weights based only on the language modality can be changed by the injection of the physiological signals into TPTr which capture SS changes that are not expressed in linguistic information. Thus, our proposed framework could be valuable for developing novel techniques for extracting representations not only from linguistic modality but physiological modality.

Chapter 7

Conclusion

This thesis addressed unresolved issues of physiological signals. Conclusions and future works are described in this chapter.

First, by considering individual physiological differences as a covariate shift, IW-LR and IW-SVM, which mitigate the accuracy degradation due to physiological individual differences in the training data, were created. As a result, most of the importance-weighted models outperform conventional models based on ECG and GSR features in emotion estimation. In the personality estimation, the IW method improves the macroaveraged F1-score for all SVM models. The best performing model (GSR model) outperformed the model with the best previously reported macroaveraged F1-score by 1.9% in personality estimation. Although physiological signals are often used in affective computing, individual physiological differences have been almost ignored in previous studies. In this work (Chapter 3), the fundamental issue is resolved by applying the IW method for the first time.

Second, although Chapter 3 focuses on physiological individual differences as a fundamental problem, Chapter 4 presented the application of physiological signals for sensing dynamic changes in user sentiment levels in human-agent interactions. Online SS estimation is known to be a relatively challenging task because current emotional states are not always expressed in a natural setting, but this thesis demonstrated the effectiveness of the physiological signals collected in a naturalistic human-agent interaction setting for the first time. Finding that exchange-level physiological signals are useful even under naturalistic human-agent interactions is valuable for the development of adaptive dialogue systems.

Third, this thesis further demonstrated the effectiveness of physiological signals and clarified their position in multimodal processing by comprehensive and thorough analysis (Chapter 5). In this work, the effects of physiological signals in multimodal sentiment analysis were investigated by evaluating

all of the fusion models for different types of sentiment estimation in naturalistic human-agent interaction settings. The results suggest that physiological features are effective in the unimodal model and that the fusion of linguistic representations with physiological features provides the best results for estimating SS labels as annotated by the users. In contrast, the tensor fusion of linguistic representations with audiovisual features is effective for estimating sentiment labels as annotated by third-party in regression tasks, which can be derived from the corresponding signals that are observable by humans. That is, it was newly revealed that different modalities play different roles in sentiment estimation during human-agent interaction.

Fourth, new physiological signal processing methods that are robust against changes in sentiment state masked by the user are proposed in Chapter 6. This work proposes an effective fusing algorithm that combines physiological features and a Transformer language model. Compared with linguistic models based on BERT representations, physiological LSTM models based on our proposed physiological signal processing method have competitive performance. Moreover, we extend our physiological signal processing method to the Transformer language model and propose TPTr, which captures sentiment changes based on both linguistic and physiological information. The ensemble with the TPTr model significantly outperforms the previous best result. Furthermore, the effectiveness of the proposed approach was also shown by applying a variety of physiological signals, further improving the performance with ensemble methods. This novel and unique model promises to be a powerful tool to realize adaptive dialogue systems and related emotionally intelligent systems.

From a higher perspective than each task (Chapter 3 to 6), further investigation of multimodal processing with physiological signals is still needed in future works. First, issues of individual differences are still not resolved in human-agent interactions, as shown in Figure 4.4 in Chapter 4, although it is shown that the IW method is effective for movie-watching tasks (Chapter 3). Thus, applying the IW method to sentiment estimation during dialogue (Chapter 4 to 6) is one of the most important future works. In this case, there is a concern that the method for importance estimation used in this thesis, KuLSIF [95], is applicable only for linear models such as SVM and LR and is not applicable for nonlinear models such as DNN, including Transformer. Alternatively, a recently proposed method called dynamic importance weighting can improve IW for deep learning under covariate shift [168]. Thus, this technique could bring further performance improvement for sentiment estimation based on physiological signals.

Additionally, it would be interesting to investigate when individual differences are prominent in sentiment estimation and how to ingeniously comple-

ment them with each modality. Obviously, human beings also have individual differences other than physiological differences. Thus, it seems that not only physiological signals but also dynamic adjustment with other modalities will be needed for further understanding of human behavior by machines.

Furthermore, there is a need to implement the proposed model in actual dialogue systems and evaluate the effectiveness of our proposed method for user satisfaction with the overall dialogue. This is a key evaluation to clarify the significance of the physiological signals in an actual system. Additionally, the potential of physiological signal modeling is not limited to spoken dialogue systems. For example, sensing techniques using wearable devices can be used for sentiment analysis in text-to-text dialogue where audiovisual data are not available. Thus, evaluation of the physiological signals in text-to-text interaction settings would further clarify the usefulness of the physiological signals.

Finally, to optimize the machine learning model for each dialogue setting, such as chit-chat or text-to-text, large-scale pretraining of the physiological signals may be plausible. The pretraining of the physiological signals can leverage the optimization for the internal state estimation, such as emotion, sentiment, attitude and engagement estimation. Similar to BERT [131] and LaMDA [169], self-supervised learning was used for emotion recognition with physiological signals, which was proposed recently [170]. Thus, a deeper understanding of multimodal processing, model architecture and learning design will lead to the realization of adaptive dialogue systems and emotionally intelligent agents.

Bibliography

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

[2] Grand View Research Inc. "Affective computing market size, share & trends analysis report by technology (touch-based, touchless), by software, by hardware, by end-use (healthcare automotive), and segment forecasts, 2020 - 2027" (2020). online.

[3] Chloe Clavel and Zoraida Callejas. "Sentiment analysis: from opinion mining to human-agent interaction". *IEEE Transactions on Affective Computing* 7.1 (2015), pp. 74–93.

[4] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. "A survey of multimodal sentiment analysis". *Image and Vision Computing* 65 (2017), pp. 3–14.

[5] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. "A review of affective computing: From unimodal analysis to multimodal fusion". *Information Fusion* 37 (2017), pp. 98–125.

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

[7] Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. "Utterance-level multimodal sentiment analysis". *Association for Computational Linguistics (ACL)* (2013), pp. 973–982.

[8] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages". *IEEE Intelligent Systems* 31.6 (2016), pp. 82–88.

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

[10] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. "IEMOCAP: Interactive emotional dyadic motion capture database". *Language Resources and Evaluation* 42.4 (2008), pp. 335–359.

[11] Khiet P Truong, David A Van Leeuwen, and Franciska MG De Jong. "Speech-based recognition of self-reported and observed emotion in a dimensional space". *Speech Communication* 54.9 (2012), pp. 1049–1063.

[12] R. Benjamin Knapp, Jonghwa Kim, and Elisabeth André. "Physiological signals and their use in augmenting emotion recognition for human–machine interaction". *Emotion-Oriented Systems* (2011), pp. 133–159.

[13] Dominik Leiner, Andreas Fahr, and Hannah Früh. "EDA positive change: A simple algorithm for electrodermal activity to measure general audience arousal during media exposure". *Communication Methods and Measures* 6.4 (2012), pp. 237–250.

[14] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. "Toward machine emotional intelligence: Analysis of affective physiological state". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23.10 (2001), pp. 1175–1191.

[15] Richard D Lane, Kateri McRae, Eric M Reiman, Kewei Chen, Geoffrey L Ahern, and Julian F Thayer. "Neural correlates of heart rate variability during emotion". *Neuroimage* 44.1 (2009), pp. 213–222.

[16] Hidetoshi Shimodaira. "Improving predictive inference under covariate shift by weighting the log-likelihood function". *Journal of Statistical Planning and Inference* 90.2 (2000), pp. 227–244.

[17] James J Gross and Lisa Feldman Barrett. "Emotion generation and emotion regulation: One or two depends on your point of view". *Emotion Review* 3.1 (2011), pp. 8–16.

[18] Paul Ekman. "Facial expression and emotion." *American Psychologist* 48.4 (1993), p. 384.

[19] Gary R VandenBos. *APA dictionary of psychology*. American Psychological Association, 2007.

[20] John D Mayer and Peter Salovey. "The intelligence of emotional intelligence". *Intelligence* 17.4 (1993), pp. 433–442.

[21] Junjie Lin, Wenji Mao, and Daniel D Zeng. "Personality-based refinement for sentiment classification in microblog". *Knowledge-Based Systems* 132 (2017), pp. 204–214.

[22] Daniel Goleman. *Emotional intelligence*. Bantam Books New York, 1995.

[23] Oswald Barral and Giulio Jacucci. "Applying physiological computing methods to study psychological, affective and motivational relevance". *International Workshop on Symbiotic Interaction (Symbiotic)* (2015), pp. 35–46.

[24] Rosalind W Picard. "Automating the recognition of stress and emotion: From lab to real-world impact". *IEEE MultiMedia* 23.3 (2016), pp. 3–7.

[25] Barbara Giżycka and Grzegorz J Nalepa. "Emotion in models meets emotion in design: Building true affective games". *Games, Entertainment, Media Conference (GEM)* (2018), pp. 1–5.

[26] Yantao Li, Gang Zhou, Daniel Graham, and Andrew Holtzhauer. "Towards an EEG-based brain-computer interface for online robot control". *Multimedia Tools and Applications* 75.13 (2016), pp. 7999–8017.

[27] Jonghwa Kim and Elisabeth André. "Emotion recognition based on physiological changes in music listening". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30.12 (2008), pp. 2067–2083.

[28] Anne-Flore Nicole Marie Perrin, He Xu, Eleni Kroupi, Martin Řeřábek, and Tourajd Ebrahimi. "Multimodal dataset for assessment of quality of experience in immersive multimedia". *ACM International Conference on Multimedia (ACM-MM)* (2015), pp. 1007–1010.

[29] Omar AlZoubi, Sidney K D'Mello, and Rafael A Calvo. "Detecting naturalistic expressions of nonbasic affect using physiological signals". *IEEE Transactions on Affective Computing* 3.3 (2012), pp. 298–310.

[30] Kevin Brady, Youngjune Gwon, Pooya Khorrami, Elizabeth Godoy, William Campbell, Charlie Dagli, and Thomas S. Huang. "Multi-modal audio, video and physiological sensor learning for continuous emotion prediction". *International Workshop on Audio/Visual Emotion Challenge (AVEC)* (2016), pp. 97–104.

[31] Martin Gjoreski, Hristijan Gjoreski, Mitja Luštrek, and Matjaž Gams. "Deep affect recognition from R-R intervals". *International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp) and International Symposium on Wearable Computers (ISWC)* (2017), pp. 754–762.

[32] Juan Abdon Miranda-Correa and Ioannis Patras. "A multi-task cascaded network for prediction of affect, personality, mood and social context using EEG signals". *International Conference on Automatic Face Gesture Recognition (FG)* (2018), pp. 373–380.

[33] Paul Ekman. "Emotions revealed". *British Medical Journal* 328.Suppl S5 (2004).

[34] Jeffrey F Cohn. "Foundations of human computing: facial expression and emotion". *International Conference on Multimodal Interfaces (ICMI)* (2006), pp. 233–238.

[35] Klaus R Scherer. "Vocal communication of emotion: A review of research paradigms". *Speech Communication* 40.1-2 (2003), pp. 227–256.

[36] Maja Pantic, Anton Nijholt, Alex Pentland, and Thomas S Huanag. "Human-Centred Intelligent Human? Computer Interaction (HCI^2): how far are we from attaining it?" *International Journal of Autonomous and Adaptive Communications Systems* 1.2 (2008), pp. 168–187.

[37] Paul Ekman and Wallace V Friesen. "Facial action coding system". *Environmental Psychology & Nonverbal Behavior* (1978).

[38] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. "Social signal processing: Survey of an emerging domain". *Image and Vision Computing* 27.12 (2009), pp. 1743–1759.

[39] Themis Balomenos, Amaryllis Raouzaiou, Spiros Ioannou, Athanasios Drosopoulos, Kostas Karpouzis, and Stefanos Kollias. "Emotion analysis in man-machine interaction systems". *International Workshop on Machine Learning for Multimodal Interaction (MLMI)* (2004), pp. 318–328.

[40] Koji Inoue, Divesh Lala, Katsuya Takanashi, and Tatsuya Kawahara. "Latent character model for engagement recognition based on multimodal behaviors". *International Workshop on Spoken Dialogue System Technology (IWSDS)* (2019), pp. 119–130.

[41] Sayaka Tomimasu and Masahiro Araki. "Assessment of users' interests in multimodal dialog based on exchange unit". *Proceedings of the Workshop on Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction (MA3HMI)* (2016), pp. 33–37.

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

[43] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent". *IEEE Transactions on Affective Computing* 3.1 (2011), pp. 5–17.

[44] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions". *International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (2013), pp. 1–8.

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

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

[47] Sidney K D'mello and Jacqueline Kory. "A review and meta-analysis of multimodal affect detection systems". *ACM Computing Surveys (CSUR)* 47.3 (2015), pp. 1–36.

[48] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge". *International Workshop on Audio/Visual Emotion Challenge (AVEC)* (2016), pp. 3–10.

[49] Yiqun Yao, Verónica Pérez-Rosas, Mohamed Abouelenien, and Mihai Burzo. "MORSE: MultimOdal sentiment analysis for Real-life SEttings". *International Conference on Multimodal Interaction (ICMI)* (2020), pp. 387–396.

[50] Nikhil Singh, Kegan James Moneghetti, Jeffrey Wilcox Christle, David Hadley, Victor Froelicher, and Daniel Plews. "Heart rate variability: An old metric with new meaning in the era of using mHealth technologies for health and exercise training guidance. Part two: Prognosis and training". *Arrhythmia & Electrophysiology Review* 7.4 (2018), pp. 247–255.

[51] Hans J Eysenck. *Dimensions of personality*. Vol. 5. Transaction Publishers, 1950.

[52] Renée M Tobin, William G Graziano, Eric J Vanman, and Louis G Tassinary. "Personality, emotional experience, and efforts to control emotions". *Journal of Personality and Social Psychology* 79.4 (2000), pp. 656–669.

[53] Emma Komulainen, Katarina Meskanen, Jari Lipsanen, Jari Marko Lahti, Pekka Jylhä, Tarja Melartin, Marieke Wichers, Erkki Isometsä, and Jesper Ekelund. "The effect of personality on daily life emotional processes". *PLoS One* 9.10 (2014), e110907.

[54] Marvin Zuckerman. "Good and bad humors: Biochemical bases of personality and its disorders". *Psychological Science* 6.6 (1995), pp. 325–332.

[55] Ada H Zohar, C Robert Cloninger, Rollin McCraty, et al. "Personality and heart rate variability: Exploring pathways from personality to cardiac coherence and health". *Open Journal of Social Sciences* 1.6 (2013), pp. 32–39.

[56] Juan Abdon Miranda-Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. "AMIGOS: A dataset for affect, personality and mood research on individuals and groups". *IEEE Transactions on Affective Computing* 12.2 (2021), pp. 479–493.

[57] Gerhard Stemmler and Jan Wacker. "Personality, emotion, and individual differences in physiological responses". *Biological Psychology* 84.3 (2010), pp. 541–551.

[58] Vitaliy Kolodyazhniy, Sylvia D Kreibig, James J Gross, Walton T Roth, and Frank H Wilhelm. "An affective computing approach to physiological emotion specificity: Toward subject-independent and stimulus-independent classification of film-induced emotions". *Psychophysiology* 48.7 (2011), pp. 908–922.

[59] Abraham Wald. "Foundations of a general theory of sequential decision functions". *Econometrica, Journal of the Econometric Society* (1947), pp. 279–313.

[60] Gerrit J.J. van den Burg and Alfred O. Hero. "Fast meta-learning for adaptive hierarchical classifier design". *arXiv preprint 1711.03512* (2017).

[61] Hatice Gunes and Maja Pantic. "Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners". *International Conference on Intelligent Virtual Agents (IVA)* (2010), pp. 371–377.

[62] Florian Eyben, Martin Wöllmer, Michel F Valstar, Hatice Gunes, Björn Schuller, and Maja Pantic. "String-based audiovisual fusion of behavioural events for the assessment of dimensional affect". *International Conference on Automatic Face and Gesture Recognition (FG)* (2011), pp. 322–329.

[63] Daniel McDuff, Rana Kaliouby, Thibaud Senechal, May Amr, Jeffrey Cohn, and Rosalind Picard. "Affectiva-MIT facial expression dataset (AM-FED): Naturalistic and spontaneous facial expressions collected". *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2013), pp. 881–888.

[64] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. "A multimodal database for affect recognition and implicit tagging". *IEEE Transactions on Affective Computing* 3.1 (2011), pp. 42–55.

[65] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. "DEAP: A database for emotion analysis; using physiological signals". *IEEE Transactions on Affective Computing* 3.1 (2011), pp. 18–31.

[66] Mojtaba Khomami Abadi, Ramanathan Subramanian, Seyed Mostafa Kia, Paolo Avesani, Ioannis Patras, and Nicu Sebe. "DECAF: MEG-based multimodal database for decoding affective

physiological responses". *IEEE Transactions on Affective Computing* 6.3 (2015), pp. 209–222.

[67] François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. "Using linguistic cues for the automatic recognition of personality in conversation and text". *Journal of Artificial Intelligence Research* 30 (2007), pp. 457–500.

[68] Ligia Batrinca, Bruno Lepri, Nadia Mana, and Fabio Pianesi. "Multimodal recognition of personality traits in human-computer collaborative tasks". *International Conference on Multimodal Interaction (ICMI)* (2012), pp. 39–46.

[69] Alexei V Ivanov, Giuseppe Riccardi, Adam J Sporka, and Jakub Franc. "Recognition of personality traits from human spoken conversations". *Annual Conference of the International Speech Communication Association (Interspeech)* (2011).

[70] Ramanathan Subramanian, Julia Wache, Mojtaba Khomami Abadi, Radu L Vieriu, Stefan Winkler, and Nicu Sebe. "ASCERTAIN: Emotion and personality recognition using commercial sensors". *IEEE Transactions on Affective Computing* 9.2 (2016), pp. 147–160.

[71] Paul T Costa and Robert R McCrae. *Revised NEO personality inventory (NEO-PI-R) and Neo five-factor inventory (NEO-FFI)*. Psychological Assessment Resources, 1992.

[72] Wouter M Kouw and Marco Loog. "A review of domain adaptation without target labels". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.3 (2019), pp. 766–785.

[73] Makoto Yamada, Masashi Sugiyama, and Tomoko Matsui. "Semi-supervised speaker identification under covariate shift". *Signal Processing* 90.8 (2010), pp. 2353–2361.

[74] Yan Li, Hiroyuki Kambara, Yasuharu Koike, and Masashi Sugiyama. "Application of covariate shift adaptation techniques in brain–computer interfaces". *IEEE Transactions on Biomedical Engineering* 57.6 (2010), pp. 1318–1324.

[75] Sicheng Zhao, Guiguang Ding, Jungong Han, and Yue Gao. "Personality-aware personalized emotion recognition from physiological signals." *International Joint Conference on Artificial Intelligence (IJCAI)* (2018), pp. 1660–1667.

[76] Ross Harper and Joshua Southern. "A Bayesian deep learning framework for end-to-end prediction of emotion from heartbeat". *IEEE Transactions on Affective Computing* 13.2 (2022), pp. 985–991.

[77] Martin Gjoreski, Blagoj Mitrevski, Mitja Luštrek, and Matjaž Gams. "An inter-domain study for arousal recognition from physiological signals". *Informatica* 42.1 (2018), pp. 61–68.

[78] Valentina Markova and Todor Ganchev. "Automated recognition of affect and stress evoked by audio-visual stimuli". *Balkan Conference on Lighting (BalkanLight)* (2018), pp. 1–4.

[79] Wenxuan Mou, Hatice Gunes, and Ioannis Patras. "Alone versus in-a-group: A multi-modal framework for automatic affect recognition". *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15.2 (2019), pp. 1–23.

[80] Kuan Tung, Po-Kang Liu, Yu-Chuan Chuang, Sheng-Hui Wang, and An-Yeu Andy Wu. "Entropy-assisted multi-modal emotion recognition framework based on physiological signals". *IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)* (2018), pp. 22–26.

[81] Sheng-Hui Wang, Huai-Ting Li, En-Jui Chang, and An-Yeu (Andy) Wu. "Entropy-assisted emotion recognition of valence and arousal using XGBoost classifier". *Artificial Intelligence Applications and Innovations (AIAI)* (2018), pp. 249–260.

[82] En-Jui Chang, Abbas Rahimi, Luca Benini, and An-Yeu Andy Wu. "Hyperdimensional computing-based multimodality emotion recognition with physiological signals". *International Conference on Artificial Intelligence Circuits and Systems (AICAS)* (2019), pp. 137–141.

[83] Hao-Chun Yang and Chi-Chun Lee. "An attribute-invariant variational learning for emotion recognition using physiology". *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2019), pp. 1184–1188.

[84] Luz Santamaria-Granados, Mario Munoz-Organero, Gustavo Ramirez-Gonzalez, Enas Abdulhay, and NJIA Arunkumar. "Using deep convolutional neural network for emotion detection on a physiological signals dataset (AMIGOS)". *IEEE Access* 7 (2018), pp. 57–67.

[85] Tzyy-Ping Jung, Terrence J Sejnowski, et al. "Multi-modal approach for affective computing". *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2018), pp. 291–294.

[86] Alessandro Vinciarelli and Gelareh Mohammadi. “A survey of personality computing”. *IEEE Transactions on Affective Computing* 5.3 (2014), pp. 273–291.

[87] James A Russell. “A circumplex model of affect”. *Journal of Personality and Social Psychology* 39.6 (1980), pp. 1161–1178.

[88] Joseph P Zbilut, Nitza Thomasson, and Charles L Webber. “Recurrence quantification analysis as a tool for nonlinear exploration of nonstationary cardiac signals”. *Medical Engineering & Physics* 24.1 (2002), pp. 53–60.

[89] Yuta Tsuboi, Hisashi Kashima, Shohei Hido, Steffen Bickel, and Masashi Sugiyama. “Direct density ratio estimation for large-scale covariate shift adaptation”. *Journal of Information Processing* 17 (2009), pp. 138–155.

[90] Masashi Sugiyama. “Learning under non-stationarity: Covariate shift adaptation by importance weighting”. *Handbook of Computational Statistics*. Springer, 2012, pp. 927–952.

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

[92] Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex J Smola. “Correcting sample selection bias by unlabeled data”. *Advances in Neural Information Processing Systems (NIPS)* (2007), pp. 601–608.

[93] Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. “Direct importance estimation for covariate shift adaptation”. *Annals of the Institute of Statistical Mathematics* 60.4 (2008), pp. 699–746.

[94] Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. “A least-squares approach to direct importance estimation”. *The Journal of Machine Learning Research* 10 (2009), pp. 1391–1445.

[95] Takafumi Kanamori, Taiji Suzuki, and Masashi Sugiyama. “Statistical analysis of kernel-based least-squares density-ratio estimation”. *Machine Learning* 86.3 (2012), pp. 335–367.

[96] Sanford Weisberg. *Applied linear regression*. Vol. 528. John Wiley & Sons, 2005.

[97] Yash Mehta, Navonil Majumder, Alexander Gelbukh, and Erik Cambria. "Recent trends in deep learning based personality detection". *Artificial Intelligence Review* 53.4 (2020), pp. 2313–2339.

[98] Ali Hassan, Robert Damper, and Mahesan Niranjan. "On acoustic emotion recognition: compensating for covariate shift". *IEEE Transactions on Audio, Speech, and Language Processing* 21.7 (2013), pp. 1458–1468.

[99] Takafumi Kanamori, Taiji Suzuki, and Masashi Sugiyama. "Computational complexity of kernel-based density-ratio estimation: A condition number analysis". *Machine Learning* 90.3 (2013), pp. 431–460.

[100] Ivor W. Tsang, James T. Kwok, and Pak-Ming Cheung. "Core vector machines: Fast SVM training on very large data sets." *Journal of Machine Learning Research* 6.13 (2005), pp. 363–392.

[101] Niels Landwehr, Mark Hall, and Eibe Frank. "Logistic model trees". *Machine Learning* 59.1-2 (2005), pp. 161–205.

[102] Garrett Wilson and Diane J Cook. "A survey of unsupervised deep domain adaptation". *ACM Transactions on Intelligent Systems and Technology* 11.5 (2020), pp. 1–46.

[103] Zhen Ren, Xin Qi, Gang Zhou, and Haining Wang. "Exploiting the data sensitivity of neurometric fidelity for optimizing EEG sensing". *IEEE Internet of Things Journal* 1.3 (2014), pp. 243–254.

[104] Kyriaki Kalimeri and Charalampos Saitis. "Exploring multimodal biosignal features for stress detection during indoor mobility". *International Conference on Multimodal Interaction (ICMI)* (2016), pp. 53–60.

[105] Phuong Pham and Jingtao Wang. "Adaptive review for mobile mooc learning via multimodal physiological signal sensing-a longitudinal study". *International Conference on Multimodal Interaction (ICMI)* (2018), pp. 63–72.

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

[107] Chuan-Yu Chang, Jeng-Shiun Tsai, Chi-Jane Wang, and Pau-Choo Chung. "Emotion recognition with consideration of facial expression and physiological signals". *Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)* (2009), pp. 278–283.

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

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

[110] Utkarsh Chauhan, Norbert Reithinger, and John R Mackey. "Real-time stress assessment through PPG sensor for VR biofeedback". *International Conference on Multimodal Interaction (ICMI): Adjunct* (2018), pp. 1–5.

[111] Yuning Qiu, Teruhisa Misu, and Carlos Busso. "Driving anomaly detection with conditional generative adversarial network using physiological and CAN-Bus data". *International Conference on Multimodal Interaction (ICMI)* (2019), pp. 164–173.

[112] Iulia Lefter and Siska Fitrianie. "The multimodal dataset of negative affect and aggression: A validation study". *International Conference on Multimodal Interaction (ICMI)* (2018), pp. 376–383.

[113] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. "Introducing WESAD, a multimodal dataset for wearable stress and affect detection". *International Conference on Multimodal Interaction (ICMI)* (2018), pp. 400–408.

[114] Dan Bohus and Eric Horvitz. "Learning to predict engagement with a spoken dialog system in open-world settings". *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)* (2009), pp. 244–252.

[115] Yukiko I. Nakano and Ryo Ishii. "Estimating user's engagement from eye-gaze behaviors in human-agent conversations". *International Conference on Intelligent User Interfaces (IUI)* (2010), pp. 139–148.

[116] Yuki Hirano, Shogo Okada, Haruto Nishimoto, and Kazunori Komatani. "Multitask prediction of exchange-level annotations for multimodal dialogue systems". *International Conference on Multimodal Interaction (ICMI)* (2019), pp. 85–94.

[117] Thierry Chaminade, Léo Biaocchi, Farah H Wolfe, Noël Nguyen, and Laurent Prévot. "Communicative behavior and physiology in social interactions". *Workshop on Modeling INTERPERsonal SynchrONy And infLuence (INTERPERSONAL)* (2015), pp. 25–30.

[118] Olga Egorow and Andreas Wendemuth. "Detection of challenging dialogue stages using acoustic signals and biosignals". *Conference on Computer Graphics, Visualization and Computer Vision (WSCG)* (2016), pp. 137–143.

[119] Kazunori Komatani, Shogo Okada, Haruto Nishimoto, Masahiro Araki, and Mikio Nakano. "Multimodal dialogue data collection and analysis of annotation disagreement". *International Workshop on Spoken Dialogue Systems (IWSDS)* (2019), pp. 201–213.

[120] Kazunori Komatani and Shogo Okada. "Collection and analysis of human-system multimodal dialogue data with subjective ratings". *IEICE Technical Report (in Japanese)* 119.179 (2019), pp. 21–26.

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

[122] Björn Schuller, Stefan Steidl, and Anton Batliner. "The INTERSPEECH 2009 emotion challenge". *Annual Conference of the International Speech Communication Association (INTERSPEECH)* (2009), pp. 312–315.

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

[124] Paul Ekman and Wallace V. Friesen. "Facial action coding system: A technique for the measurement of facial movement". *Palo Alto* 3 (1978).

[125] Klaus Weber, Hannes Ritschel, Ilhan Aslan, Florian Lingenfelser, and Elisabeth André. "How to shape the humor of a robot-social behavior adaptation based on reinforcement learning". *International Conference on Multimodal Interaction (ICMI)* (2018), pp. 154–162.

[126] Ashima Yadav and Dinesh Kumar Vishwakarma. "Sentiment analysis using deep learning architectures: a review". *Artificial Intelligence Review* 53.6 (2020), pp. 4335–4385.

[127] Yoon Kim. "Convolutional neural networks for sentence classification". *Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Oct. 2014), pp. 1746–1751.

[128] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. "Recursive deep models for semantic compositionality over a sentiment treebank". *Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Oct. 2013), pp. 1631–1642.

[129] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. "Deep contextualized word representations". *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)* (June 2018), pp. 2227–2237.

[130] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need". *Advances in Neural Information Processing Systems (NIPS)* 30 (2017), pp. 5998–6008.

[131] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "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)* (2019), pp. 4171–4186.

[132] Ye-Yi Wang, Li Deng, and Alex Acero. "Spoken language understanding". *IEEE Signal Processing Magazine* 22.5 (2005), pp. 16–31.

[133] Robert R McCrae and Oliver P John. "An introduction to the five-factor model and its applications". *Journal of Personality* 60.2 (1992), pp. 175–215.

[134] David Watson, Lee Anna Clark, and Auke Tellegen. "Development and validation of brief measures of positive and negative affect: The PANAS scales." *Journal of Personality and Social Psychology* 54.6 (1988), p. 1063.

[135] Ali Bakhshi and Stephan Chalup. "Multimodal emotion recognition based on speech and physiological signals using deep neural networks". *International Conference on Pattern Recognition (ICPR)* (2021), pp. 289–300.

[136] Kevin Doherty and Gavin Doherty. "Engagement in HCI: conception, theory and measurement". *ACM Computing Surveys (CSUR)* 51.5 (2018), pp. 1–39.

[137] Mounia Lalmas, Heather O'Brien, and Elad Yom-Tov. "Measuring user engagement". *Synthesis Lectures on Information Concepts, Retrieval, and Services* 6.4 (2014), pp. 1–132.

[138] Shun Katada, Shogo Okada, Yuki Hirano, and Kazunori Komatani. "Is she truly enjoying the conversation?: Analysis of physiological signals toward adaptive dialogue systems". *International Conference on Multimodal Interaction (ICMI)* (2020), pp. 315–323.

[139] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. "Learning word vectors for 157 languages". *arXiv preprint 1802.06893* (2018).

[140] Michael E Dawson, Anne M Schell, and Diane L Filion. "The electrodermal system". *Handbook of psychophysiology* (2000), pp. 200–223.

[141] Nadine Glas and Catherine Pelachaud. "Definitions of engagement in human-agent interaction". *International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE. 2015, pp. 944–949.

[142] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. "Applying conditional random fields to Japanese morphological analysis". *Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2004), pp. 230–237.

[143] Nozomi Kobayashi, Kentaro Inui, Yuji Matsumoto, Kenji Tateishi, and Toshikazu Fukushima. "Collecting evaluative expressions for opinion extraction". *International Conference on Natural Language Processing (ICNLP)* (2004), pp. 596–605.

[144] Masahiko Higashiyama, Kentaro Inui, and Yuji Matsumoto. "Learning sentiment of nouns from selectional preferences of verbs and adjectives". *Annual Meeting of the Association for Natural Language Processing* (2008), pp. 584–587.

[145] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. "Improving language understanding with unsupervised learning". *Technical report, OpenAI* (2018).

[146] Tatsuki Akahori, Kohji Dohsaka, Masaki Ishii, and Hidekatsu Ito. "Efficient Creation of Japanese Tweet Emotion Dataset Using Sentence-Final Expressions". *Global Conference on Life Sciences and Technologies (LifeTech)* (2021), pp. 501–505.

[147] Wolfram Boucsein. *Electrodermal activity*. Springer Science & Business Media, 2012.

[148] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. "Multimodal machine learning: A survey and taxonomy". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.2 (2018), pp. 423–443.

[149] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. "Tensor fusion network for multimodal sentiment analysis". *Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2017), pp. 1103–1114.

[150] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." *Journal of Machine Learning Research* 9.11 (2008), pp. 2579–2605.

[151] Sidney K D'Mello, Nia Dowell, and Art Graesser. "Unimodal and multimodal human perception of naturalistic non-basic affective states during human-computer interactions". *IEEE Transactions on Affective Computing* 4.4 (2013), pp. 452–465.

[152] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. "Advancements in noncontact, multiparameter physiological measurements using a webcam". *IEEE Transactions on Biomedical Engineering* 58.1 (2010), pp. 7–11.

[153] John D. Mayer and Peter Salovey. "The intelligence of emotional intelligence". *Intelligence* 17.4 (1993), pp. 433–442.

[154] Haifeng Chen, Dongmei Jiang, and Hichem Sahli. "Transformer encoder with multi-modal multi-head attention for continuous affect recognition". *IEEE Transactions on Multimedia* 23 (2021), pp. 4171–4183.

[155] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. "Integrating multimodal information in large pretrained Transformers". *Annual Meeting of the Association for Computational Linguistics (ACL)* (2020), pp. 2359–2369.

[156] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. "MISA: Modality-invariant and -specific representations for multimodal sentiment analysis". *ACM International Conference on Multimedia (ACM-MM)* (2020), pp. 1122–1131.

[157] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. "Multimodal Transformer for unaligned multimodal language sequences". *Annual Meeting of the Association for Computational Linguistics (ACL)* (2019), pp. 6558–6569.

[158] Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. "Multimodal affective analysis using hierarchical attention strategy with word-level alignment". *Annual Meeting of the Association for Computational Linguistics (ACL)* (2018), pp. 2225–2235.

[159] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. "Words can shift: Dynamically adjusting word representations using nonverbal behaviors". *Association for the Advancement of Artificial Intelligence Conference (AAAI)* (2019), pp. 7216–7223.

[160] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". *Neural Computation* 9.8 (1997), pp. 1735–1780.

[161] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition". *Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 770–778.

[162] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. "Layer normalization". *arXiv preprint 1607.06450* (2016).

[163] Heath Yates, Brent Chamberlain, Greg Norman, and William H. Hsu. "Arousal Detection for Biometric Data in Built Environments using Machine Learning". *International Joint Conference on Artificial Intelligence (IJCAI) Workshop on Artificial Intelligence in Affective Computing* 66 (2017), pp. 58–72.

[164] Marco Maier, Daniel Elsner, Chadly Marouane, Meike Zehnle, and Christoph Fuchs. "DeepFlow: Detecting optimal user experience from physiological data using deep neural networks". *International Joint Conference on Artificial Intelligence (IJCAI)* (2019), pp. 1415–1421.

[165] Jessica Sharmin Rahman, Tom Gedeon, Sabrina Caldwell, Richard Jones, Md Zakir Hossain, and Xuanying Zhu. "Melodious micro-frissons: detecting music genres from skin response". *International Joint Conference on Neural Networks (IJCNN)* (2019), pp. 1–8.

[166] Shun Katada, Shogo Okada, and Kazunori Komatani. "Effects of physiological signals in different types of multimodal sentiment estimation". *IEEE Transactions on Affective Computing* (2022). Online Available.

[167] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "RoBERTa: A robustly optimized BERT pretraining approach". *arXiv preprint 1907.11692* (2019).

[168] Tongtong Fang, Nan Lu, Gang Niu, and Masashi Sugiyama. "Rethinking importance weighting for deep learning under distribution shift". *Advances in Neural Information Processing Systems (NeurIPS)* 33 (2020), pp. 11996–12007.

[169] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. "LaMDA: Language models for dialog applications". *arXiv preprint 2201.08239* (2022).

[170] Juan Vazquez-Rodriguez, Grégoire Lefebvre, Julien Cumin, and James L Crowley. "Transformer-based self-supervised learning for emotion recognition". *arXiv preprint 2204.05103* (2022).

[171] Hajime Mochizuki. "Investigation of words in a Japanese closed caption TV corpus". *2019 STEAM Education PROCEEDINGS* (2019).

Appendix

Distribution of parts of speech

We investigated whether the user utterances in the Hazumi dataset include over/underrepresentation of certain types of words or have a similar distribution as another Japanese spoken language corpus. The distribution of the types of words is shown in Table S1. The numbers and ratios of tokens and types in user utterances are depicted. The distribution of the parts of speech (PoSs) in the total sample is similar to another Japanese spoken language corpus [171], although the ratios of nouns and symbols are slightly different since [171] was based on closed captioning. In addition, a similar distribution was observed by stratified analysis of positive and negative SS labels. As shown in Table S1, the utterances with positive SS labels had a larger number of tokens than the utterances with negative SS labels, probably because the participants enjoyed talking with the dialogue system and talked with various expressions.

Table S1: User utterance distribution. PNA, pre-noun adjectival.

PoS	Total sample				positive SS label				negative SS label				[171]
	Token		Type		Token		Type		Token		Type		Token
	#	%	#	%	#	%	#	%	#	%	#	%	%
Noun	8934	21.5	2050	63.3	5624	22.0	1545	61.7	3310	20.7	1031	59.0	33.7
Verb	4086	9.8	629	19.4	2530	9.9	485	19.4	1556	9.7	350	20.0	12.5
Adjective	967	2.3	160	4.9	576	2.2	127	5.1	391	2.4	81	4.6	1.6
Adverb	1689	4.1	148	4.6	1080	4.2	123	4.9	609	3.8	89	5.1	2.6
Interjection	2242	5.4	38	1.2	1165	4.5	31	1.2	1077	6.7	32	1.8	1.1
Symbol	7872	18.9	12	0.4	4634	18.1	10	0.4	3238	20.3	6	0.3	10.3
Particle	9517	22.9	81	2.5	6165	24.1	73	2.9	3352	21.0	68	3.9	25.5
Auxiliary verb	4736	11.4	43	1.3	2874	11.2	39	1.6	1862	11.6	34	1.9	9.8
Conjunction	224	0.5	27	0.8	141	0.6	25	1.0	83	0.5	19	1.1	0.9
PNA	454	1.1	21	0.6	293	1.1	20	0.8	161	1.0	16	0.9	1.1
Prefix	230	0.6	17	0.5	153	0.6	13	0.5	77	0.5	10	0.6	0.7
Filler	646	1.6	12	0.4	374	1.5	12	0.5	272	1.7	12	0.7	0.2
Total	41597	100.0	3238	100.0	25609	100.0	2503	100.0	15988	100.0	1748	100.0	100.0

Comparison of Multimodal Language Models

The unimodal BERT representation (L_c) models outperformed models trained on handcrafted features (L_a model) and fastText word vectors (L_b model). However, there is no guarantee that the better unimodal model will be a better multimodal model. We therefore performed an extensive investigation to explore the potential of other multimodal language models. The fusion model with either L_a or L_b did not outperform the fusion with L_c in the case of our study (Tables S2 and S3), although there were a few cases in which the L_b fusion model outperformed the L_c model when the same architecture and modality were compared (for example, in the SS estimation, the L_b+P FNN model with EF had an MAE of 1.082 (Table S2), whereas the L_c+P FNN model with EF had an MAE of 1.088 (Table 3 in the main text)). The highest estimation performance was achieved by using multimodal models with L_c in SS and TS estimation (please see the main text).

Confirming the reproducibility of our results, FNN with LF_2 of the linguistic (L_a or L_b) and physiological (P) modalities had good performance in SS estimation (Table S2). The L_b+P FNN model with LF_2 achieved an accuracy of 0.627, and the L_a+P FNN model with LF_2 achieved an MAE of 1.051 in the SS estimation. On the other hand, in the TS estimation, the L_b+V TFN model achieved an accuracy of 0.820, and the L_b+A+V TFN model achieved an MAE of 0.436 in the TS estimation (Table S3). Thus, the effective modality and architecture were the same among the three multimodal language models (fusion of L_a, L_b or L_c with another non-verbal modality) for different types of sentiment estimation.

Table S2: SS estimation results for multimodal models with handcrafted features or fastText word vectors. L_a, handcrafted linguistic features; L_b, fastText word vectors; P, physiological features; A, audio features; and V, visual features.

Feature	L-SVM			R-SVM			FNN EF			FNN LF_1			FNN LF_2			TFN		
	Acc	F1	MAE	Acc	F1	MAE	Acc	F1	MAE	Acc	F1	MAE	Acc	F1	MAE	Acc	F1	MAE
L_a+P	0.579	0.507	1.320	0.325	0.262	**1.083**	0.581	0.511	1.105	0.575	0.513	1.135	0.597	0.528	**1.051**	0.565	0.504	1.110
L_a+A	0.557	0.503	1.162	0.482	0.406	1.116	0.561	0.504	1.148	0.566	0.507	1.109	0.584	0.523	1.104	0.570	0.515	1.104
L_a+V	0.566	0.510	1.186	0.467	0.366	1.093	0.571	0.508	1.141	0.584	0.523	1.117	0.584	0.522	1.097	0.577	0.517	1.110
L_b+P	0.600	0.506	1.193	0.538	0.427	1.124	**0.604**	**0.536**	**1.082**	0.616	0.520	**1.098**	**0.627**	0.538	1.062	0.587	0.497	**1.069**
L_b+A	0.567	0.516	1.229	0.524	0.475	1.107	0.562	0.507	1.156	0.568	0.511	1.184	0.577	0.518	1.107	0.581	0.514	1.130
L_b+V	0.573	0.525	1.454	0.542	0.478	1.107	0.562	0.515	1.129	0.568	0.517	1.165	0.581	0.530	1.096	0.547	0.471	1.113
L_a+P+A	0.577	0.510	1.179	0.494	0.417	1.121	0.566	0.510	1.110	0.562	0.503	1.110	0.617	0.541	1.056	0.567	0.511	1.127
L_a+P+V	0.577	0.509	1.199	0.497	0.379	1.094	0.570	0.508	1.108	0.586	0.524	1.119	0.611	0.540	1.054	0.577	0.518	1.119
L_a+A+V	0.559	0.504	1.186	0.498	0.422	1.139	0.566	0.510	1.114	0.574	0.515	1.114	0.599	0.538	1.089	0.584	0.523	1.124
L_b+P+A	0.601	0.528	1.214	0.543	0.474	1.117	0.567	0.512	1.144	0.595	0.518	1.130	0.610	0.537	1.063	0.594	0.518	1.111
L_b+P+V	**0.619**	0.533	1.337	0.525	0.468	1.104	0.564	0.511	1.134	**0.622**	**0.535**	1.132	0.626	**0.550**	1.062	0.585	0.499	1.101
L_b+A+V	0.579	0.531	1.210	**0.560**	**0.503**	1.112	0.577	0.520	1.132	0.567	0.513	1.176	0.588	0.531	1.091	0.579	0.521	1.146
L_a+P+A+V	0.587	0.519	**1.156**	0.464	0.403	1.139	0.563	0.505	1.108	0.575	0.515	1.104	0.620	**0.550**	1.057	0.574	0.517	1.132
L_b+P+A+V	0.611	**0.540**	1.217	0.553	0.494	1.107	0.581	0.521	1.124	0.599	0.526	1.121	0.616	0.548	1.062	**0.616**	**0.536**	1.097

Table S3: TS estimation results for multimodal models with handcrafted features or fastText word vectors. L_a, handcrafted linguistic features; L_b, fastText word vectors; P, physiological features; A, audio features; and V, visual features.

Feature	L-SVM			R-SVM			FNN EF			FNN LF_1			FNN LF_2			TFN		
	Acc	F1	MAE	Acc	F1	MAE	Acc	F1	MAE	Acc	F1	MAE	Acc	F1	MAE	Acc	F1	MAE
L_a+P	0.781	0.737	0.656	0.686	0.405	0.770	0.797	0.763	0.502	0.795	0.756	0.558	0.800	0.758	0.548	0.769	0.733	0.655
L_a+A	0.764	0.716	0.651	0.689	0.417	0.618	0.776	0.727	0.504	0.785	0.743	0.515	0.786	0.735	0.513	0.781	0.734	0.641
L_a+V	0.783	0.739	0.643	0.683	0.405	0.645	0.784	0.746	0.502	0.798	0.759	0.522	0.803	**0.760**	0.529	0.785	0.743	0.653
L_b+P	**0.808**	**0.756**	0.668	0.704	0.573	0.775	**0.819**	**0.768**	0.506	**0.820**	**0.776**	0.500	**0.813**	0.759	0.538	0.809	0.758	0.458
L_b+A	0.759	0.695	0.717	**0.727**	0.593	0.725	0.765	0.711	0.517	0.771	0.719	0.504	0.777	0.718	**0.497**	0.810	0.766	0.446
L_b+V	0.797	0.736	0.788	0.691	0.432	0.988	0.806	0.750	0.507	0.809	0.767	0.497	0.811	0.755	0.514	**0.820**	**0.775**	0.447
L_a+P+A	0.770	0.720	0.649	0.687	0.414	**0.616**	0.777	0.730	0.498	0.786	0.743	0.516	0.788	0.731	0.532	0.783	0.737	0.644
L_a+P+V	0.790	0.746	0.645	0.682	0.403	0.651	0.790	0.750	0.496	0.800	0.761	0.524	0.801	0.745	0.545	0.789	0.747	0.653
L_a+A+V	0.767	0.720	**0.639**	0.686	0.417	0.617	0.780	0.732	0.489	0.789	0.749	0.504	0.798	0.748	0.519	0.789	0.741	0.639
L_b+P+A	0.756	0.691	0.732	0.719	0.578	0.737	0.758	0.701	0.519	0.772	0.721	0.501	0.788	0.720	0.524	0.808	0.763	0.441
L_b+P+V	0.797	0.735	0.768	0.692	0.427	1.010	0.804	0.743	0.508	0.806	0.762	0.494	0.805	0.738	0.537	0.818	0.774	0.451
L_b+A+V	0.764	0.704	0.720	0.723	0.595	0.698	0.775	0.719	0.510	0.775	0.724	**0.492**	0.801	0.746	0.507	0.809	0.763	**0.436**
L_a+P+A+V	0.769	0.720	0.650	0.688	0.415	**0.616**	0.779	0.729	**0.487**	0.789	0.747	0.497	0.797	0.741	0.533	0.790	0.742	0.642
L_b+P+A+V	0.766	0.707	0.712	0.722	**0.599**	0.670	0.781	0.725	0.508	0.773	0.721	0.503	0.800	0.736	0.526	0.815	0.772	0.453

Publication List

International Journal (peer reviewed)

Shun Katada and Shogo Okada. "Biosignal-based user-independent recognition of emotion and personality with importance weighting". Multimedia Tools and Applications, 2022. (corresponded to Chapter 3 in this thesis)

Shun Katada, Shogo Okada, and Kazunori Komatani. "Effects of physiological signals in different types of multimodal sentiment estimation". IEEE Transactions on Affective Computing. 2022. Online Available. (corresponded to Chapter 5 in this thesis)

International Conference (peer reviewed)

Shun Katada, Shogo Okada, Yuki Hirano, and Kazunori Komatani. "Is she truly enjoying the conversation?: Analysis of physiological signals toward adaptive dialogue systems". International Conference on Multimodal Interaction (ICMI) (2020). (corresponded to Chapter 4 in this thesis)

Shun Katada, Shogo Okada, and Kazunori Komatani. "Transformer-based physiological feature learning for multimodal analysis of self-reported sentiment". International Conference on Multimodal Interaction (ICMI) (2022), Accepted. (corresponded to Chapter 6 in this thesis)

Shun Katada, Kiyoaki Shirai, and Shogo Okada. "Incorporation of contextual information into BERT for dialog act classification in Japanese". International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP) (2021).