Multimodal User Satisfaction Recognition for Non-task Oriented Dialogue Systems

Wenqing Wei, Sixia Li, Shogo Okada
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, Japan
wqwei@jaist.ac.jp,lisixia@jaist.ac.jp,okada-s@jaist.ac.jp

Kazunori Komatani
The Institute of Scientific and Industrial Research
Osaka University
Ibaraki, Osaka, Japan
komatani@sanken.osaka-u.ac.jp

ABSTRACT

Multimodal dialogue systems (MDSs) are needed to allow users to converse with virtual agents that use natural language by sensing the multimodal behavior of users. One crucial step in the development of an MDS is measuring how well the dialogue system performs. Though previous research focused on the user satisfaction modeling from linguistic modality in text-to-text dialogue systems, the user satisfaction is observed by not only spoken dialogue contents but also the acoustic and visual nonverbal behaviors of users. Multimodal social signal sensing provides a solution that automatically measures dialogue systems based on subjective evaluation. With this background, we proposed a multimodal recognition model of the user using sequence modeling algorithms (RNN, LSTM, and GRU). It is a novel challenge to recognize the user satisfaction label at the dialogue level. Each label was annotated by the user based on the overall dialogue. We extracted both verbal features and nonverbal features at the exchange level (the unit is a pair of system and user utterances) and analyzed the contributions of multimodal features and unimodal features to recognize user satisfaction labels at the dialogue level. We used a multimodal user-system dialogue data corpus with user satisfaction labels at the dialogue level. To validate the recognition accuracy of the proposed multimodal modeling approach, we compared the proposed method with two models based on human perception by external human coders and the system operator (called "Wizard") with whom the user talks. The experimental results showed that the multimodal model achieved a better performance in both classification and regression tasks. The results indicated that the performance of the multimodal model was higher than that of the human models.

CCS CONCEPTS

• **Human-centered computing** → **HCI theory, concepts and models**; • **Computing methodologies** → *Discourse, dialogue and pragmatics.*

ICMI '21, October 18–22, 2021, Montréal, QC, Canada

ACM ISBN 978-1-4503-8481-0/21/10…$15.00
https://doi.org/10.1145/3462244.3479928

KEYWORDS

Multimodal interaction, User satisfaction, Spoken dialogue system, Recurrent Neural Networks

ACM Reference Format:
Wenqing Wei, Sixia Li, Shogo Okada and Kazunori Komatani. 2021. Multimodal User Satisfaction Recognition for Non-task Oriented Dialogue Systems . In *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3462244.3479928

1 INTRODUCTION

With the development of natural language processing and speech recognition, spoken dialogue systems, such as those of Amazon Alexa, SIRI, and Google Assistant, are used in many fields. There is great interest in developing non-task oriented dialogue systems such as chatbots and open-domain dialogue systems [12][22]. While improving the quality of the non-task oriented dialogue system is important for the user dialogue experience, it is not easy to evaluate how well the system works; therefore, an automatic evaluation of whether a user could satisfy through the dialogue experience is crucial for developing and improving dialogue systems. Two unexplored problems exist in the current satisfaction recognition models. First, almost all previous research [8, 24] has focused on user satisfaction modeling in text-to-text dialogue systems rather than multimodal systems. On evaluating user satisfaction in multimodal dialogue systems such as embodied conversational agents (ECAs) and social robots, the satisfaction level is observed from spoken dialogue contents and the acoustic and visual nonverbal behaviors of users. Second, most previous work [6, 8, 23] recognized satisfaction label at the turn level (per utterance or exchange) to ensure natural interactions. However, a user essentially feels satisfaction throughout the whole conversation, thus, the system designer needs to analyze not only satisfaction at turn level, but also the overall satisfaction concerning the whole conversation. We define overall satisfaction as the "dialogue-level satisfaction".

This study presents a multimodal model to recognize user satisfaction at the dialogue level by using multimodal features observed from users, which is suitable for evaluating non-task oriented dialogue systems. The multimodal features extracted from each exchange (a pair of system and user utterances) are input to each unit of the sequence models (RNN, LSTM, and GRU). The output is set as the dialogue-level satisfaction annotated by the user who has talked with the system. We utilized a novel multimodal dialogue data corpus to construct these sequence models, including dialogue-level (overall) satisfaction labels, exchange-level sentiment annotation, and multimodal data including spoken dialogue transcription, audio

signals, face images, and body motion data. We used five feature sets to recognize user satisfaction. Meanwhile, this study analyzes the contributions of different features to user satisfaction.

To validate the proposed methods based on the machine learning (ML models), we compared the performance of the proposed model with two types of human methods. The first method is a sequence model that recognizes dialogue-level user satisfaction from exchange-level impression annotations (**Human model (1)**). The second method, the system operator (called "Wizard") directly recognizes the dialogue-level user satisfaction (**Human model (2)**). The main contributions of this study can be summarized in the following three aspects.

Multimodal user satisfaction recognition: This task is unexplored, and it is a new challenge in the multimodal human-agent interaction domain. We proposed a multimodal approach utilizing sequence modeling algorithms to recognize user satisfaction at the dialogue level in multimodal interactions. In this study, we combined audio, visual, and text features to recognize user satisfaction. We demonstrate that multimodal features performed better than unimodal features in Section 6.1.

Comparison between the contribution of multimodal features and exchange-level annotation: Many studies have focused on proposing multimodal models for recognizing the exchange-level sentiment label, how exchange-level label is correlated with the dialogue-level satisfaction is still unclear. We first explored the relationship between exchange-level and dialogue-level annotations. Then, this study used exchange-level annotation scores as manual features to recognize user satisfaction and compared the results with those obtained multimodal (automatically obtained) features. The comparison between the two feature types is described in Section 6.2.

Comparison between the ML models and human model: To validate the effectiveness of the multimodal ML models, we compared the recognition result of the ML models with the user satisfaction score annotated by a system operator (Wizard). The comparative analysis in Section 6.3 demonstrates the challenging nature of the task and the contribution of the automatic multimodal recognition technique on estimating user satisfaction.

2 RELATED WORK

Intelligent conversational agents have become widely used in daily life. Measuring the performance quality of a dialogue system is a critical component during the development process. Initially, some researchers used the dialogue efficiency and dialogue costs, which are related to the length of the dialogue or task success, to measure the performance [13]. However, there is no task success information in non-task oriented conversations (such as small talk and multi-domain dialogue) when interacting with simulated or recruited users.

To develop an appropriate and correct system, recent studies have focused on user-centered criteria that are defined based on human judgments to approximate the usability of dialogue systems. An annotated score, such as "user satisfaction", is recognized by using machine learning techniques. For example, Engelbrecht et al. [6] used dialogue actions as input features to recognize user satisfaction at the exchange-level. Higashinaka et al. [9] used annotations by experts who observed the dialogue as target variables to model the user satisfaction. Since most input features are annotated manually, this method is inconvenient and inefficient for online applications. Schmitt and Ultes [23] used dialog manager-related parameters, the semantic meanings of which were extracted automatically as input features, to recognize the median rating of several expert ratings at the exchange-level.

In terms of experimental methods, some researchers have regarded the user satisfaction recognition task as a sequence problem. Hara et al. proposed an N-gram model trained using sequences consisting of dialog acts to recognize user satisfaction [8]. A hidden Markov model (HMM) was also used to model user satisfaction transitions in dialogues [10]. However, the experiment has shown that Support Vector Machines methods that did not use sequence information were performed better than HMMs [23]. User satisfaction recognition is a temporal task that should benefit from time-series dialogue data. To investigate the effect of temporal information, Ultes et al. extended the set of temporal features to different levels, and the results showed that interaction parameters (e.g., ASR performance) at the window and dialogue levels that provide temporal information have major effects on interaction quality [25]. Recently, deep learning techniques have also been applied for user satisfaction recognition tasks. Ultes et al. [24] proposed a recurrent neural network (RNN) to achieve improved recognition accuracy. To eliminate the heavy reliance on handcrafted temporal features, they presented a deep learning-based Interaction Quality (IQ) estimation model that utilizes recurrent neural networks' capabilities to automatically learn temporal information.

Concerning features, all previous researches focused on using linguistic features and dialogue content to recognize user satisfaction with a text-based dialogue system. In a multimodal dialogue system, it is well known that sensing multimodal information from the user is useful in recognizing the inner states of the users, such as the sentiment level. Recently, many studies have proposed that multimodal information including visual and acoustic features improves sentiment recognition accuracy. The temporally selective attention model [26], multi-attention recurrent network [29], memory fusion network [28], and tensor fusion network [27] were proposed for multimodal sentiment analysis. Hirano et al. proposed a multitask deep learning neural network model (MT-DNN) using multimodal features to recognize the user exchange-level sentiment toward spoken dialogue systems [11].

The main difference between this work and previous works is summarized as follows. Though most of the previous studies have focused on recognizing user satisfaction at the exchange level, the main work focus of this study recognizes user satisfaction at the dialogue level to evaluate the non-task oriented dialogue systems. We also present a multimodal modeling method based on the user's performance in the conversation.

3 DATA AND ANNOTATIONS

Figure 1 shows the overview of this research. In this section, we describe the multimodal dialogue dataset and annotations to it.

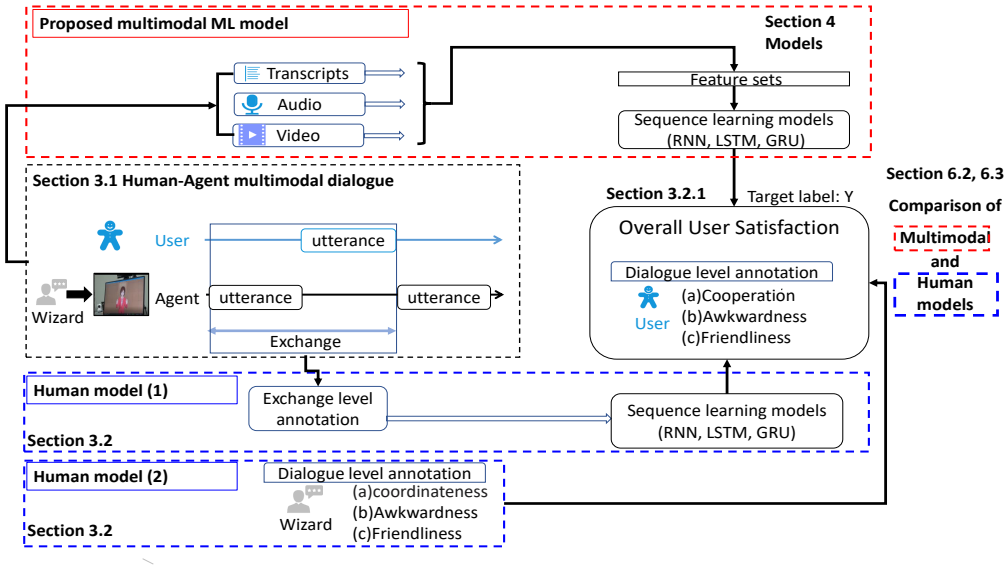

Figure 1: Overview of the estimation of the user's satisfaction at the dialogue level

3.1 Data recording

This study was conducted on two shared multimodal dialogue datasets [1] named Hazumi1902 and Hazumi1911, in which recording settings were almost the same [18]. Both corpora were arranged to record facial videos, audio data, and upper body data and used a virtual agent called MMD-Agent as the interface to communicate with participants who were manipulated by an operator (Wizard) in another room. In this system, the operator could select a topic, utterances on the topic, and general responses used in conversation. To shorten the time interval before the machine responded, the operator was well trained and had time to select the next utterance while the participant was speaking (approximately 10 seconds).

Regarding acoustic signals and body posture, audio and posture of the upper body were recorded by a Kinect sensor. The posture information was recorded at 30 fps. Each participant's voice was recorded as a 16 kHz WAV file. The number of participants in the two corpora was 60, which included 25 males and 35 females. The participants' ages ranged from 20 to 70 years, and they were recruited from the public through a recruitment agency.

3.2 Annotations

The data corpus included two kinds of annotations. One was annotation at the dialogue level, and the other was annotation at the exchange-level. First, the dialogue-level annotations were used as the target labels in this study to develop the recognition model of user satisfaction. Second, the exchange-level annotations, including the user sentiment, indicate the user's perceptions of the system; therefore these annotations were used as partial information for understanding the dialogue-level satisfaction.

3.2.1 Dialogue-level annotations. For dialogue-level annotations, it is difficult to define user satisfaction based on one criterion. For this reason, this study used a questionnaire with 18 labels relating to the user's impression of the dialogue proposed in [2].[2] The questionnaire measured interpersonal communication cognition as a social skill. The 18 items were "well-coordinated", "boring", "cooperative", "harmonious", "unsatisfying", "uncomfortably paced", "cold", "awkward", "engrossing", "unfocused", "involving", "intense", "unfriendly", "active", "positive", "dull", "worthwhile", and "slow". Kimura et al. [17] analyzed the rapport in dyadic interactions (60 pairs, with 120 subjects) and reported that three labels ("well-coordinated", "awkward" and "unfriendly") were carefully extracted as representative labels from the 18 labels by conducting a factor analysis. Based on this finding, we defined the scores of the three labels as the user satisfaction level, and used these three values as grand-truth values for machine learning. We use the three labels ("well-coordinated", "awkward", and "unfriendly") as "coordinateness", "awkwardness" and "friendliness" in this study. Each label was evaluated on an eight-point scale from 1 to 8.

We asked the participants to annotate all 18 labels, but we asked the Wizard to annotate only the three labels ("well-coordinated", "awkward" and "unfriendly") to reduce the burden on the Wizard in annotating the rapport scores of all 60 participants after the dialogue[3].

3.2.2 Exchange-level annotations. The exchange was defined as the section beginnings from the start time of a system utterance and endings at the start time of the next system utterance. Exchange-level annotations were collected to analyze the user's internal state per in each exchange unit. Hirano et al. [11] and Katada et al. [14] presented multimodal models to recognize the exchange-level annotations. Three types of annotations were given at the exchange-level as follows:

[1] The doi is doi/10.32130/rdata.4.1

[2] We used the Japanese version [17] translated from the original questionnaire.

[3] Though we also conducted similar questionnaires before each dialogue, we did not use them

Table 1: Usage of different annotations

Unit	Type	Annotator	Usage
Exchange-level annotations	Topic continuance Third sentiment	Third party coders	Input features for Human model (1)
	Self sentiment	Dialogue users	
Dialogue-level annotations	Coordinateness Awkwardness Friendliness	(i): Dialogue users (ii): Wizard	(i): Target label (ii): Using as a human model to recognize (i) (Human model (2))

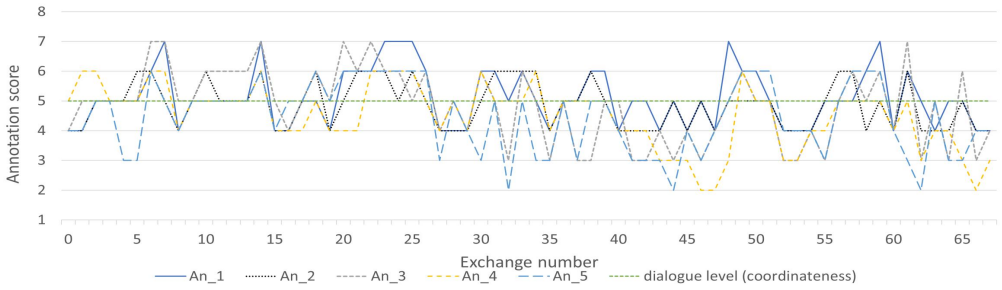

Figure 2: Example of annotation in a conversation. The An_1 to An_5 denote the topic continuance level annotated per each exchange by the five annotators.

Topic continuance: The topic continuance label was a degree indicating whether the topic should be changed. Five human coders assigned such labels depending on whether the system should have continued the current topic or changed the topic in the next system's utterance. The labels of the scores ranged from "strongly change the topic" 1 to "strongly continue the topic" 7, as shown in Figure 2.

External sentiment: When a participant communicated with the dialogue system, the participant had different sentiments during each turn. Human coders annotated the external sentiment level per exchange with scores ranging from 1 (the participants seemed bored with the dialogue) to 7 (participants seemed to enjoy the dialogue) while watching recorded videos of the dialogues.

Self-sentiment: This annotation was similar to the external sentiment annotation. Self-sentiment labels were assigned as scores ranging from 1 to 7, which were divided into two categories. Positive sentiments included "enjoy talking" and "satisfied with the talk", and negative sentiments included "want to stop talking" and "confused about the system utterances".

Based on these definitions, in total, 5373 exchanges obtained from 60 participants were annotated. The agreement scores of the annotators measured by Cronbach's alpha were 0.83 for the topic continuance and 0.86 for the external sentiment.

3.2.3 Usage of difference annotation. As described in Sections 3.2.1 and 3.2.2, two types of annotations were used in this study. As shown in Table 1, three types of exchange-level labels were used as input features to recognize the user satisfaction on the dialogue level. The details of the experiments are presented in Section 5.2. For dialogue-level annotation, both the user and Wizard annotated the user satisfaction labels at dialogue level after the conversation.

Table 2: Pearson correlation coefficient between exchange annotations and the dialogue-level annotations

	Topic continuance	External sentiment	Self-sentiment
Coordinateness	0.17	0.24	0.30
Awkwardness	−0.16	−0.18	−0.36
Friendliness	0.07	0.05	0.29

In this study, user annotations at the dialogue level were used as target labels. We used the Wizard's annotation to evaluate the user satisfaction as a "human" model based on Wizard's subjectivity. The results of estimations by the Wizard and models trained with multimodal features facilitated the comparison of the performances of humans and ML models.

3.3 Relation between exchange-level annotations and dialogue-level annotation

To explore the relationship between exchange labels and dialogue labels, we used the Pearson correlation coefficient to calculates the correlations between dialogue-level annotation and the average value of all exchange-level annotations in one dialogue. Generally, it belongs to weak correlation when the correlation coefficient is higher than 0.1; and if the correlation is higher than 0.3, it is a moderate correlation. Table 2 shows the correlation matrix between the dialogue level of user satisfaction after dialogue annotations and the average value of all exchange-level annotations. Compared with the correlation between the third-party (topic continuance and external sentiment) annotations on the exchange and dialogue-level annotations, the self-sentiment annotations on the exchange had a

higher correlation with dialogue-level annotations. We also found that all exchange annotations were positively correlated with the coordinateness and friendliness labels and negatively correlated with the awkwardness labels. The self-sentiment annotation had the highest correlation with the dialogue-level annotation.

However, we found that the correlation between exchange-level labels and dialogue-level labels was not strong, indicating that exchange-level annotation cannot accurately express user satisfaction at the dialogue level. For this reason, it is necessary to recognize user satisfaction at the dialogue level directly. In this study, the exchange-level annotation feature was used as a manual feature to identify the user's satisfaction at the dialogue level. We analyzed the exchange-level annotation feature results and compared them with the multimodal results in Section 6.2.

4 MULTIMODAL USER SATISFACTION MODELING

4.1 Multimodal feature extraction

4.1.1 Audio feature. This study extracted acoustic features at the exchange level as the emotional information in speech by using the speech feature extractor OPENSMILE [7]. The acoustic features corresponded to the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), which achieves high performance in emotion-related fields. These features were extracted for each speaker turn and normalized by each speaker after extraction. Finally, we obtain a 88 dimensions vector.

4.1.2 Linguistic feature. We extracted two types of linguistic features from the manual transcription of spoken dialogue contents:
Part of speech: The sentences were segmented into words and annotated with universal part-of-speech (POS) tags using Stanza NLP [4]. The PoS tag set was composed of 17 types: "adjective", "adposition", "adverb", "auxiliary", "coordinating conjunction", "determine", "interjection", "noun", "numeral", "particle", "pronoun", "proper noun", "punctuation", "subordinating conjunction", "symbol", "verb", "other". The PoS categories (nouns, verbs, etc.) in a user's utterance were counted.
BERT (Bidirectional Encoder Representations from Transformers [5]): In this study, we used a model pre-trained on only Japanese text (using Wikipedia) [16]. We used this model to extract features from the text at the exchange level, and finally, we obtained a 768-dimensional text representation vector.

4.1.3 Visual feature. We extracted body activity and facial features as visual features using an RGB camera and Kinect V2 with a depth sensor.
Body activity features: This study used three-dimensional coordinate data for each joint of the upper body, which was estimated from a Microsoft Kinect v2, to extract motion features. We used five points of body motion, which included the left shoulder, right shoulder, left hand, right hand, and head. We denoted the three-dimensional coordinate data of each body point at tth-frame as $w(t) = x, y, z$. We calculated the absolute value of velocity between two frames as $|v(t)| = |w(t+1) - w(t)|$ and calculated the absolute value of acceleration between frames as $|a(t)| = |v(t) - v(t-1)|$. After $v(t)$ and $a(t)$ were calculated, we used the maximum value of acceleration, and the maximum, mean, and standard deviation of velocity in the user turn as body activity features. Finally, the body activity feature set included 20 dimensions in total.
Facial landmark feature: OpenFace [1] software output the three dimensional coordinates of 68 facial landmarks in each frame. This study chose ten facial landmarks, including 2 points on each eye, 4 points around the mouth, and two on the eyebrows. We adopted the same method used for body feature tracking. We extracted the maximum acceleration value and the maximum, mean, and standard deviation of the velocity for each user exchange turn as facial features. Finally, we obtained a 40-dimensional vector.
Action units: Facial expressions display emotional states, which help regulate turn-taking during the conversation. This is often represented using facial action units (AUs), which objectively describe facial muscle activations [4]. In this study, we used OpenFace to obtain 18 types of AUs which were rated between 0 and 1, indicating absence and presence, respectively. Then we calculated the average of each AU in exchange for facial AU features (18 dim). Overall, 58 dimensions of facial features were used in this study.

4.2 Models

To recognize the user satisfaction at dialogue level, a machine learning model needs to capture dynamical change in multimodal behaviors while the user is talking with the system. To model the sequence of multimodal behaviors, we utilized the following three sequence models; Recurrent neural network (RNN), long short term memory (LSTM) and gated recurrent unit (GRU) models.

The multimodal features extracted from each exchange were input to each unit of the RNN, LSTM, and GRU in the proposed multimodal model. This study used the early fusion method and the unimodal features (audio a_t: 88 dim., linguistic l_t: 785 dim., and visual v_t: 78 dim.) extracted from the t-th exchange were concatenated into one vector x_t (951 dim.). The input of these recurrent neural network models was x_t $(1 \leq t \leq T)$. In all models, a two-recurrent (hidden) layers with 128 units ($h_t^{(1)}$, $h_t^{(2)}$) are used to extract the features from the sequence input vector x_t with T exchanges. We obtained two final hidden states $h_T^{(1)}$, $h_T^{(2)}$ (2 (layers)*128 (units)) from the recurrent layers. A fully connected layer followed the recurrent layer to project the output (2*128) from the recurrent layers into the output layer. For a classification task, the output layer containing two units and the log-SoftMax function was used to output the probabilities of the different user satisfaction S_c. For a regression task, the sigmoid unit was used to output the estimated value S_r of the user satisfaction level (1-8).

5 EXPERIMENTS

The purpose of the experiments was to recognize user satisfaction at the dialogue level. We evaluated the user satisfaction recognition accuracy through both classification and regression tasks. Three research questions were addressed, each of which corresponds to a subsection in Section 6.
RQ1: Do multimodal features contribute to improving user satisfaction recognition?
RQ2: Which is more effective in user satisfaction recognition multimodal features or exchange-level annotation features?

[4]https://github.com/stanfordnlp/stanza

RQ3: Compared with human subjective perception, how does the recognition of multimodal models perform?

5.1 Experimental setting

5.1.1 Regression task setting. The regression tasks aimed to fit the labels of dialogue base on different feature sets. The mean squared error (MSE) was calculated using the square of the difference between the actual and estimated values, which were then summed and averaged. It was convenient to take the squared derivative of the results. In this work, we used the MSE as the loss function for all regression tasks.

5.1.2 Classification task setting. The binary classification datasets were developed as follows. All dialogue-level label annotated scores (1-8) were converted into binary values (high and low) with a threshold of 4 (neutral state). The numbers of high/low data points for the three target labels at dialogue level were 38/22 for the coordinateness label, 32/28 for the awkwardness label, and 49/11 for the friendliness label, respectively. We used the F1-score as a metric to evaluate the accuracy of imbalanced datasets in which the number of samples was different between the two classes.

5.1.3 Hyperparameter setting and evaluation. To evaluate the comparative models under equivalent conditions, we used the same parameters in all models. We used the Adam optimizer, set the learning rate to 0.001, and set the total number of epochs to 30. Five-fold cross-validation was conducted, and their average F1-score is reported.

5.1.4 Combination of multimodal futures. According to the findings in previous works, linguistic features were the key descriptors in recognizing user satisfaction. For this reason, we set the unimodal model with a linguistic feature set as the baseline model. In addition to the baseline model, we prepare four combinations of unimodal features (audio, visual and linguistic) to analyze the effectiveness of the verbal-nonverbal multimodal models and nonverbal multimodal models (without linguistic features).
(1) L: model trained with Linguistic features (baseline)
(2) A+V: model trained with Acoustic + Visual features
(3) A+L: model trained with Acoustic + Linguistic features
(4) V+L: model trained with Visual + Linguistic features
(5) ALL: model trained with Acoustic + Visual + Linguistic features

5.2 Comparative methods

We prepared two human models as comparative methods with the proposed multimodal models.

5.2.1 Human model (1) using exchange-level annotations. In this group experiment, five experts annotated two labels (external sentiment and topic continuance) in each exchange turn with scores ranging from 1 (the participants seemed bored with the dialogue) to 7. For the external sentiment and topic continuance labels, the averaged annotated scores were calculated per the t-th exchange and then combined with the self-sentiment annotation at t-th exchange as manually annotated features total three dimensions a_t $(1 \leq t \leq T)$. a_t is input features for the sequence models. The network architecture is the same as the multimodal models described in Section 4.2.

5.2.2 Human model (2) using subjective evaluation of Wizard. In the second group experiment, the Wizard's annotation was regarded as the result of human recognition. For the classification methods, similar to the annotation procedure described in Section 5.1, the Wizard's annotation results were divided into high and low satisfaction categories, then F1-score was calculated. For the regression, we computed the MSE of the Wizard's annotations and user annotations.

6 RESULTS

Tables 3 and 4 show the regression and binary classification results of satisfaction label recognition by the RNN, LSTM, and GRU, respectively, based on the five feature sets. In this section, first, we compared both the regression performance: MSE and classification F1-score of these models. Second, we compare the results of the proposed multimodal models with that of the model trained with the exchange-level annotation features. Finally, we compare the performance of the multimodal models to that of the "human model" based on subjective evaluation by the Wizard.

6.1 Comparison between unimodal and multimodal features (RQ1):

Columns 3 to 7 in Table 3 show the regression results: the MSE values of user satisfaction labels generated by training with different multimodal feature sets.

We observed that the multimodal models yielded the best performance for the coordinateness and awkwardness labels among all models. The ALL feature set (A+V+L) and L+V feature set produced the lowest MSE values (2.93 and 4.00) in the LSTM method for the coordinateness and awkwardness labels. Most of the feature fusion models (A+L and A+V) performed better in terms of MSE than those using linguistic features. The unimodal set (L) achieved the best result for the friendliness label, with an MSE of 2.87. For the awkwardness label, L+V produced the lowest MSE for all methods.

Columns 3 to 7 in Table 4 present the F1-scores of the classification of user satisfaction labels (except friendliness) generated by training with different multimodal feature sets. Due to the imbalance in the friendliness label (49/11), all models overfitted this label. For the coordinateness dialogue label, in the LSTM and GRU methods, the ALL feature set (A+V+L) produced the best F1-scores (0.76 and 0.75) among all feature sets, and similarly, in the regression task, for the awkwardness label, L+V yielded the best F1-score for all methods.

6.2 Comparison between multimodal and exchange-level-annotation features (RQ2):

Column 8 in Table 3 presents the regression results based on the features of the exchange-level annotation (annotation features). We report the lowest MSE of each label result for the annotation features as follows. The MSE was 3.56 for the coordinateness label with LSTM, 4.33 for awkwardness with the GRU, and 3.09 for the friendliness with the LSTM. Although the performance of the annotation features was better in some cases, the multimodal feature sets achieved the best performance considering all the results.

Column 8 in Table 4 shows the user satisfaction classification results based on the annotation features. The results demonstrate

Table 3: Regression results of each multimodal combination for three user satisfaction labels (Acoustic (A), Visual (V), Linguistic (L), and Exchange annotation (Exchange An). The accuracy denotes the mean squared error (MSE). The bold values indicate the best MSE for the performance index.)

		Proposed					Human models	
Labels	**Model**	**L**	**A+L**	**A+V**	**L+V**	**ALL**	**(1) Exchange An.**	**(2) Wizard**
Coordinateness	**RNN**	3.66	**3.14**	3.28	4.18	3.56	3.66	3.38
	LSTM	3.58	3.05	3.17	4.14	**2.93**	3.56	
	GRU	3.35	**3.21**	3.28	3.92	3.32	3.75	
Awkwardness	**RNN**	4.46	4.43	4.59	**4.31**	4.48	4.41	5.70
	LSTM	4.58	4.63	4.58	**4.00**	4.54	4.44	
	GRU	4.57	4.52	4.52	**4.22**	4.34	4.33	
Friendliness	**RNN**	**2.87**	3.23	3.26	3.14	3.45	3.13	4.15
	LSTM	**2.88**	3.06	3.29	3.03	3.13	3.09	
	GRU	3.02	3.25	3.28	**2.87**	3.14	3.32	

Table 4: Binary classification F1-score of each multimodal combination of three user satisfaction labels (Acoustic (A), Visual (V), Linguistic (L), and Exchange-level annotation (Exchange An). The bold values indicate the best F1-score)

		Proposed Methood					Human models	
Labels	**Model**	**L**	**A+L**	**A+V**	**L+V**	**ALL**	**(1) Exchange An.**	**(2) Wizard**
Coordinateness	**RNN**	0.67	0.72	**0.73**	0.71	0.72	0.61	0.72
	LSTM	0.70	0.74	0.75	0.65	**0.76**	0.55	
	GRU	0.69	0.68	0.74	0.63	**0.75**	0.55	
Awkwardness	**RNN**	0.59	0.53	0.61	**0.72**	0.59	0.60	0.58
	LSTM	0.66	0.58	0.61	**0.68**	0.63	0.54	
	GRU	0.58	0.62	0.56	**0.63**	0.61	0.56	

that the RNN achieved the best performance for the coordinateness dialogue label, with an F1-score of 0.61. All multimodal features performed better than the annotation features for the coordinateness label. Similar to the coordinateness label, the RNN achieved the best performance (0.60) for the awkwardness label, which was better than that obtained by the RNN method trained using the A+V+L (0.59) feature set. For the other methods (LSTM and GRU), the multimodal feature performance was better in all cases. Overall, the results show that multimodal features can improve recognition performance. In this evaluation, we used the two-layered RNN based models for comparing models. However, the network architecture is not optimized for the Human model (1) with the low dimensional input (three dimensions), so the fair evaluation using the optimized network architecture per each model is future work.

6.3 Comparison of human model and ML models (RQ3):

Column 9 in Table 3 lists the regression results of the human model, in which the Wizard estimated user satisfaction. For the coordinateness label, the human model yielded an MSE of 3.38, which was better than some feature sets (A, L+V) but worse than the best result (2.93) achieved by LSTM. The regression results for the awkwardness and friendliness labels were worse than almost all ML model results.

Column 9 in Table 4 shows the classification results of the human model, in which the Wizard estimated user satisfaction. For the classification task, we calculated the F1-scores of binary classifications based on the annotations by the Wizard. Similar to the regression results, the human model obtained a better F1-score (0.72) than some feature sets (A, L+V) for the coordinateness label. In contrast, the result (0.76) of the LSTM model was higher than that of the human annotator. Most ML models performed better than the human models for the friendliness label. Overall, both regression and the classification results indicate that the performance of the multimodal model was higher than that of the human model.

7 DISCUSSION

7.1 Feature analysis

7.1.1 Contribution of each modality. To analyze the contribution of each modality to three satisfaction labels on classification tasks. We use ablation experiments, in which a GRU model was trained by removing feature sets one by one. If the F1-score decreased, the removed feature set was effective for the classification. On the contrary, if the F1-score improved, the removed feature set was not effective for classification. Table 5 shows the binary classification recognition performance of the GRU model on user satisfaction labels (except friendliness) trained with feature sets after each feature set was excluded. This table shows that the acoustic feature set was the most effective (+0.12) for the coordinateness label. The second most effective feature set was facial features (+0.10). The linguistic features were less effective. The results indicated that non-linguistic features performed better than linguistic features in identifying the coordinateness label. For the awkwardness label, the body features (+0.07) and linguistic features (+0.05) yielded better

Table 5: Contribution of each modality feature to two labels in GRU (Diff denotes the difference in F1-scores for cases in which a specific modality was removed)

Modality	Label			
ALL (A+V+L)	Coordinateness		Awkwardness	
	0.75		0.61	
Remove modality	F1	Diff	F1	Diff
Acoustic	0.63	0.12	0.63	-0.02
Facial	0.71	0.04	0.61	0.00
Action Unit	0.68	0.07	0.60	-0.01
Body	0.77	-0.02	0.54	0.07
Linguistic	0.74	0.01	0.56	0.05

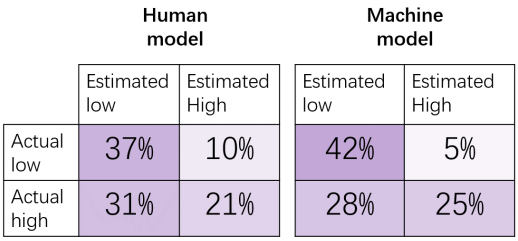

Figure 3: Confusion matrix of the binary classification task for the awkwardness label (ML models: LSTM regression result using the L+V feature set), and human model (annotation by the Wizard))

values, which means that body motion and linguistic features were effective in recognizing the awkwardness label. The acoustic features were less effective. For the friendliness label, the model with linguistic features achieved a better performance than the model with multimodal features (refer to Table 3). The results demonstrate that linguistic features can improve the performance more than other modality features. However, the difference is 0.12 in the maximum case, and the difference is not significant, so analyzing the specific features or frames in the sequence data, which significantly improves accuracy, is essential for future work.

7.1.2 Comparison between unimodal and multimodal. For the coordinateness and awkwardness labels, the recognition performance was improved with multimodal features. For the friendliness label, the difference in accuracy between the unimodal and the best multimodal models was insignificant (refer to Tables 3 and 5). We analyzed the results of Section 6.1 and the ablation experiments. For the coordinateness label, communication is a cooperative activity involving coordinated behaviors [19]. Participants in conversation spontaneously adjust facial expressions, postures, pronunciation and speech rates [3, 20, 21]. In this study, the multimodal fusion set (ALL) produced a better performance for the coordinateness label. For the awkwardness label, participants show a very negative attitude when they are embarrassed in dialogue with the agent. They do not often physically express their feelings and communicate with the agent. At the same time, the degree of embarrassment has an important relationship with the participation attitude in the dialogue. Body features were the most effective (refer to Table 5). The result partially aligned to the finding that body features (hand and head movements) are closely related to embarrassment[15]. For the friendliness label, the linguistic feature achieved a better performance in most cases in this study.

7.2 Comparison between Multimodal recognition and human perception

As shown in Section 6.3, we found that the performance of our proposed model was better than that of the human model (2). To further analyze the difference in accuracy between the human model and ML models, we evaluated the regression accuracy of 60 participants for the awkwardness label with the LSTM model using 5-fold cross-validation. In this study, we divided the regression values into binary values (high and low). We considered the threshold of 4.5 due to the regression result is continuous values. The evaluation using other threshold values is set as future work. To observe the overall classification, we calculated the confusion matrix of the machine and human (Wizard) scores. As shown in Figure 3, the recognition performance of the ML models was better than that of the human model for the high and low awkwardness labels. However, both the ML and human models showed false low-level recognition results (true high embarrassment was mistaken for low embarrassment), accounting for 31 % and 28 % of the total samples, respectively. This result suggests that both humans and machines have difficulty identifying high participant awkwardness at the dialogue level. In addition, compared to other labels, the regression MSE result for the awkwardness label was larger (refer to Table 3).

8 CONCLUSIONS

This paper proposed a multimodal user satisfaction recognition model suitable for evaluating non-task oriented dialogue systems at the dialog level by utilizing a novel multimodal user-system dialogue data corpus. To consider the contextual information in the dialogue, LSTM, RNN, and GRU structures were applied in this study. The results of the three different models indicated that multimodal features achieved a better performance than unimodal features, exchange annotation, and human models in user satisfaction recognition, which shows that our proposed model is reliable for identifying user satisfaction at the dialogue level. However, there is still room for improvement in multimodal user satisfaction. In this study, the feature vectors of different modalities were concatenated into one feature vector and used for training. In future work, we will focus on integrating multiple features better to improve the performance in these tasks and investigate the relevance of specific multimodal features in user satisfaction recognition.

ACKNOWLEDGMENTS

This work was partially supported by "Five-star Alliance", the Cooperative Research Program of NJRC Materials and Devices, KAKENHI: Grant-in-Aid for Scientific Research (A) (grant no. 19H01120), and JST AIP Trilateral AI Research (grant no. JPMJCR20G6), Japan.

REFERENCES

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

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

[3] Joseph N Cappella and Sally Planalp. 1981. Talk and silence sequences in informal conversations III: Interspeaker influence. *Human Communication Research* 7, 2 (1981), 117–132.

[4] Jeffrey F Cohn, Zara Ambadar, and Paul Ekman. 2007. Observer-based measurement of facial expression with the Facial Action Coding System. *The handbook of emotion elicitation and assessment* 1, 3 (2007), 203–221.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. (2018), 4171––4186.

[6] Klaus-Peter Engelbrecht, Florian Gödde, Felix Hartard, Hamed Ketabdar, and Sebastian Möller. 2009. Modeling user satisfaction with hidden Markov models. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. 170–177.

[7] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing* 7, 2 (2015), 190–202.

[8] Sunao Hara, Norihide Kitaoka, and Kazuya Takeda. 2010. Estimation Method of User Satisfaction Using N-gram-based Dialog History Model for Spoken Dialog System. In *Proc. International Conference on Language Resources and Evaluation (LREC)*.

[9] Ryuichiro Higashinaka, Yasuhiro Minami, Kohji Dohsaka, and Toyomi Meguro. 2010. Issues in predicting user satisfaction transitions in dialogues: Individual differences, evaluation criteria, and prediction models. In *Proc. International Workshop on Spoken Dialogue Systems (IWSDS)*. 48–60.

[10] Ryuichiro Higashinaka, Yasuhiro Minami, Kohji Dohsaka, and Toyomi Meguro. 2010. Modeling user satisfaction transitions in dialogues from overall ratings. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. 18–27.

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

[12] Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)* 38, 3 (2020), 1–32.

[13] Ute Jekosch. 2005. *Voice and Speech Quality Perception: Assessment and Evaluation (Signals and Communication Technology)*. Springer-Verlag, Berlin, Heidelberg.

[14] 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. ACM International Conference on Multimodal Interaction (ICMI)*. 315–323.

[15] Dacher Keltner. 1995. Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. *Journal of personality and social psychology* 68, 3 (1995), 441.

[16] Y Kikuta. 2019. BERT pretrained model trained on Japanese Wikipedia articles.

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

[18] Kazunori Komatani and Shogo Okada. 2021. Multimodal Human-Agent Dialogue Corpus with Annotations at Utterance and Dialogue Levels. In *Proc. International Conference on Affective Computing and Intelligent Interaction (ACII)*.

[19] Nida Latif, Adriano V Barbosa, Eric Vatiokiotis-Bateson, Monica S Castelhano, and KG Munhall. 2014. Movement coordination during conversation. *PLoS one* 9, 8 (2014), e105036.

[20] Gregory J McHugo, John T Lanzetta, Denis G Sullivan, Roger D Masters, and Basil G Englis. 1985. Emotional reactions to a political leader's expressive displays. *Journal of Personality and Social Psychology* 49, 6 (1985), 1513.

[21] Jennifer S Pardo. 2006. On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America* 119, 4 (2006), 2382–2393.

[22] Dunlu Peng, Ming Zhou, Cong Liu, and Jun Ai. 2020. Human–machine dialogue modelling with the fusion of word-and sentence-level emotions. *Knowledge-Based Systems* 192 (2020), 105319.

[23] Alexander Schmitt and Stefan Ultes. 2015. Interaction quality: assessing the quality of ongoing spoken dialog interaction by experts—and how it relates to user satisfaction. *Speech Communication* 74 (2015), 12–36.

[24] Stefan Ultes. 2019. Improving Interaction Quality Estimation with BiLSTMs and the Impact on Dialogue Policy Learning. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. 11–20.

[25] Stefan Ultes, Alexander Schmitt, and Wolfgang Minker. 2017. Analysis of temporal features for interaction quality estimation. In *Dialogues with Social Robots*. 367–379.

[26] Hongliang Yu, Liangke Gui, Michael Madaio, Amy Ogan, Justine Cassell, and Louis-Philippe Morency. 2017. Temporally selective attention model for social and affective state recognition in multimedia content. In *Proc. ACM International Conference on Multimedia*. 1743–1751.

[27] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1114–1125.

[28] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. In *Proc. Association for the Advancement of Artificial Intelligence (AAAI)*, Vol. 32.

[29] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018. Multi-attention recurrent network for human communication comprehension. In *Proc. Association for the Advancement of Artificial Intelligence (AAAI)*, Vol. 2018. 5642.