Estimating User Sentiment at Sub-exchange Granularity from Exchange-level Annotations

Daichi Yukizawa, Kenta Yamamoto, Ryu Takeda, and Kazunori Komatani
SANKEN, The University of Osaka, Japan
E-mail: {daichi-yukizawa@ei., kentayamamoto@, rtakeda@, komatani@}sanken.osaka-u.ac.jp
Tel: +81-6-6879-8416

***Abstract*—It is essential for spoken dialogue systems to estimate user sentiment. Conventional estimation models typically rely on datasets annotated at the exchange-level (i.e., a pair of system and user utterances) and estimate the overall sentiment only at the end of the exchange. This problem setting causes two major problems: (1) the user sentiment cannot be estimated until the exchange is complete, resulting delayed system responses, and (2) it assumes that the user sentiment remains uniform throughout the exchange. To address these problems, we proposes a novel problem setting: sentiment estimation at sub-exchange granularity, which aims to capture sentiment fluctuations within shorter time segments of an exchange. However, the absence of datasets annotated at the sub-exchange-level presents a challenge for supervised learning. To overcome this challenge, we focus on the relationship between exchange-level sentiment labels and the proportions of sentiment at the sub-exchange-level. Based on this relationship, we construct a model that estimates user sentiment at the sub-exchange granularity using only exchange-level labels. Evaluation results demonstrate that the model achieves a certain level of effectiveness in estimating sentiment at sub-exchange granularity. This study serves as a fundamental step toward realizing sentiment estimation at sub-exchange granularity and represents the first stage of future developments.**

I. Introduction

A. Background and Motivation

Recognizing user sentiment during dialogue is crucial for spoken dialogue systems. In such systems, it is essential to flexibly adapt both the content and flow of dialogue in response to the user sentiment. For example, if a user is enjoying the dialogue, the system should continue the current topic; if not, it should switch topics to maintain user engagement and satisfaction.

Previous studies have proposed methods for estimating psychological states such as sentiment and emotion at the exchange-level [1]–[6]. These studies typically train sentiment estimation models based on labels annotated at the exchange-level and estimate the overall sentiment at the end of each exchange. In fact, many publicly available datasets provide sentiment labels annotated for the each exchange [7], [8].

The conventional problem setting causes two major problems. First, because sentiment can only be estimated after an exchange has concluded, conventional estimation models cannot capture dynamic changes in sentiment during the dialogue. This limitation can lead to delays or inappropriate system responses. Second, they assume that sentiment remains uniform throughout the exchange. In reality, user sentiment often fluctuates over time, and the label assigned to the exchange-level may not accurately reflect the sentiment of its individual segments. This issue has also been pointed out in prior work showing discrepancies between sub-exchange emotion and exchange-level labels [2], where exchange-level estimation was conducted while considering temporal changes in emotion.

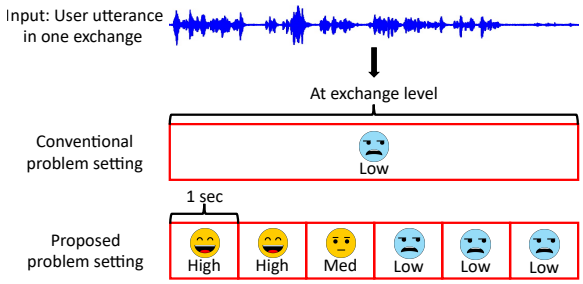

Fig. 1. Estimation Unit Comparison in Sentiment Estimation: Conventional (Exchange-level) vs. Proposed (Sub-exchange-level)

To address these problems, this study proposes a novel problem setting: estimating sentiment at sub-exchange granularity. This enables more flexible and adaptive dialogue strategies—for instance, immediately modifying the system response policy as soon as user sentiment begins to decline. In this paper, we clarify the necessity of sub-exchange sentiment estimation through data analysis.

However, performing sentiment estimation at the sub-exchange-level requires datasets annotated at short time intervals. Currently, most datasets only provide sentiment labels at the exchange-level, and no ground-truth labels are available for short segments at sub-exchange granularity. This makes it difficult to apply conventional supervised learning approaches directly. Therefore, new methods are needed to effectively leverage existing datasets annotated at the exchange-level.

In this study, we construct a method for estimating sentiment at the sub-exchange-level, even under the constraint that only exchange-level sentiment labels are available, as illustrated in Fig. 1. We first analyze how sentiment labels are assigned in the dataset. Then, we investigate the relationship between the exchange-level sentiment label of an exchange and the proportion of sentiment labels at sub-exchange granularity, and construct a model that estimates sub-exchange sentiment based on this relationship.

979-8-3315-7206-8/25/$31.00

The main contributions of this study are as follows:

- We identify the problems of conventional sentiment estimation at the exchange-level and propose a new problem setting that enables estimation at sub-exchange granularity.
- We develop and evaluate a method that enables sub-exchange sentiment estimation even in datasets annotated only at the exchange-level.

B. Related Work

Although there has been extensive research on both sentiment estimation and emotion recognition, comprehensive studies that address these tasks at sub-exchange granularity have not yet been conducted. While recent work has made significant progress in multimodal sentiment analysis—leveraging speech, facial, and linguistic information—most existing studies still treat each exchange as a single unit and do not account for sub-exchange fluctuations in user states.

In both sentiment estimation and emotion recognition tasks, the input modalities are often the same, such as speech and facial information. However, the output differs: sentiment estimation typically represents user states as continuous scores or discrete levels like low, medium, and high [7], while emotion recognition classifies user states into predefined emotional categories such as joy, anger, and sadness [8]. Given this conceptual similarity and shared input data, techniques from emotion recognition are also relevant to sentiment estimation. Since there has been little to no prior work addressing sentiment estimation at sub-exchange granularity, this study refers to existing research on emotion recognition that considers temporal variation at sub-utterance-level.

Some studies have attempted to account for temporal variation at the sub-utterance-level [2], [3], yet these still perform estimation at the utterance-level. Qifei et al. [2] extract short-frame speech features at sub-utterance granularity and use attention mechanisms to emphasize emotionally salient frames. Deeksha et al. [3] propose an emotion recognition method using graph structures, where utterances are nodes and temporal/speaker relationships form edges, enabling the model to capture dynamic emotion changes at sub-utterance granularity. Still, their predictions are made only after the utterance ends.

As described above, previous studies have mainly used information at the exchange-level—such as the entire utterance's speech and text—as input to estimate sentiment or emotion for the whole exchange. These methods leverage integrated information over the exchange-level and achieve reasonable estimation performance. However, they face limitations in capturing subtle fluctuations or changes in sentiment that occur at the sub-exchange-level.

Building on these prior findings, our study differs in its novel problem setting of estimating sentiment at sub-exchange granularity. This enables capturing fine-grained variations in user sentiment at the sub-exchange-level during a dialogue. Additionally, our study uses only speech and visual modalities. This is because speech and visual information contain sentiment-related cues even in short time segments, whereas text at the level of a single word often lacks sufficient information related to sentiment.

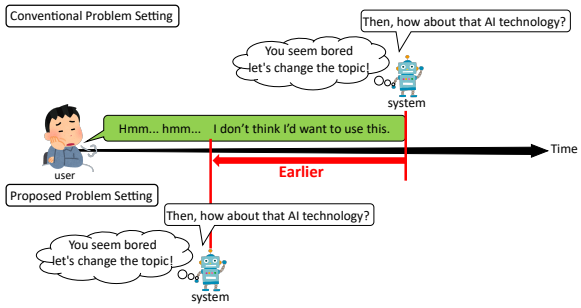

Fig. 2. Comparison of system response timing between conventional exchange-level and proposed sub-exchange-level problem setting

II. Proposed Problem Setting for Sentiment Estimation at Sub-exchange Granularity

This section introduces the novel problem setting of estimating sentiment at sub-exchange granularity. We then annotate the data at the sub-exchange-level and analyze it.

A. Necessity of Sentiment Estimation at Sub-exchange Granularity

1) Advantages of Sentiment Estimation at Sub-exchange Granularity: This study proposes a new problem setting: estimating sentiment at sub-exchange granularity. In conventional problem setting, sentiment is typically estimated at the exchange-level. In contrast, in proposed problem setting, sentiment is continuously estimated throughout the exchange, allowing for the detection of sentiment fluctuations that conventional problem setting may overlook, thereby improving dialogue quality.

Estimating sentiment at sub-exchange granularity can significantly impact how a spoken dialogue system responds. Traditional system estimates sentiment only after user utterance has ended, resulting in delayed responses. This delay can cause the user to continue speaking despite experiencing boredom or disengagement.

The proposed problem setting enables immediate adaptation when user sentiment declines. As illustrated in Fig. 2, conventional problem setting waits until the end of the user utterance to estimate sentiment, whereas proposed problem setting continuously estimates user sentiment throughout the utterance. This allows the system to detect decreases in user sentiment in real time and initiate responses at more appropriate moments. Such adaptability can enhance user satisfaction and contribute to more natural interactions.

2) Challenge and Approach of Sub-exchange Sentiment Estimation: To build a sentiment estimation model at sub-exchange granularity, a dataset annotated with sentiment labels at the sub-exchange-level would ideally be required. However, such datasets are not currently available.

Therefore, in this study, we aim to make effective use of existing datasets in which sentiment labels are assigned at the exchange-level. In these datasets, the exchange-level

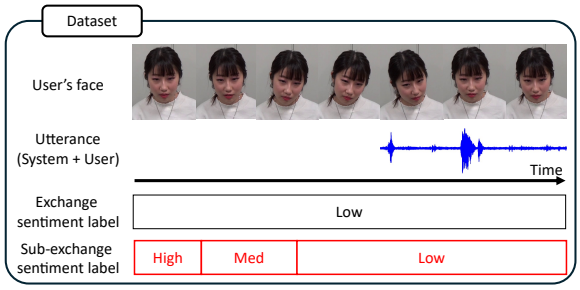

Fig. 3. Sentiment fluctuations in one exchange

sentiment label for an exchange is determined by the most frequently occurring sentiment observed during that exchange. By leveraging this structure, we construct a model capable of estimating sentiment at sub-exchange granularity even in the absence of fine-grained annotations.

Moreover, our proposed model is not restricted to a specific dataset structure, but can be applied to various datasets with exchange-level sentiment annotations, thereby enabling sentiment estimation at sub-exchange granularity more broadly. Although this approach may not be optimal, it represents a practical and feasible solution under current labeling constraints and serves as a step toward enabling finer-grained sentiment estimation.

B. Sub-exchange Annotation and Dataset Analysis

Before building the proposed sentiment estimation model at sub-exchange granularity, we first analyze the characteristics of the dataset used in our experiments. This section describes how sentiment labels were annotated at sub-exchange granularity and presents insights from its analysis, which inform the design of the model.

1) Annotation of Sub-exchange Sentiment Labels: In this study, we utilized Hazumi1911[1] corpus, a multimodal dataset consisting of 26 dialogues between a dialogue agent and human users. The corpus includes video recordings of the users, transcriptions of the utterances, and third-party sentiment annotations on a 7-point scale. Since the system is expected to estimate sentiment from a third-party perspective, we use these third-party annotations in our work. Five annotators watched the user videos and rated the sentiment of each exchange as a whole on a 7-point scale.

In this dataset, sentiment labels are provided on a 7-point scale at the exchange-level, where each exchange is defined as a pair consisting of a system utterance and the corresponding user utterance. Higher scores indicate more positive sentiments (e.g., enjoyment, willingness to continue talking, satisfaction), while lower scores represent more negative sentiments (e.g., boredom, reluctance to continue, dissatisfaction). However, as shown in the top panel of Fig. 3, such exchange-level annotations make it difficult to capture fluctuations in sentiment within an exchange.

We conducted additional annotations specifically for sub-exchange sentiment estimation. Focusing on dialogue videos of a female user in her twenties, the author annotated sentiment labels at sub-exchange granularity, as illustrated in the bottom panel of Fig. 3. Labels were assigned to segments defined between identified points of sentiment change. The evaluation considered both the original Hazumi criteria and additional vocal and facial cues, such as pitch, brightness, and expressions like boredom, lack of emotion, or smiling.

To enable sub-exchange sentiment estimation, the author manually annotated additional sentiment labels on a 7-point scale at sub-exchange granularity. Labels were assigned to segments defined between points where noticeable changes in sentiment occurred. In addition to the original Hazumi criteria, the annotation also took into account vocal and facial cues such as pitch, brightness, and expressions indicating boredom, emotional flatness, or smiling. As shown in the bottom panel of Fig. 3, the annotated sentiment fluctuates at sub-exchange granularity, transitioning between levels such as High, Medium, and Low.

2) Analysis and Application of Sub-exchange Sentiment: Based on the newly annotated data, we analyzed sentiment fluctuations at sub-exchange granularity. The dataset consists of 113 exchanges, among which 56 exhibited observable changes in sentiment during the exchange. This result underscores the necessity of sentiment estimation at the sub-exchange-level and suggests that estimation at the exchange-level may fail to adequately capture dynamic changes in sentiment.

Furthermore, we analyzed the relationship between the exchange-level sentiment label and the proportion of sentiment labels observed during that exchange. Table I shows the internal sentiment proportion corresponding to each exchange-level sentiment label ("low," "medium," and "high"). The results reveal a clear trend: for "low" labels, low internal sentiments dominate; for "high" labels, high internal sentiments are more frequent. These findings suggest a certain degree of correlation, indicating that the exchange-level sentiment label to some extent reflects the proportion of sub-exchange sentiment.

In addition, since sentiment is closely linked to both vocal and facial features, we focused on the correspondence between the exchange-level sentiment label and the distribution of feature representations at sub-exchange granularity. However, the features used in this study are high-dimensional (either 768 or 136 dimensions), and the limited amount of data poses a challenge for building an estimation model. To address this challenge, we applied clustering to compress the feature space by converting high-dimensional features into one-dimensional cluster labels. This approach enables the estimation of sub-

TABLE I
PROPORTION OF INTERNAL SENTIMENT LABELS FOR EACH EXCHANGE-LEVEL SENTIMENT LABEL

		Internal sentiment proportion		
		Low	Med	High
exchange-level sentiment label	Low	0.62	0.38	0.00
	Med	0.46	0.53	0.00
	High	0.08	0.71	0.21

[1]https://github.com/ouktlab/Hazumi1911/

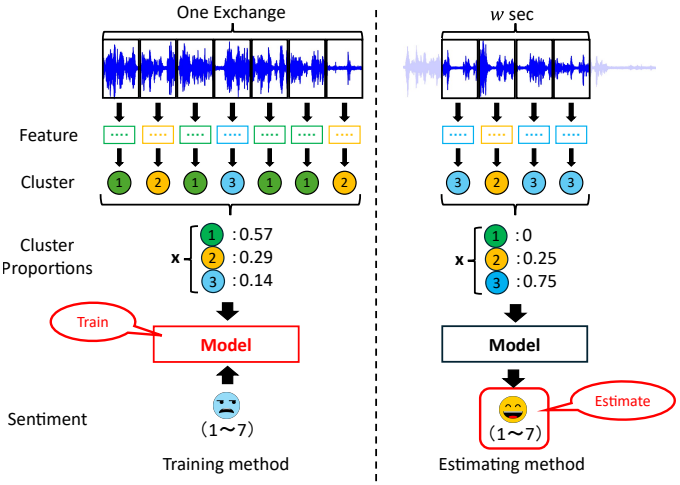

Fig. 4. Training and Estimation Methods of the Estimation Model Used in the Experiment

exchange sentiment from datasets labeled only at the exchange-level.

III. Model for Estimating Sentiment at Sub-exchange Granularity

Based on the procedure outlined in Section II-B2, a model is constructed to estimate sentiments at sub-exchange granularity every second by utilizing the relationship between cluster proportions at sub-exchange granularity and the sentiment label at the exchange-level. The training and estimation methods are described below.

1) Training: As shown on the left side of Fig. 4, the training method involves extracting features from each one-second segment of speech waveforms and images at sub-exchange granularity and then clustering them (denoted as clusters 1, 2, and 3). The model then learns the relationship between cluster proportions at sub-exchange granularity and the sentiment label at the exchange-level. The specific procedure is detailed below.

First, the steps to calculate the cluster proportions at sub-exchange granularity are as follows:

1) Segment the speech and image data of each exchange into one-second intervals. Speech segments correspond to the user's utterances, while image frames are extracted every second during both system and user utterances.
2) Extract speech and image features from each one-second segment as follows: let $\mathbf{x}_t$ denote the speech waveform or image input for the t-th one-second segment, and $\mathcal{F}$ be the feature extraction function. Then, the resulting feature vector $\mathbf{f}_t$ for that segment is given by $\mathbf{f}_t = \mathcal{F}(\mathbf{x}_t)$. An exchange $\mathcal{E}$ consists of a sequence of T such segments, i.e., $\mathcal{E} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$. Accordingly, the feature sequence for the exchange is $\mathcal{F}(\mathcal{E}) = \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_T\}$.
3) Apply k-means clustering to the extracted feature vectors $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_T\}$ to assign each one-second segment to one of c clusters.
4) Compute the proportion of each cluster at sub-exchange granularity using the following equation:

$$P_i = \frac{C_i}{T}, \tag{1}$$

where P_i is the proportion of cluster i in the exchange, C_i is the number of one-second segments assigned to cluster i.

Next, the relationship between cluster proportions P_i at sub-exchange granularity and the exchange-level sentiment score for the exchange is modeled using linear multiple regression to learn regression coefficients β_i and bias terms b, as shown in Eq. (2). Linear regression is employed instead of more complex models such as deep neural networks (DNNs) because complex models require large amounts of data and tend to overfit with limited data, whereas linear regression offers lower model complexity and stable performance in data-scarce settings.

$$y = \sum_{i=1}^{c} P_i \beta_i + b, \tag{2}$$

where y represents the sentiment rating for the exchange-level (a continuous value between 1 and 7), P_i denotes the proportion of the i-th cluster at sub-exchange granularity, β_i is the regression coefficient corresponding to the i-th cluster, b is the bias term.

2) Estimation: The estimation method, as illustrated on the right side of Fig. 4, involves inputting speech waveforms and images at one-second intervals and estimating sentiments using the model. Since cluster proportions over a temporal window are required for estimation, we consider clusters within a sliding window of w seconds. The specific steps are as follows:

1) Input speech waveforms and images at one-second intervals and extract features from each segment.
2) Perform clustering of the extracted features using the k-means model obtained during training.
3) Calculate the proportion of clusters over the past w seconds. If $t < w$, calculate the proportion over the past t seconds instead. This is expressed as

$$\hat{P}_i = \begin{cases} \dfrac{\hat{C}_i}{w} & (t \geq w) \\ \dfrac{\hat{C}_i}{t} & (t < w), \end{cases} \tag{3}$$

where t denotes the elapsed time (in seconds) from the beginning of the current exchange to the present moment, $\hat{C}_i$ is the number of segments classified into the i-th cluster over the past w or t seconds, and $\hat{P}_i$ is the proportion of the i-th cluster among those segments.
4) Estimate sentiment scores at one-second intervals using the cluster proportions over the past window, the learned regression coefficients, and the bias term:

$$y_{\text{pred}} = \sum_{i=1}^{c} \hat{P}_i \beta_i + b, \tag{4}$$

TABLE II
AVERAGE THIRD-PARTY SENTIMENT ANNOTATIONS AT THE EXCHANGE-LEVEL (TRAINING DATA)

Sentiment	Low			Med	High			Total
	1.0–1.5	1.5–2.5	2.5–3.5	3.5–4.5	4.5–5.5	5.5–6.5	6.5–7.0	
Raw counts in 7 bins	0	30	146	840	869	430	25	2340
Counts in 3 classes		176		840		1324		

TABLE III
SENTIMENT ANNOTATIONS BY THE AUTHOR AT THE SUB-EXCHANGE-LEVEL FOR EACH ONE-SECOND INTERVAL (TEST DATA)

Sentiment	Low			Med	High			Total
	1	2	3	4	5	6	7	
Entire exchange	9	45	155	712	136	36	2	1095
		209		712		174		
User's utterance	8	39	128	405	131	36	2	749
		175		405		169		

where y_{pred} denotes the predicted sentiment score (a continuous value between 1 and 7), $\hat{P}_i$ is the proportion of the i-th cluster over the past window, β_i is the learned regression coefficient corresponding to the i-th cluster, b is the bias term, and c is the number of clusters.

IV. EXPRIMENTS

In this section, we evaluate the performance of sentiment estimation using the method. First, Section IV-A describes the dataset, followed by Section IV-B outlining the experimental conditions, including feature extraction models and hyper-parameters. Finally, Section IV-C presents the experimental results of sentiment estimation using the method and discusses the findings.

A. Dataset

We conducted experiments using the dataset introduced in Section II-B1. The training data consisted of 25 dialogues (2,340 exchanges) with sentiment labels assigned at the exchange-level, as summarized in Table II. For testing, one dialogue with sub-exchange sentiment annotations was used, with a total exchange duration of 1,095 seconds and a total user speech duration of 749 seconds. The distributions of the training and testing data are shown in Table III.

B. Experimental Settings

We describe the feature extraction methods and model structures for speech and image modalities, as well as the baseline methods.

1) Speech Settings: Speech features were extracted from user utterances using HuBERT[2] [9] pretrained on the Japanese speech corpus ReasonSpeech[3]. The input was speech waveforms, and the output was 768-dimensional feature vectors. According to Qifei et al. [2], the 9th layer of HuBERT performs best among the 6th, 9th, and 12th layers. To confirm this, features from all three layers were extracted and compared.

[2]https://huggingface.co/rinna/japanese-hubert-large
[3]https://huggingface.co/datasets/reazon-research/reazonspeech

Experiments varied the number of clusters c and the input window length w. The number of clusters was set to 3, 5, 10, and 20. The average duration of user utterances in the training data was approximately 9.6 seconds. Since an excessively large w would approximate exchange-level estimation, input lengths shorter than the average duration were used: 1, 2, 4, 6, and 8 seconds.

2) Image Settings: Image features were extracted during both system and user utterances using OpenFace [10]. The input was images, and the output was 136-dimensional feature vectors comprising 68 facial landmarks' x and y coordinates.

Experiments varied the number of clusters c and the input window length w, consistent with speech settings. The number of clusters was set to 3, 5, 10, and 20. Since the average exchange duration in the training data was approximately 13.2 seconds, input lengths shorter than this average were used: 1, 3, 6, 9, and 12 seconds.

3) Baseline Sentiment Estimation: Two baseline methods were used for comparison purposes. Baseline 1 assumes that sentiment remains constant within sub-exchange granularity, and it is used to verify whether sentiment fluctuations at sub-exchange-level can be estimated. Baseline 2 estimates sentiment at sub-exchange granularity without clustering, directly modeling the relationship between high-dimensional feature distributions and sentiment labels. This baseline serves to evaluate the effectiveness of incorporating clustering in the method. Accuracy was computed based on ground truth labels.

Baseline 2 differs from the proposed method by replacing the cluster proportions P_i in Eq. (2) with feature distributions $\mathbf{f}$, and models the sentiment score as $y = \sum_{i=1}^{d} f_i \beta_i + b$, where y is the exchange-level sentiment score (a continuous value between 1 and 7), f_i is the i-th element of the feature distribution vector $\mathbf{f}$, β_i is the corresponding regression coefficient, b is the bias term, and d is the dimensionality of the feature vector.

During estimation, the feature distributions over the past t or w seconds, denoted as $\hat{\mathbf{f}}$, are used instead of cluster proportions, and the estimated sentiment score is calculated as $y_{\text{pred}} = \sum_{i=1}^{d} \hat{f}_i \beta_i + b$, where y_{pred} is the estimated sentiment score, $\hat{f}_i$ is the i-th element of the feature distribution vector $\hat{\mathbf{f}}$, β_i is the learned regression coefficient corresponding to the i-th feature, b is the bias term, and d is the feature dimensionality.

C. Evaluation and Discussion of Proposed Method

We first examined the optimal combination of the number of clusters c and input window length w for both speech- and image-based sentiment estimation. The highest accuracy for speech, measured by micro accuracy, was achieved with features from the 9th layer of HuBERT, $c = 10$, and $w = 6$,

TABLE IV
ACCURACY OF SENTIMENT ESTIMATION USING SPEECH FEATURES FROM HUBERT'S 9TH LAYER

		w				
	c	1	2	4	6	8
Baseline 1		0.42	0.42	0.42	0.42	0.42
Baseline 2		0.33	0.30	0.33	0.33	0.33
Proposed	3	0.38	0.47	0.48	0.48	0.49
	5	0.41	0.42	0.45	0.46	0.45
	10	0.36	0.47	0.49	***0.50***	0.49
	20	0.31	0.40	0.41	0.42	0.41

w: window size (seconds), c: number of clusters

TABLE V
ACCURACY OF SENTIMENT ESTIMATION USING IMAGE FEATURES

		w				
	c	1	3	6	9	12
Baseline 1		0.37	0.37	0.37	0.37	0.37
Baseline 2		0.29	0.25	0.25	0.25	0.25
Proposed	3	0.50	0.50	0.49	0.50	0.50
	5	0.50	0.50	0.50	0.50	0.50
	10	0.52	0.56	0.56	***0.57***	0.57
	20	0.51	0.52	0.54	0.55	0.55

w: window size (seconds), c: number of clusters

yielding an accuracy of 0.50 (Table IV). For image-based estimation, the highest accuracy was 0.57 at $c = 10$ and $w = 9$ (Table V).

Next, we compare the highest accuracies achieved by each method. Comparing the highest accuracies, the method improved speech-based estimation accuracy by 0.08 over Baseline 1 (0.42) and 0.17 over Baseline 2 (0.33). For images, improvements were 0.20 over Baseline 1 (0.37) and 0.28 over Baseline 2 (0.29). These results indicate that the proposed method better captures sentiment variations at sub-exchange granularity. Additionally, clustering proves effective compared to directly using high-dimensional features for sentiment estimation.

However, it should be noted that the test data consisted of recordings from a single speaker, and thus inter-speaker variability was not taken into account. The distributions of true and estimated sentiment ratings at the point of highest accuracy are as follows:

Speech: true distribution (low, medium, high) = (175, 405, 169); predicted distribution = (0, 375, 374)

Images: true distribution (low, medium, high) = (209, 712, 174); predicted distribution = (1, 924, 170)

It can be observed that "low" sentiment ratings were rarely estimated in both the speech- and image-based models. Furthermore, based on the observed data, the lower accuracy at the early stages of the exchange is likely due to the limited amount of information available.

V. CONCLUSION

In this paper, (1) we clarified the problems of the conventional problem setting that performs estimation at the exchange-level, and proposed the necessity of a new problem setting focusing on estimation at sub-exchange granularity. (2) Moreover, we developed a method to estimate sentiment at sub-exchange granularity by utilizing a dataset labeled with sentiment at the exchange-level. As a result, we demonstrated that it is possible to construct a model and estimate sentiment at sub-exchange-level even in the absence of datasets labeled at that finer granularity.

Future work includes incorporating temporal information into the model and fine-tuning the feature extraction model to further improve estimation accuracy. Although this study used a one-second unit as the basis for estimation, we will explore alternative unit lengths to assess their impact on performance.

ACKNOWLEDGMENT

This work was supported by JST Moonshot R&D Grant Number JPMJMS2011 and JSPS KAKENHI Grant Number JP22H00536.

REFERENCES

[1] S. Katada, S. Okada, and K. Komatani, "Transformer-based physiological feature learning for multimodal analysis of self-reported sentiment," in *ICMI*, 2022, pp. 349–358.

[2] Q. Li, Y. Gao, C. Wang, Y. Deng, J. Xue, and Y. Han, "Frame-level emotional state alignment method for speech emotion recognition," in *ICASSP*, 2024, pp. 11 486–11 490.

[3] D. Chandola, E. Altarawneh, M. Jenkin, and M. Papagelis, "SERC-GCN: Speech emotion recognition in conversation using graph convolutional networks," in *ICASSP*, 2024, pp. 76–80.

[4] Y. Wang, D. Li, and J. Shen, "Inter-modality and intra-sample alignment for multi-modal emotion recognition," in *ICASSP*, 2024, pp. 8301–8305.

[5] J. Liu, S. Chen, L. Wang, *et al.*, "Multimodal emotion recognition with capsule graph convolutional based representation fusion," in *ICASSP*, 2021, pp. 6339–6343.

[6] P. Kumar, V. Khokher, Y. Gupta, and B. Raman, "Hybrid fusion based approach for multimodal emotion recognition with insufficient labeled data," in *ICIP*, 2021, pp. 314–318.

[7] K. Komatani and S. Okada, "Multimodal human-agent dialogue corpus with annotations at utterance and dialogue levels," in *ACII*, 2021, pp. 1–8.

[8] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, and S. Kim, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[9] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[10] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *WACV*, 2016, pp. 1–10.