Integrating Physiological, Speech, and Textual Information Toward Real-Time Recognition of Emotional Valence in Dialogue

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

Abstract

Accurately estimating users' emotional states in real time is crucial for enabling dialogue systems to respond adaptively. While existing approaches primarily rely on verbal information, such as text and speech, these modalities are often unavailable in non-speaking situations. In such cases, non-verbal information, particularly physiological signals, becomes essential for understanding users' emotional states. In this study, we aimed to develop a model for real-time recognition of users' binary emotional valence (high-valence vs. low-valence) during conversations. Specifically, we utilized an existing Japanese multimodal dialogue dataset, which includes various physiological signals, namely electrodermal activity (EDA), blood volume pulse (BVP), photoplethysmography (PPG), and pupil diameter, along with speech and textual data. We classify the emotional valence of every 15-second segment of dialogue interaction by integrating such multimodal inputs. To this end, time-series embeddings of physiological signals are extracted using a self-supervised encoder, while speech and textual features are obtained from pre-trained Japanese HuBERT and BERT models, respectively. The modality-specific embeddings are integrated using a feature fusion mechanism for emotional valence recognition. Experimental results show that while each modality individually contributes to emotion recognition, the inclusion of physiological signals leads to a notable performance improvement, particularly in non-speaking or minimally verbal situations. These findings underscore the importance of physiological information for enhancing real-time valence recognition in dialogue systems, especially when verbal information is limited.

1 Introduction

Understanding and responding to users' emotional states is a fundamental capability for user-adaptive dialogue systems. Such systems must be able to detect and adapt to users' emotional states in real time to foster natural and engaging interactions. While prior work has made considerable progress in emotion recognition by leveraging verbal information (Majumder et al., 2019; Ghosal et al., 2019), notably text and speech, these approaches often fall short in non-speaking or minimally verbal contexts, which are common in real-world interactions. In these cases, relying solely on verbal information can hinder the system's ability to maintain emotional awareness and responsiveness.

To address this limitation, non-verbal modalities, especially physiological signals, offer valuable cues about users' internal states. Physiological signals, such as EDA and BVP, have been shown to reflect changes in emotional arousal and valence (Komatani and Okada, 2021; Saffaryazdi et al., 2022; Singh et al., 2024; Jiang et al., 2024). These signals are continuous, language-independent, and inherently real-time, making them suitable for supplementing verbal inputs during emotionally relevant but verbally sparse moments in dialogue.

In this study, we tackle binary emotional valence recognition (high-valence vs. low-valence) in conversational settings by integrating physiological signals with speech and text. As a preliminary step toward developing real-time emotion recognition systems, we develop a model to classify the emotional valence of each 15-second dialogue segment using a Japanese multimodal dialogue dataset that includes multiple physiological signals alongside speech recordings and textual transcripts. To this end, we adopt a multimodal recognition framework that combines (1) time-series embeddings of physiological signals extracted via a self-supervised encoder, (2) speech features from a pre-trained Japanese HuBERT model, and (3) text embeddings from a pre-trained Japanese BERT model. The modality-specific representations are then fused via a feature fusion mechanism and used for downstream classification.

Proceedings of the 26th Annual Meeting of SIGDIAL, pages 591–600
Aug 25–27, 2025. ©2025 SIGDIAL

Our findings demonstrate the effectiveness of integrating physiological signals with verbal information for emotional valence recognition and provide an assessment of the individual contributions of EDA, BVP, PPG, and pupil diameter. The results underscore the potential of physiological signals to enhance the emotional intelligence of dialogue systems, enabling robust performance even during verbally limited or silent phases of interaction.

2 Related Work

Recognizing the emotional states of speakers during dialogue is essential for achieving smooth and adaptive communication. Toward this goal, a wide range of approaches leveraging textual, speech, and facial expression information have been proposed (Shi and Huang, 2023; Ma et al., 2024). Recent studies have explored multimodal fusion to improve emotion recognition performance. For example, Hazarika et al. (2020) proposed MISA, a framework that integrates text, audio, and visual information via contextual inter-modal attention. Zadeh et al. (2018) introduced the CMU-MOSEI dataset along with the Dynamic Fusion Graph for modeling cross-modal interactions. Majumder et al. (2019) developed DialogueRNN, which tracks speaker states and conversational context using multimodal features. However, most of these approaches primarily rely on verbal and visual modalities.

Meanwhile, physiological signals, particularly cardiac-related signals such as electrocardiograms (ECG), have emerged as promising modalities for emotion recognition. Traditional methods extract handcrafted features such as heart rate variability (HRV) and apply classifiers such as SVMs or random forests. More recently, the dynamic nature of physiological data has motivated the application of time-series modeling techniques, including LSTM-based recurrent networks and Transformer architectures, to capture the temporal dependencies and improve prediction accuracy (Katada et al., 2022). In addition, self-supervised learning methods, such as Ts2Vec (Yue et al., 2022) and contrastive learning frameworks (Wu et al., 2023), have been proposed to learn rich physiological representations without extensive labeled data.

Several studies have explored the integration of physiological signals with other modalities to improve emotion recognition performance in human-computer interactions. For instance, Wang et al. (2022) proposed a multimodal framework combining EEG and speech signals, demonstrating that EEG can significantly enhance speech-based emotion classification. Similarly, Katada et al. (2020) examined EDA and visual features to assess users' binary sentiment states in dialogue, achieving an accuracy of 63.2% and emphasizing the improvements gained through physiological signal integration.

Despite these advancements, many prior studies continue to rely heavily on verbal information and often assume access to complete utterances or entire conversational segments. This overlooks the challenges of real-time emotion recognition during ongoing interactions, especially in non-speaking or minimally verbal situations.

To address this limitation, we propose a multimodal recognition framework that integrates physiological signals, speech, and textual information. By predicting emotional valence over short, fixed time intervals, our approach would enable robust emotion recognition, even in conversational environments with limited linguistic information.

3 Dataset

To investigate multimodal recognition of emotional valence, we used an existing Japanese human-human multimodal dialogue dataset that we previously created (Jiang et al., 2024), which contains multimodal information recorded by heterogeneous devices. Table 1 provides an overview of the dataset. The dataset consists of dyadic interactions involving 40 Japanese participants, grouped into 20 pairs, each engaging in dialogues across three topics: "Chit-chat," "Narrative," and "Discussion," yielding a total of 60 dialogue sessions. Each session lasts approximately 10 minutes and records a range of multimodal data, including speech, video, physiological signals, and motion information. Manually annotated transcriptions of the spoken conversations are also provided. The emotional content of the dialogues was annotated by the interlocutors themselves at each moment in the dialogue, using continuous ratings from 0 to 10 (higher values indicate higher emotional valence), collected at a sampling rate of 4 Hz.

In this study, we focus on leveraging physiological signals and verbal information for emotional valence recognition. Among the available modalities in the dataset (text, speech, video, physiological signals, and motion information), we selected physiological data (EDA, BVP, PPG, and pupil

Description	Details
Overview	Japanese Multimodal dialogues between two human interlocutors
Participants	40 individuals forming 20 dyadic groups
Dialogue Duration	10 minutes per dialogue
Number of Dialogues	60 sessions
Annotations	Subjective emotional valence ratings of interlocutor at 4 Hz (Continuous scale of 0 to 10 represents low to high emotional valence)

Table 1: Overview of Japanese multimodal dialogue dataset used in this study.

diameter), speech data, and text data for training the recognition models. A detailed description of the selected data is provided in Table 2. The subjective emotional valence annotations included in the dataset were used as the target labels for the recognition task (see Sec. 3.2 for details).

3.1 Data Preprocessing

Among the data collected from the 40 participants, one participant's pupil data was missing. Therefore, we excluded all multimodal data associated with this participant. To enable emotional valence prediction at any point during the conversation without being restricted to utterance boundaries, we segmented the physiological, speech, and text data, along with the corresponding subjective emotional annotations of the remaining 39 participants, into 15-second intervals based on timestamps (Fig. 1). The 15-second window was selected to strike a balance between capturing local emotional fluctuations and maintaining the stability of multimodal information, particularly physiological signals. It is short enough to reflect momentary changes in emotional state while providing sufficient data within each segment for reliable prediction. Additionally, we removed segments with missing pupil diameter data due to eye blinks, which accounted for less than 1.5% of the entire dataset.

As a result, we obtained a total of 4,831 multimodal samples, each comprising 15-second segments of physiological, speech, text data, and subject emotional valence. In the following sections, we describe the detailed preprocessing procedures for each modality.

3.1.1 Physiological Signals

Following the methodology described in (Jiang et al., 2024), we used the NeuroKit2 toolbox[1] to preprocess and extract features from the EDA, BVP, and PPG signals. Preprocessing involved noise reduction and filtering, after which relevant features were extracted for each signal.

The EDA signals, acquired at a low sampling rate of 4 Hz using EmbracePlus wrist-worn sensors[2] (hereafter, EmbracePlus), underwent no additional filtering. Two key features were extracted: the skin conductance level (SCL), representing the tonic component and reflecting the general activity of sweat glands, and the skin conductance response (SCR), representing the phasic component and indicating rapid changes in skin conductance triggered by specific stimuli, useful for assessing short-term physiological reactions.

BVP was recorded at 64 Hz using EmbracePlus, while PPG was captured at 120 Hz using Shimmer3 GSR+ ear-mounted sensors[3]. Both BVP and PPG signals were processed with a bandpass filter to minimize noise and smooth the signals. Feature extraction involved calculating the Heart Rate (beats per minute), Peak (the maximum amplitude of the BVP or PPG waveform, representing heartbeat intensity), and RRI (R-R interval, the variability between successive heartbeats). These features were obtained by applying a peak detection algorithm to the time-series data.

Pupil diameter measurements were recorded at 13-26 Hz for both eyes using Pupil Core eye trackers[4]. Each measurement was accompanied by a confidence score indicating its quality. We considered data points with a confidence score below 0.6 to be unreliable and treated them as missing values. To standardize the data for subsequent processing, all pupil diameter measurements were resampled to a uniform rate of 4 Hz.

3.1.2 Speech

Although the speech of each speaker was recorded separately using unidirectional microphones (DPA 4088 microphones[5]), the audio recordings still contained background speech from the other interlocu-

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

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

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

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

[5] https://www.hibino-intersound.co.jp/dpa_microphones/5394.html

Modality	Data Type	Device and Description
Physiology	EDA waveform (4 Hz)	Electrodermal activity recorded using EmbracePlus wrist-worn sensors
	BVP waveform (64 Hz)	Blood Volume Pulse recorded using EmbracePlus wrist-worn sensors
	PPG waveform (120 Hz)	Photoplethysmography recorded using Shimmer3 GSR+ sensors
	Pupil diameter (13–26 Hz)	Pupil size of both eyes recorded using Pupil Core eye trackers
Speech	Monologue audio	Speech recordings of individual interlocutors captured with DPA 4088 mics
Text	Text Transcriptions	Manual annotated textual transcriptions of spoken conversations

Table 2: Overview of multimodal data used in this study, including physiological, speech, and textual modalities, along with their respective recording devices and specifications.

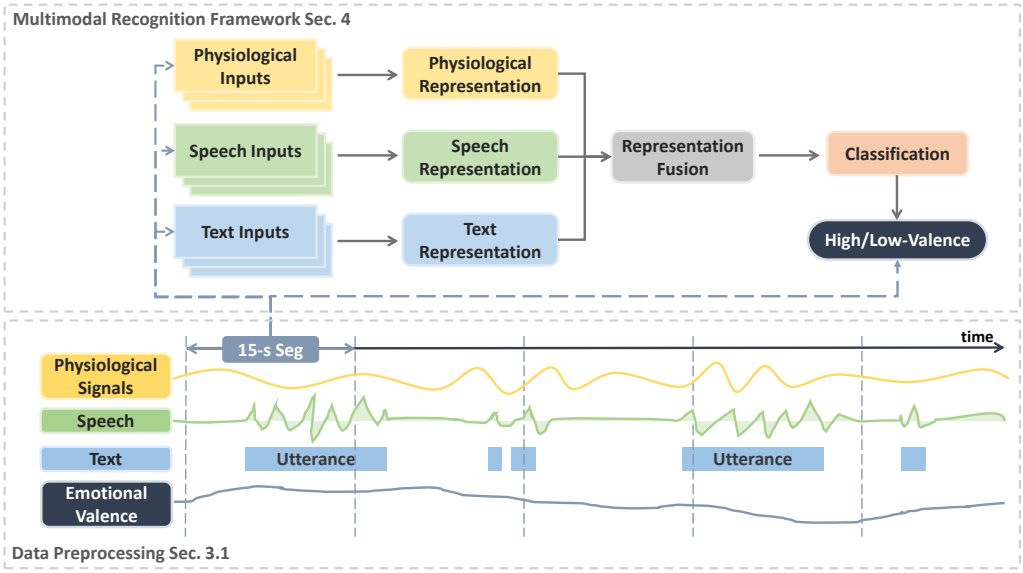

Figure 1: Overview of multimodal recognition framework of emotional valence. Physiological signals, speech, and transcription data are segmented into 15-second intervals during preprocessing (Sec. 3.1). Each modality is independently processed to extract modality-specific representations (Sec. 4.1), which are then fused and input into classification model to predict high or low emotional valence (Sec. 4.2).

tor. Therefore, to suppress the unintended speaker's voice, we applied a low-pass filter with a cutoff frequency of 3,000 Hz to attenuate noise using the pydub[6] library in Python.

3.1.3 Text

To obtain finer-grained temporal information, we generated word-level timestamps from the original utterance-level transcriptions using a forced alignment approach, following the methodology of Pratap et al. (2024). Specifically, we used the Japanese wav2vec2 model[7] to extract high-quality token-level acoustic representations from the audio recordings. By aligning these acoustic features with the utterance-level transcriptions, we predicted the timing boundaries of each word, yielding precise word-level timestamps. On the basis of these timestamps, we extracted the corresponding text content by collecting all words spoken within each 15-second interval. Tokens that spanned the segment boundary were included in the 15-second segment.

3.2 Emotional Valence Labels

Previous studies have indicated an imbalance in the distribution of subjective evaluation annotations, with a predominance of high-valence data (Jiang et al., 2024). To address this bias, ensure stable model learning, and facilitate balanced classification, we formulated the emotion recognition task as a binary classification problem, distinguishing between "high-valence" and "low-valence" emotional states.

To achieve a more balanced distribution of emotional labels, we evaluated classification thresholds of 5 and 6. Specifically, the average subjective valence score was calculated for each 15-second segment; segments with an average score higher than the threshold were labeled as "high-valence,"

[6] https://pydub.com/

[7] https://huggingface.co/reazon-research/japanese-wav2vec2-large-rs35kh

while those with an average score equal to or below the threshold were labeled as "low-valence." This resulted in a class distribution of 3,981 high-valence instances and 850 low-valence instances ($4.7 : 1$) for threshold 5, and 3,170 high-valence instances and 1,661 low-valence instances ($1.9 : 1$) for threshold 6. Since threshold 5 produced highly imbalanced distributions, we selected threshold 6 as the cutoff point.

4 Multimodal Recognition Framework

To recognize emotional valence from both verbal and non-verbal cues, we developed a multimodal recognition framework that integrates text, speech, and physiological data, as illustrated in Fig. 1.

4.1 Multimodal Feature Extraction

Physiological, speech, and text representations were extracted from the 15-second segments processed in Section 3.1. The following details the procedures used to extract modality-specific feature representations.

4.1.1 Physiological Representations

To reduce the influence of individual differences between different interlocutors on the physiological signals, min-max normalization was applied to all physiological data on a speaker-by-speaker basis, scaling the physiological values of each speaker to a consistent range (0 to 1).

For each physiological signal (EDA, BVP, PPG, and pupil diameter), the time-series data were constructed by stacking the extracted features along the feature dimension. Specifically, each sample is represented as:

$$\text{EDA} \in \mathbb{R}^{n_{\text{sample}} \times n_{\text{timestamps}} \times 3}, \quad (1)$$

$$\text{BVP} \in \mathbb{R}^{n_{\text{sample}} \times n_{\text{timestamps}} \times 4}, \quad (2)$$

$$\text{PPG} \in \mathbb{R}^{n_{\text{sample}} \times n_{\text{timestamps}} \times 4}, \quad (3)$$

$$\text{Pupil} \in \mathbb{R}^{n_{\text{sample}} \times n_{\text{timestamps}} \times 2}, \quad (4)$$

where n_{sample} denotes the number of 15-second segments, and $n_{\text{timestamps}} = 15$ seconds $\times$ sampling rate represents the number of timestamps for each signal. The feature dimensions correspond to:

EDA: Raw EDA signal, SCL, SCR;

BVP: Cleaned BVP signal, Heart Rate, Peak, RRI;

PPG: Cleaned PPG signal, Heart Rate, Peak, RRI;

Pupil: Pupil diameter of the right and left eyes.

In this study, we used Ts2Vec (Yue et al., 2022), a general-purpose framework for time-series representation learning, as the physiological encoder. Ts2Vec leverages hierarchical contrastive learning over extended contextual views to capture and vectorize multi-scale temporal dependencies. Each constructed time-series input was fed into Ts2Vec to obtain a fixed-dimensional embedded representation in $\mathbb{R}^{n_{\text{sample}} \times d}$, where d denotes the embedding dimensionality for each physiological signal.

We trained the TS2Vec model on the training data using three hyperparameters: a batch size of 4, an embedding dimensionality of 200, and 15 training epochs.

4.1.2 Speech Representations

We utilized a pre-trained HuBERT model optimized for Japanese (hereafter, Japanese HuBERT) (Sawada et al., 2024) to obtain the speech representations. Japanese HuBERT was pre-trained on a Japanese speech corpus, ReazonSpeech (Yin et al., 2023), by using the self-supervised speech representation learning approach of HuBERT (Hsu et al., 2021). To align with the input requirements of Japanese HuBERT, we downsampled each 15-second audio clip from 44.1 kHz to 16 kHz. The model generated frame-level speech representations for each 50-ms segment of the speech waveform, resulting in embeddings with a shape of $\mathbb{R}^{n_{\text{sample}} \times n_{\text{frame}} \times d}$, where n_{sample} denotes the number of 15-second speech clips, n_{frame} the number of frames per clip, and $d = 768$ the dimensionality of each frame-level embedding.

For subsequent representation fusion, we applied mean pooling over the frame dimension, yielding fixed-size speech representations of shape $\mathbb{R}^{n_{\text{sample}} \times d}$.

4.1.3 Text Representations

We utilized a pre-trained Japanese BERT model[8] to extract text representations. Each 15-second text segment contained a varying number of words depending on the interlocutors' speaking duration. For each segment, the initial text representations were obtained in the form of $\mathbb{R}^{n_{\text{token}} \times d}$, where n_{token} refers to the number of tokens in the text sequence, and $d = 768$ represents the dimensionality of each token embedding.

To generate a fixed-size text representation for each segment, we applied mean pooling across

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

the token dimension, resulting in text representations of shape $\mathbb{R}^{n_{\text{sample}} \times d}$, where n_{sample} indicates the number of text segments.

4.2 Multimodal Representation Fusion

To effectively integrate information from multiple modalities, we used a feature fusion mechanism that assigns learnable weights to modality-specific representations. This enables the model to focus on more informative modalities while down-weighting less relevant ones.

Let $h_1, h_2, \ldots, h_M$ denote the input representations from M modalities, each projected to a shared dimensional space $\mathbb{R}^d$ via modality-specific linear layers. The fusion module learns a set of weights $w_1, w_2, \ldots, w_M$ through a softmax operation applied to trainable parameters.

The final fused representation o is computed as a weighted sum of the modality-specific representations:

$$o = \sum_{m=1}^{M} w_m h_m \tag{5}$$

where w_m represents the normalized weight for modality m.

4.3 Classification

The fused multimodal representation is then fed into a Multi-Layer Perceptron (MLP) classifier to predict the emotional valence (i.e., high-valence or low-valence). The MLP consists of a single hidden layer with 128 units, followed by a ReLU activation and a final output layer for binary classification. The model was trained using the Adam optimizer with a learning rate of 0.001 for 20 epochs. The batch size was set to the full sample size in each training.

To address class imbalance in the training set, we oversampled the minority class by using the Adaptive Synthetic (ADASYN) algorithm (He et al., 2008) before training.

4.4 Evaluation Procedure

We applied the 5-fold cross-validation with an interlocutor-pair open setting. Specifically, the dataset was partitioned such that in each fold, the samples from certain interlocutor pairs were exclusively assigned to the test set, while the samples from the remaining interlocutor pairs were used for training. The interlocutor pairs were randomly assigned to each fold using a fixed random seed of 42. This procedure ensured that the test data of an interlocutor pair was completely excluded from the training data, thereby preventing any potential data contamination. The number of high-valence, low-valence, and total samples across 5 test folds is summarized in Table 3.

Test Fold	High	Low	Total
Fold 1	579	383	962
Fold 2	558	452	1,010
Fold 3	591	309	900
Fold 4	804	180	984
Fold 5	638	337	975
All	**3,170**	**1,661**	**4,831**

Table 3: Sample counts for each test fold, including number of high-valence and low-valence samples across test folds. Total number of segments is shown per fold, as well as aggregated totals across all folds.

For the imbalanced classification task, we used Balanced accuracy (B-Acc) and Macro F1 score (M-F1) evaluation metrics, which are calculated as follows:

$$\text{B-Acc} = \frac{1}{2}\left(\text{TPR} + \text{TNR}\right) \tag{6}$$

$$\text{M-F1} = \frac{1}{2}\left(F1_{\text{high}} + F1_{\text{low}}\right) \tag{7}$$

where TPR, TNR denote the True Positive Rate and True Negative Rate, and $F1_{\text{high}}$, $F1_{\text{low}}$ denote the $F1$ scores for high-valence and low-valence classes. The reported performance was obtained by averaging the B-Acc and M-F1 across the five folds.

5 Experiment

To evaluate the effectiveness of different input modalities in the recognition of emotional valence, we conducted two rounds of classification tasks:

Unimodal and multimodal recognition: We assessed both unimodal and multimodal models. We conducted unimodal experiments individually for the physiological, speech, and text modalities. Subsequently, we evaluated the performance of the multimodal models by combining different pairs of modalities (Text + Speech, Text + Physio, Speech + Physio), and a model incorporating all three modalities (Text + Speech + Physio). For the physiological modality, all physiological signals (EDA, BVP, PPG, and pupil diameter) were used as input.

Ablation study on physiological signal contribution: We conducted an ablation study to further

analyze the contribution of each physiological signal. This was done by taking the best-performing physiological-signal-including model and removing one type of physiological signal at a time from the input.

Additionally, to examine the performance of each model across different speaking duration percentages, we further investigated how recognition performance varies with these percentages and conducted a case study.

5.1 Performance of Unimodal and Multimodal Recognition

Table 4 presents the B-Acc and M-F1 for the unimodal and multimodal recognition models. Among the unimodal models, the Speech model achieved the best performance, with a B-Acc of 0.597 and an M-F1 of 0.578. For the multimodal models, all combinations of two or three modalities outperformed their respective unimodal counterparts, demonstrating the effectiveness of the multimodal recognition framework. In particular, the Speech + Physio model achieved the highest recognition performance, with a B-Acc of 0.626 and an M-F1 of 0.615. The All (Text + Speech + Physio) model achieved the second-best performance, with a B-Acc of 0.615 and an M-F1 of 0.606, suggesting that while integrating additional modalities can be beneficial, the optimal fusion may depend on the complementarity of specific modalities.

To evaluate the effectiveness of different models, we conducted McNemar's statistical tests (Gillick and Cox, 1989), comparing the best-performing model (Speech + Physio) against each of the other six models: Text, Speech, Physio, Text + Speech, Text + Physio, and All. Holm's correction was applied to control for multiple comparisons. The Speech + Physio model significantly outperformed all unimodal models (Text, Speech, and Physio model) as well as the Text + Speech and the Text + Physio models ($p < 0.01$).

5.2 Results of Physiological Signal Ablation Study

We conducted an ablation study by removing one physiological signal input (EDA, BVP, PPG, and pupil diameter) at a time from the best-performing Speech + Physio model. The results are summarized in Table 5. When EDA was removed, the performance dropped most significantly, with decreases of 0.013 in B-Acc and 0.016 in M-F1, indicating that EDA contributes more substantially to emotional valence recognition compared with the other physiological signals. Removing BVP or PPG signals led to moderate declines in both metrics. Interestingly, excluding pupil diameter slightly improved B-Acc to 0.627, although it caused a small decrease in M-F1 (-0.001). This suggests that pupil diameter may not contribute as effectively to recognition performance as the other physiological features in the context of our multimodal recognition framework.

Modality		B-Acc	M-F1	Sig.
Uni	Text	0.542	0.524	**
	Speech	0.597	0.578	**
	Physio	0.552	0.529	**
Multi	Text + Speech	0.596	0.584	**
	Text + Physio	0.559	0.548	**
	Speech + Physio	**0.626**	**0.615**	
	All	<u>0.615</u>	<u>0.606</u>	

Table 4: 5-fold cross-validation results for unimodal and multimodal emotion recognition models. B-Acc: Balanced Accuracy; M-F1: Macro F1 Score. Best and second-best performances are highlighted in bold and underlined, respectively. Models marked with ** exhibit statistically significant performance difference ($p < 0.01$) from best-performing model, as determined by McNemar's test with Holm correction.

5.3 Performance Across Speaking Duration Percentages

To examine how the proportion of spoken content influences recognition performance, we grouped all 15-second samples according to their Speaking Duration Percentage (%), which was computed using the following formula:

$$\text{Speaking Duration Percentage} = \frac{\text{Speaking Time}}{\text{15s}} \times 100 \quad (8)$$

We then evaluated the M-F1 of four models: Speech + Physio, Speech, Text, and Physio, across these percentage intervals. The results are presented in Figure 2.

Across all speaking duration percentages, the Speech + Physio model consistently achieved the highest performance, outperforming all unimodal models. Notably, its performance peaked within the 40-70% speaking duration range, with the Speech-only model ranking second, highlighting the substantial contribution of spoken content to recognition of the emotional valence.

A significant proportion of samples fell within the 0-20% speaking duration range. Recognition

Modality	B-Acc	M-F1	ΔAcc	ΔF1
Speech + Physio (Full)	**0.626**	**0.615**	–	–
w/o EDA	0.613	0.599	**-0.013**	**-0.016**
w/o BVP	0.616	0.605	-0.010	-0.010
w/o PPG	0.615	0.603	-0.011	-0.012
w/o Pupil Diameter	0.627	0.614	+0.001	-0.001

Table 5: 5-fold cross-validation results for ablation study of individual physiological signals. B-Acc: Balanced Accuracy; M-F1: Macro F1 Score. ΔAcc and ΔF1 represent performance decrease in B-Acc and M-F1 compared with full Speech + Physio model. Largest performance drop is highlighted in bold.

performance in this range was generally weaker across all models. Among the unimodal models, the Physio model outperformed both the Speech and Text models when the speaking duration was below 20%. This suggests that physiological signals are especially valuable in low-speech or non-speech segments, serving as a complementary modality to verbal information and enhancing recognition in scenarios where speech is sparse or absent.

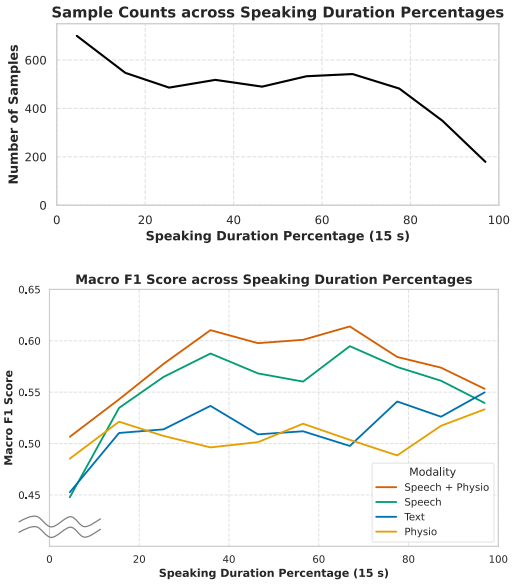

Figure 2: Sample counts (upper) and recognition performance (Macro F1 score) of Speech + Physio, Speech, Text, and Physio models (lower) across 15-second speaking duration percentages.

5.4 Case Study

We further investigated the complementarity between different modalities, focusing on scenarios with limited verbal content. As shown in Table 6, we present three representative examples highlighting the strengths of the physiological modality and the fused Speech + Physio model in such situations.

In the first two examples (Duration: 10% and 11%), the verbal content is minimal or emotionally subtle. There was a lack of clear emotional expressions, leading both the text and speech models to misclassify them as low-valence. However, the physiological signals captured subtle emotional valence patterns that correctly indicated a high-valence state. When fused with speech, the combined model also succeeded, demonstrating the complementary nature of the physiological modality.

The third case (Duration: 4%) consisted almost entirely of non-verbal vocalizations. Here, both the Physio and Speech models correctly predicted the low-valence label, while the text modality failed—likely due to the absence of meaningful lexical input. The Speech model's correct prediction can be attributed to its ability to capture intonation cues.

These cases illustrate the robustness of physiological signals and their synergistic role in multimodal recognition, particularly in emotionally ambiguous or verbally limited contexts.

6 Conclusion and Future Work

In this study, we proposed a multimodal recognition framework for short-time binary emotional valence recognition by integrating physiological signals, speech, and textual information. Leveraging a Japanese multimodal dialogue dataset, we extracted time-series representations of physiological signals using a self-supervised encoder and combined them with features from pre-trained Japanese HuBERT and BERT models. We used a feature fusion mechanism to dynamically integrate modality-specific embeddings. Experimental results demonstrate that incorporating both speech information and physiological signals significantly enhances the recognition performance of emotional valence, especially in scenarios with limited or no verbal

Duration	Transcriptions (15s)	T	S	P	S + P	GT
10%	Almost all of them are mail-order. Ummm... Ah....	low	low	**high**	**high**	**high**
11%	No, there is no such thing. That is true. It is a little **embarrassing**.	low	low	**high**	**high**	**high**
4%	Mm-hmm. Mmm-hmm. Mm-hmm.	high	**low**	**low**	**low**	**low**

Table 6: Example from case study where Text or Speech models failed to classify emotional valence, while Physio and Speech + Physio models succeeded. Examples were translated from original Japanese to English by authors. Duration: Percentage of speaking duration in 15s, T: Text model, S: Speech model, P: Physio model, S + P: Speech + Physio model, GT: Ground Truth.

input. Furthermore, our ablation study reveals that among the physiological signals (EDA, BVP, PPG, and pupil diameter), EDA contributes the most to valence recognition during dialogue.

While our findings highlight the value of physiological signals, the recognition performance remains modest and may not be sufficient for real-world deployment without further enhancement. In addition, splitting dialogue into 15-second fragments may have been disadvantageous for the models relying on speech or text. In future work, we aim to extend our framework beyond 15-second segments toward real-time processing. We also aim to explore regression-based methods for finer-grained emotional valence recognition in order to capture more subtle affective dynamics. Furthermore, we plan to investigate improved integration methods for modalities that are less suited to short segments, such as text, to better leverage contextual information across temporal boundaries. Ultimately, we intend to integrate the proposed framework into dialogue systems, allowing conversational agents to adaptively respond to users' emotional cues in real-world interactive environments.

Acknowledgments

This work was supported by JST Moonshot R&D Grant Number JPMJMS2011. We used the computational resources of the supercomputer "Flow" at the Information Technology Center, Nagoya University.

References

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. In *Proc. EMNLP-IJCNLP*, pages 154–164.

L. Gillick and S.J. Cox. 1989. Some statistical issues in the comparison of speech recognition algorithms. In *Proc. ICASSP*, pages 532–535 vol.1.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proc. ACM–MM*, pages 1122–1131.

Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *Proc. IJCNN*, pages 1322–1328.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3451—-3460.

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

Shun Katada, Shogo Okada, Yuki Hirano, and Kazunori Komatani. 2020. Is She Truly Enjoying the Conversation? Analysis of Physiological Signals toward Adaptive Dialogue Systems. In *Proc. ICMI*, page 315–323.

Shun Katada, Shogo Okada, and Kazunori Komatani. 2022. Transformer-Based Physiological Feature Learning for Multimodal Analysis of Self-Reported Sentiment. In *Proc. ICMI*, page 349–358.

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

Hui Ma, Jian Wang, Hongfei Lin, Bo Zhang, Yijia Zhang, and Bo Xu. 2024. A Transformer-Based Model With Self-Distillation for Multimodal Emotion Recognition in Conversations. *IEEE Transactions on Multimedia*, 26:776–788.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. DialogueRNN: an attentive RNN for emotion detection in conversations. In *Proc. AAAI*, volume 33, pages 6818–6825.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, and 1 others. 2024. Scaling speech technology to

1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

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

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

Tao Shi and Shao-Lun Huang. 2023. MultiEMO: An Attention-Based Correlation-Aware Multimodal Fusion Framework for Emotion Recognition in Conversations. In *Proc. ACL*, pages 14752–14766.

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

Qian Wang, Mou Wang, Yan Yang, and Xiaolei Zhang. 2022. Multi-modal emotion recognition using EEG and speech signals. *Computers in Biology and Medicine*, 149:105907.

Yujin Wu, Mohamed Daoudi, and Ali Amad. 2023. Transformer-Based Self-Supervised Multimodal Representation Learning for Wearable Emotion Recognition. *IEEE Transactions on Affective Computing*, 15(1):157–172.

Yue Yin, Daijiro Mori, and Seiji Fujimoto. 2023. ReazonSpeech: A Free and Massive Corpus for Japanese ASR. In *Proceedings of the Annual Meeting of the Association for Natural Language Processing*, pages 1134–1139.

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

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proc. ACL*, pages 2236–2246.