IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 April 2023, accepted 12 June 2023, date of publication 21 June 2023, date of current version 21 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3287984

RESEARCH ARTICLE

A Ranking Model for Evaluation of Conversation Partners Based on Rapport Levels

TAKATO HAYASHI[1], CANDY OLIVIA MAWALIM[1], RYO ISHII[2], AKIRA MORIKAWA[2], ATSUSHI FUKAYAMA[2], TAKAO NAKAMURA[2], AND SHOGO OKADA[1], (Member, IEEE)

[1]Japan Advanced Institute of Science and Technology, Nomi, Ishikawa 923-1292, Japan
[2]Human Informatics Laboratories, NTT Corporation, Yokosuka-shi, Kanagawa 239-0847, Japan

Corresponding author: Shogo Okada (okada-s@jaist.ac.jp)

This work was supported in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant 22H04860 and Grant 22H00536; in part by the Japan Science and Technology Agency (JST) Advanced Integrated Intelligence Platform Project (AIP) Trilateral Artificial Intelligence (AI) Research, Japan, under Grant JPMJCR20G6; and in part by the JST Moonshot Research and Development Program under Grant JPMJMS2237.

ABSTRACT Our proposed ranking model ranks conversation partners based on self-reported rapport levels for each participant. The model is important for tasks that recommend interaction partners based on user rapport built in past interactions, such as matchmaking between a student and a teacher in one-to-one online language classes. To rank conversation partners, we can use a regression model that predicts rapport ratings. It is, however, challenging to learn the mapping from the participants' behavior to their associated rapport ratings because a subjective scale for rapport ratings may vary across different participants. Hence, we propose a ranking model trained via preference learning (PL). The model avoids the subjective scale bias because the model is trained to predict ordinal relations between two conversation partners based on rapport ratings reported by the same participant. The input of the model is multimodal (acoustic and linguistic) features extracted from two participants' behaviors in an interaction. Since there is no publicly available dataset for validating the ranking model, we created a new dataset composed of online dyadic (person-to-person) interactions between a participant and several different conversation partners. We compare the ranking model trained via preference learning with the regression model by using evaluation metrics for the ranking. The experimental results show that preference learning is a more suitable approach for ranking conversation partners. Furthermore, we investigate the effect of each modality and the different stages of rapport development on the ranking performance.

INDEX TERMS Affective state, dyadic interaction, multimodal signal processing, preference learning, ranking model, rapport.

I. INTRODUCTION

The term *rapport* can be defined as a feeling of connection and harmony with someone else [1]. Building rapport plays an essential role in cultivating good relations with other people. Previous studies have shown that a high level of rapport improves learning gain in peer tutoring [2] and leads to successful negotiations [3]. Much research in recent years has focused on automatically measuring rapport levels from social signals in human-human [4] and human-agent interactions [5]. These rapport estimators can be applied to analyzing interpersonal relationships in interactions and to developing socially aware conversational agents. Matsuyama et al. [6], for example, proposed a robot assistant that can generate socially aware behavior due to an incorporated rapport estimator.

We address the novel task of **RA**nking **CO**nversation **P**artners based on self-reported rapport levels (RACOP). Many applications in rapport recognition can be formulated as ranking conversation partners. In one-to-one online language lesson services, for example, the evaluation of teachers can be based on user rapport built in past lessons; this is an important application area for RACOP. In these services, a user is automatically assigned a teacher available

The associate editor coordinating the review of this manuscript and approving it for publication was Wai-Keung Fung.

at the requested time. To recommend a teacher, service providers can use an ordered list of teachers created from a user's past lessons. RACOP is important for other applications, such as for the evaluation of virtual agents with various personalities and for matchmaking in an online game where a player communicates with other players via voice chat.

To rank conversation partners based on self-reported rapport levels, we can use a regression model for directly predicting rapport ratings. However, there are two concerns with this approach. One concern is that it is challenging to learn the mapping from participants' behavior to rapport ratings because regression models learn biases arising from individual differences in rapport ratings. The second concern is that the predicted rapport scores do not always correspond to the order of ground-truth rapport scores because regression does not learn ordinal relations. Martínez et al. [7] noted that regression for predicting affect ratings should be avoided because this approach introduces two biases—nonlinear scale and subjectivity of ratings. As with affect ratings, the difference between each point of rapport ratings may not be uniform (nonlinear scale); the evaluation criteria of the rapport ratings may vary across different participants (subjectivity of ratings).

Preference learning (PL) is an attractive alternative framework for avoiding two concerns and developing reliable and valid models. Therefore, we propose a deep learning model trained via PL for RACOP. The input of the model is multimodal (acoustic and linguistic) features extracted from two participants' behaviors in an interaction. The PL model is a more suitable approach for RACOP than regression because the PL model is directly trained to predict ordinal relations between two conversation partners based on rapport ratings reported by the same participants. Furthermore, transforming rapport ratings into ordinal relations avoids the bias of different subjective scales across participants because each participant has consistent evaluation criteria to some extent. In addition, the PL model is not affected by the nonlinear scale bias because the model does not directly use scalar values of rapport ratings.

Previous studies in affective computing [8], [9] showed that ranking models trained via PL have considerable advantages over regression models. In these studies, they constructed models that rank samples according to levels of emotional attributes; however, no studies have included the application of PL to a rapport recognition model.

Since there was no suitable dataset for evaluating RACOP, we collected online dyadic (person-to-person) interactions between a participant and several different conversation partners. To analyze the effect of the various stages of rapport development on ranking performance, we recorded three interactions for each pair of participants based on various topics: 1) self-introductions, 2) introduction of positive and negative experiences, and 3) introduction of self-shortcomings. After every interaction, participants reported rapport ratings for their conversation partner.

The main contributions of this paper are as follows:

1) To our knowledge, this is the first study to address ranking conversation partners based on rapport levels.
2) We create a dataset composed of interactions between a participant and several different conversation partners with self-reported rapport ratings.
3) We propose a ranking model to rank conversation partners trained via preference learning. Then, we show that preference learning is a more suitable approach than regression for RACOP.
4) To understand RACOP more thoroughly, we clarify the effect of each modality and the various stages of rapport development on ranking performance.

Section II presents a survey of the works related to our study. Section III introduces our dataset and annotation methods. Section IV presents the methodology to address RACOP. Section V describes our comparison method and experimental settings. Section VI shows the experimental results, and we discuss them.

II. RELATED WORKS

First, we introduce the research related to analyzing and to predicting rapport in interactions (Section II-A). Then, we introduce the works that address annotation methods for affective states and appropriate processing of the annotations (Section II-B).

A. RAPPORT

In social psychology, rapport is considered to play an essential role in building good relationships with a conversation partner. Early studies focused on illuminating nonverbal cues that indicate rapport. Tickle-Degnen and Rosenthal [10] investigated nonverbal behavior that correlated with rapport. They also described rapport in terms of three components: mutual attentiveness, positivity, and coordination. Bernieri et al. [11] analyzed observable cues of rapport in two contexts—adversarial and cooperative. Furthermore, Grahe and Bernieri [12] showed that observers who accessed nonverbal information evaluated rapport more accurately than observers who accessed verbal information.

Much research in recent years has focused on automatically measuring rapport levels from social signals in human–human and human–agent interactions. Visual information such as posture [4] and facial expressions [13] are commonly used for predicting rapport. Furthermore, Cerekovic et al. [5] used verbal and nonverbal cues to measure user rapport in human-agent interactions. Müller et al. [14] proposed a model to detect low rapport in group interactions. Sinha and Cassel [2] showed that high rapport with a student improves learning gains in peer tutoring. Previous studies [15], [16], therefore, addressed the automatic prediction of rapport in peer tutoring. Raphalen et al. [17] also constructed a computational framework for identifying hedges that are important for managing rapport in peer tutoring.

Attractive applications for the use of rapport estimators are socially aware conversational agents and recommendation systems. Previous studies [1], [18] developed virtual agents that promote a sense of rapport with a human speaker. Furthermore, Matsuyama et al. [6] proposed a socially aware robot assistant (SARA) to achieve both a task goal (recommending information) and a social goal (building rapport). SARA can generate socially aware behavior due to an incorporated rapport estimator. Abulimiti et al. [19] hypothesized that off-task episodes raised rapport levels in peer tutoring. They, therefore, proposed a planning model that allows a virtual agent to generate off-task episodes according to user rapport levels.

B. AFFECTIVE COMPUTING AND PREFERENCE LEARNING

To capture participants' affective states, choosing an appropriate measurement is a key problem in affective computing. An interval and an ordinal scale are often used to measure levels of affective states. A popular tool for measuring the interval scale is the FeelTrace software [20]; popular tools for measuring the ordinal scale are the Likert scale questionnaire [21] and the Self-Assessment Manikin [22].

To automatically recognize the affective state reported by these tools, many researchers have developed models to predict an intensity or a class via the regression/classification framework. This approach, however, is problematic. The regression model to predict affect ratings is unreliable because the evaluation criteria of the annotation may vary across different people [7]. In the previous studies [23], [24], they noted that the self-reported affective evaluation process is biased due to the factors of the environment, personal experience, and individual perception. Furthermore, the ordinal scale (e.g., Likert scale) is often treated as the ratio scale for regression; however, Martínez et al. [7] discussed that the implicit transformation from the ordinal scale to the ratio scale introduces a nonlinear scale bias. Considering the 5-point Likert scale questionnaire, affect ratings are not linear because the difference between each point may not be uniform. For the above reasons, it is challenging to learn the mapping from the participants' behavior to their affect ratings. The transformation from affect ratings to class may mitigate the subjective and the nonlinear scale bias, but Martínez et al. [7] also discussed that this practice adds a new type of bias due to the class splitting criteria. As these studies show, it is questionable whether the regression and classification framework is a suitable method for predicting affective states.

Preference learning (PL) is an appealing alternative framework for developing reliable and valid models in affective computing [25]. PL models are trained to predict the preference among paired samples with ordinal labels. Given two samples (s_A and s_B), the ordinal labels are represented as follows: $s_A \succ s_B$ or $s_A \prec s_B$. The symbols "$\prec$"/"$\succ$" express the preceding/succeeding order of the samples. The PL model is not affected by the nonlinear scale bias because the model does not directly use scalar values of levels of affective states. Furthermore, when levels of affective states are transformed into ordinal relations for each participant, the bias of different subjective scales across participants is avoided because each participant has consistent evaluation criteria to some extent.

There are two approaches to collecting ordinal labels: direct and indirect [25]. The direct approach is that annotators are asked to report their preference between paired samples. This approach has been applied to many tasks in affective computing, such as music [26], sound [27], and facial expression [28]. The indirect approach is that levels of affective states (reported by the interval or the ordinal scale) are transformed into ordinal labels. This approach has also been applied in many studies [8], [9], [29]; then, ranking models were trained via preference learning. Previous studies [8], [9] showed that ranking models via preference learning have significant advantages over conventional regression models. Martínez et al. [7] also indicated that transforming affect ratings into ordinal labels leads to more generalized models when compared to transforming the same ratings into a class. Furthermore, Zoumpourlis and Patras [30] showed that incorporating an auxiliary task of ordinal ranking leads to consistent performance gains for the regression and classification tasks.

Inspired by studies in preference learning for affective computing, we apply the PL framework to the model for rapport recognition. Rapport ratings are affected by the subjective scale bias as well as affect ratings. Nevertheless, no studies have attempted to explore preference learning in rapport recognition. We transform rapport ratings to ordinal labels for each participant and develop a PL model to predict the preference between two conversation partners.

TABLE 1. Dataset summary.

No. of participants	69
No. of pair of participants	96
No. of interactions	288
No. of utterances	163,026
Gender of participants	35 / 34 (male / female)
Age of participants	51 / 18 (twenties / thirties)

III. A DATASET FOR DYAD INTERACTIONS

Since there was no suitable public dataset for evaluating RACOP, we created a new dataset composed of online dyad interactions with rapport ratings. The unique point of this dataset is that we recorded dyad interactions between a participant and several different conversation partners. Our dataset consists of 288 interactions in Japanese. Each interaction lasted approximately 20 minutes, resulting in a total of more than 96 hours. Table 1 summarizes the statistics of our dataset. Since the dataset collected in this study contains self-disclosure regarding the personal information of the participants, we do not make the dataset public.

A. INTERACTION SETTING

We recruited 69 Japanese-speaking participants (35 male, 34 female) through a recruitment agency. Participants were divided into two categories according to recruitment methods. Participants in the first category took part in the experiment with three friends, and the number of these participants was 32 (16 male, 16 female); participants in the second category took part in the experiment alone, and the number of these participants was 37 (19 male, 18 female). The purpose of recruiting according to two methods is not relevant to the current work and is not discussed further.

Each participant in the first category was combined with participants in the second category randomly to form a same-gender pair of participants, resulting in a total of 96 pairs. We ensured that pairs of participants did not know each other prior to the recording. Every participant in the first category communicated with only three conversation partners. The number of partners for participants in the second category depended on the specific person and ranged from one to six.

They communicated with each other in different rooms through the video communication system. The data recording took place in a quiet room equipped with a camera and a microphone. They were able to recognize their partners' facial expressions and voices through a display and an earphone. During the recording, we placed the camera to show a participant's entire face. Some visual-based social signals—gestures and postures—are less easily conveyed to a conversation partner in online interactions than in face-to-face interactions; however, it is worth measuring rapport levels in online interactions because the frequency of usage of video communication tools has increased during the COVID-19 pandemic. All participants provided written informed consent to participate, and the study was reviewed and approved by the Research Ethics Committee of the NTT Corporation.

B. CONVERSATION TOPICS AND SELF-DISCLOSURE

Tickle-Degnen and Rosenthal [10] suggested that the importance of three behavioral components—mutual attentiveness, positivity, and coordination—for building rapport differs according to the stage of rapport development, for example, the first meeting or not. To investigate relationships between the stage of rapport development and ranking performance, we recorded three interactions based on various topics for each pair of participants.

We selected three topics—a self-introduction, an introduction of positive and negative experiences, and an introduction of self-shortcomings—to help pairs of participants develop interpersonal relationships through self-disclosure. Essential to developing interpersonal relationships is breadth, the variety of the topics discussed and depth, the degree of intimacy that guides these interactions [31]. In the early stages of a relationship, people share superficial information such as self-introductions. As the relationship progresses, people share more intimate information, such as thoughts and emotions [32]. Sharing self-shortcomings is a particularly intimate topic because of the fear of their partners' negative appraisal [33].

In the first interaction, both participants introduced themselves and discussed subjects such as how they liked to spend their days off, their favorite foods, and their favorite artists. In the second interaction, they told each other stories about happy and sad moments in their life. In the last interaction, they spoke about their personal shortcomings. Each interaction lasted 20 minutes, and there were a few minutes of break time between interactions. To enhance interactions, we instructed them to not only listen but also to actively react and to ask questions while their conversation partners spoke.

C. SELF-REPORTED ANNOTATIONS

We instructed them to complete a questionnaire with 18 items after every interaction. The questionnaire was proposed by Bernieri et al. [11] to measure participants' rapport levels for their conversation partners. Translations of 18 items for Japanese speakers were created in a previous study, and its reliability is sufficient (α = 0.92) [34]. The 18 items are ''well-coordinated'', ''boring'', ''cooperative'', ''harmonious'', ''unsatisfying'', ''uncomfortably paced'', ''cold'', ''awkward'', ''engrossing'', ''unfocused'', ''involving'', ''intense'', ''unfriendly'', ''active'', ''positive'', ''dull'', ''worthwhile'', and ''slow''. They rated each item on an 8-point Likert scale as in the original study [11]. A value of 1 corresponds to ''strongly disagree'', and a value of 8 corresponds to ''strongly agree''. We summed the values of 18 items after the values of negative questions were reversed. We defined a *rapport score* as the total score.

The Pearson correlation coefficient between rapport scores of participants in the first and the second category is 0.25. This value indicates a weak positive correlation among pairs of participants.

The mean values of rapport scores increase as the number of interactions increases. The mean value of the first topic is 108.60 (SD = 20.81), the second topic is 114.03 (SD = 19.80), and the last topic is 118.38 (SD = 20.45). Post hoc comparisons using the t test with Bonferroni correction were conducted to examine the statistical significance in the mean values of rapport scores between three topics (significance level is $p < 0.001$). The mean value of the first topic is significantly different than the mean value of the second topic ($t = 7.41$, $p = 0.00$, $df = 191$). The mean value of the second topic is also significantly different than the mean value of the last topic ($t = 5.21$, $p = 0.00$, $df = 191$).

We assume that there are two reasons for the results. One is that the total interaction time of the pair of participants increased as the number of interactions increased. Participants show an increased liking for their conversation partners as they are exposed to their partners more. This phenomenon is called the mere-exposure effect [35]. Another reason is that they were required to reveal intimate information about

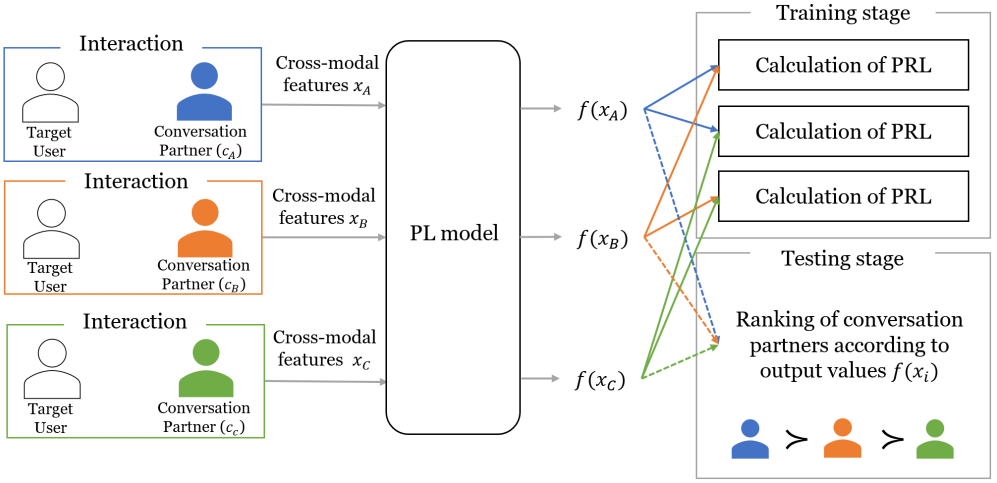


FIGURE 1. An overview of our proposed method. During the training stage, output values of the preference learning (PL) models are used for calculating a pairwise ranking loss (PRL). During the testing stage, we rank conversation partners according to the output values.

themselves as the number of interactions increased. A previous study [36] demonstrated that self-disclosure contributes to building rapport. However, not all participants benefited from the three topics because there are individual differences in the extent to which self-disclosure contributes to rapport building [37].

IV. METHODOLOGY

In this study, we develop models that rank conversation partners based on self-reported rapport levels. This problem can be formulated as pairwise comparisons between two conversation partners via the preference learning (PL) framework. We use a PL algorithm inspired by RankNet [38] and multimodal (acoustic and linguistic) features for the model's input. Figure 1 presents an overview of our proposed method.

We first propose a problem definition (Section IV-A). We then describe a loss function and a model architecture (Section IV-B). Finally, we explain the details of the multimodal features used in this study (Section IV-C).

A. PROBLEM DEFINITION

We define a *target user* as a participant who gives rapport ratings to their partner; we define a *conversation partner* as a participant for whom the target user gives rapport ratings. In the dyad interaction, rapport ratings are bidirectional; accordingly, if we regard one participant as the target user, we regard the other participant as the conversation partner and vice versa.

$\mathcal{C} = [c_1, c_2, \cdots, c_n]$ is defined as the list of conversation partners, where c_i is the i-th partner of a target user, and n is the number of their partners. Because the list $\mathcal{C}$ is created individually for each target user and each topic (see Section IV-B), all data $\mathcal{D}$ can be denoted as

$$\mathcal{D} = \{\mathcal{C}_{jk} \mid j = 1, 2, \cdots, m, k = 1, 2, 3\}, \tag{1}$$

where j and k are the j-th target user and the k-th topic, respectively. Let m be the number of target users. For conciseness of notation, we omit jk in $\mathcal{C}_{jk}$ in the following section.

Each list $\mathcal{C}$ is associated with a list of features $\mathcal{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n]$ and a list of scores $\mathcal{Y} = [y_1, y_2, \cdots, y_n]$. Features $\boldsymbol{x}_i$ are created from the target user's features $\boldsymbol{x}_i^{\text{user}}$ and their partner's features $\boldsymbol{x}_i^{\text{partner}}$ in an interaction; therefore, $\boldsymbol{x}_i = (\boldsymbol{x}_i^{\text{user}}, \boldsymbol{x}_i^{\text{partner}})$. The score y_i is defined as the rapport score that a target user gives to their i-th partner.

In this study, we develop ranking models that rank conversation partners for each list $\mathcal{C}$ in the order of the rapport scores. The training set $\mathcal{T}$ is constructed as follows: if two samples c_A and c_B are chosen from the same list C, then a paired sample $((\boldsymbol{x}_A, y_A), (\boldsymbol{x}_B, y_B))$ is added to $\mathcal{T}$. An ordinal label ($c_A \succ c_B$ or $c_A \prec c_B$) is determined according to ordinal relations among y_A and y_B. During the training stage, the PL model learns the mapping from the participants' behavior in each interaction ($\boldsymbol{x}_A$ and $\boldsymbol{x}_B$) to the ordinal labels.

B. PREFERENCE LEARNING

1) PAIRWISE RANKING LOSS FUNCTION (PRL)

We use a pairwise ranking loss function proposed by Burges et al. [38]. We consider a model f that maps the feature vector x to the real value $f(x)$. Given two samples c_A and c_B, the probability that c_A is preferred over c_B is given by P_{AB}:

$$P_{AB} = \frac{\exp(o_{AB})}{1+\exp(o_{AB})}, \tag{2}$$

where $o_{AB} = f(x_A) - f(x_B)$. During the training stage, the target probability $\bar{P}_{AB}$ is set according to the ordinal labels between two samples. $\bar{P}_{AB} = 0$ implies that c_B is preferred over c_A; $\bar{P}_{AB} = 1$ implies that c_A is preferred over c_B. We use the cross-entropy loss function $\mathcal{L}_{AB}$:

$$\mathcal{L}_{AB} = -\bar{P}_{AB} \log P_{AB} - \left(1 - \bar{P}_{AB}\right) \log \left(1 - P_{AB}\right). \tag{3}$$

The loss is backpropagated to the network parameters.

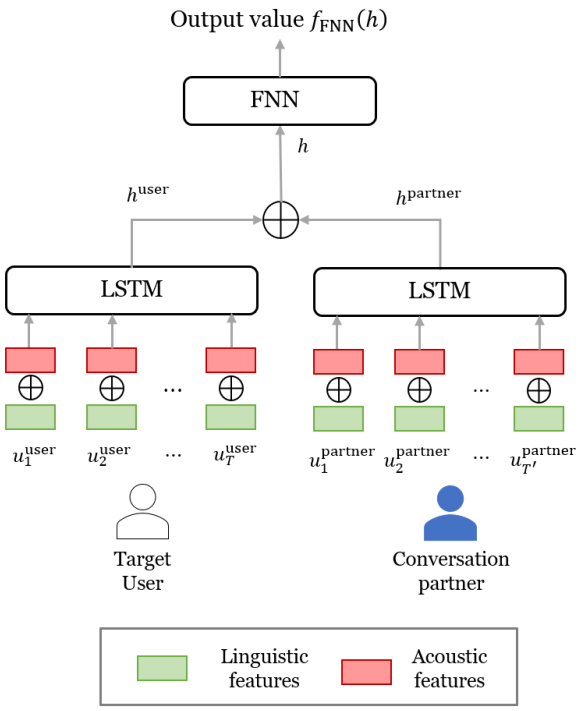


FIGURE 2. An architecture of our proposed model. Our PL model consisted of unidirectional long short-term memory (LSTM) networks and feedforward neural networks (FNN). The input of the LSTM is multimodal—linguistic and acoustic—features extracted for each utterance in an interaction.

2) MODEL ARCHITECTURE

Our PL model consists of unidirectional long short-term memory (LSTM) networks and feedforward neural networks (FNNs). Figure 2 illustrates the overview of the model architecture. To model the sequence of multimodal features, we used two-layer LSTM networks separately for two participants in an interaction. In this study, we used the early fusion method. Unimodal feature vectors (linguistic: 768 dim., acoustic: 88 dim.) were extracted from the participant's t-th utterances; then, these vectors were concatenated into a multimodal feature vector $\boldsymbol{u}_t$ (856 dim.). The inputs of the LSTM networks were

$$\boldsymbol{x}^{\mathrm{user}} = \left[\boldsymbol{u}_1^{\mathrm{user}}, \boldsymbol{u}_2^{\mathrm{user}}, \cdots, \boldsymbol{u}_T^{\mathrm{user}}\right], \tag{4}$$

$$\boldsymbol{x}^{\mathrm{partner}} = \left[\boldsymbol{u}_1^{\mathrm{partner}}, \boldsymbol{u}_2^{\mathrm{partner}}, \cdots, \boldsymbol{u}_{T'}^{\mathrm{partner}}\right], \tag{5}$$

where T is the number of users' utterances and T' is the number of their partner's utterances in an interaction. We used the output vector corresponding to the last utterance as the embedding vector. The target user's embedding vector $\boldsymbol{h}^{\mathrm{user}}$ and the conversation partner's embedding vector $\boldsymbol{h}^{\mathrm{partner}}$ were concatenated into the embedding vector $\boldsymbol{h}$.

$$\boldsymbol{h}^{\mathrm{user}} = \mathrm{LSTM}\left(\boldsymbol{x}^{\mathrm{user}}\right), \tag{6}$$

$$\boldsymbol{h}^{\mathrm{partner}} = \mathrm{LSTM}\left(\boldsymbol{x}^{\mathrm{partner}}\right), \tag{7}$$

$$\boldsymbol{h} = \boldsymbol{h}_T^{\mathrm{user}} \oplus \boldsymbol{h}_{T'}^{\mathrm{partner}}. \tag{8}$$

To map the vector $\boldsymbol{h}$ to the output value $f_{\mathrm{FNN}}(\boldsymbol{h})$, we used a two-layer FNN:

$$f_{\mathrm{FNN}}\left(\boldsymbol{h}\right) = \mathrm{FNN}\left(\boldsymbol{h}\right). \tag{9}$$

We represent equations (6)-(9) as one function $f(\boldsymbol{x})$.

During the training stage, this output value was used for calculating the loss (see IV-B). During the testing stage, we considered that $f(\boldsymbol{x}_A) > f(\boldsymbol{x}_B)$ implies $c_A \succ c_B$; therefore, if $f(\boldsymbol{x}_A) > f(\boldsymbol{x}_B) > f(\boldsymbol{x}_C)$, then the predicted global order list is $c_A \succ c_B \succ c_C$.

C. FEATURE EXTRACTION

1) ACOUSTIC FEATURES

We used OpenSMILE [39] software to extract acoustic features from each utterance. The acoustic features correspond to eGeMAPS [40], the de facto standard preset in speech emotion recognition. The preset contains 88 parameters, such as pitch and loudness. The acoustic features were extracted from each utterance and normalized for each person using z score normalization.

2) LINGUISTIC FEATURES

BERT [41] is a language representation model that achieves state-of-the-art performance on many natural language processing tasks. Recent studies have shown that BERT is also helpful in emotion recognition in conversation [42], [43]. A model pretrained on only Japanese text was applied in this study; the Japanese-BERT was developed at Tohoku University.[1] The participants' utterances were transcribed into text data by an automatic speech recognition system; then, we used the Japanese-BERT to extract features from each utterance. We used the output vector corresponding to the first token (the [CLS] token) as utterance features. This output vector is 768-dimensional.

V. EXPERIMENTAL SETTINGS

A. COMPARISON MODEL (REGRESSION)

To compare the results with the ranking performance of the preference learning (PL) model, we developed a regression model built with neural networks. The architecture of the regression model was the same as the PL model, and the regression model also consisted of two-layer LSTM networks and two-layer FNN. The regression model, however, predicts the exact values of the rapport score for each interaction. We used the mean squared error (MSE) as the loss function in the regression. During the testing stage, we ranked conversation partners for each target user in the order of predicted rapport scores because predicted rapport scores of an ideal regression model correspond to the order of ground-truth rapport scores.

B. HYPERPARAMETER SETTINGS

For PL and regression, we set the batch size as 32 and the number of epochs as 30 without early stopping. We also

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

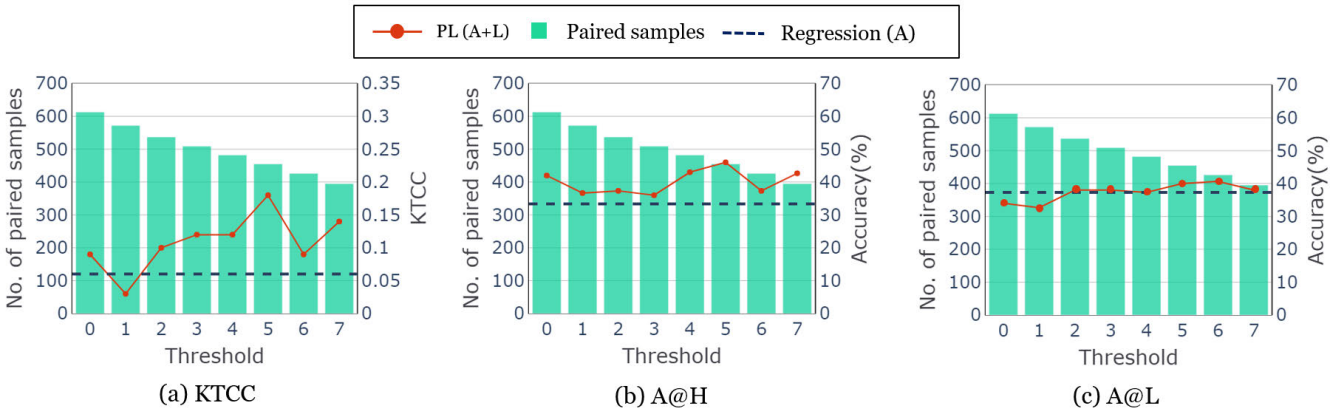


FIGURE 3. Ranking the performance of PL models (orange marker, right y-axis) and No. of paired samples (green, left y-axis) for various margin thresholds (x-axis). The dotted lines indicate the performance of the unimodal model trained on acoustic features (A); it is the best regression model among the regression models (see lines 6-8 of Table 2).

used the Adam optimizer. Hyperparameter optimization was performed via Optuna; Optuna is an automatic hyperparameter optimization software framework [44]. The number of trials for searching a combination of hyperparameters was 10. We defined the range of possible values as follows: the learning rate = $[5e^{-5}, 1e^{-5}, 5e^{-6}]$, the drop rate = $[0, 0.1, 0.3]$, and the number of hidden units (LSTM first-layer, LSTM second-layer, FNN first-layer, FNN second-layer) = $[128, 256, 516]$.

C. EVALUATION METRIC

To evaluate ranking performance, we calculated Kendall's tau correlation coefficient (KTCC), the accuracy at the highest-rapport conversation partner (A@H), and the accuracy at the lowest-rapport conversation partner (A@L). KTCC measures the correlation between the predicted ordered list and the ground-truth ordered list. A@H measures the accuracy of retrieving the highest-rapport conversation partner in the ground-truth ordered list, and A@L measures the accuracy of retrieving the lowest-rapport conversation partner.

D. EVALUATION PROCEDURE

We evaluated models by a double cross-validation approach. As the outer fold, we used leave-one-person-out cross-validation (LOPOCV); as the inner fold, we used hold-out validation. LOPOCV and hold-out validation ensure that all interactions that were engaged in by a target user or their conversation partners in the testing (validation) set were excluded from the training set. In hold-out validation, we randomly chose two participants—male and female—as target users from the training set, and we used their interactions as the validation set for hyperparameter optimization. Fixed seed values determined the combination of a target user for the testing set and target users for the validation set. The combination was the same throughout a series of experiments. The reason we used not cross-validation but hold-out validation as the inner fold was to reduce computational cost.

Nineteen out of 69 participants communicated with two or fewer conversation partners. We did not consider them as the target user because short, ordered lists cause ranking performance for the models to be overestimated or underestimated. Three lists (three topics) were created from each fold (50 target users), resulting in 150 (50×3) lists. We reported the average ranking performance of 50 folds to evaluate the generalization performance for the models.

For PL, we used the accuracy of pairwise comparison (AP) as the evaluation metric for hyperparameter optimization. AP is the accuracy for binary classification of ordinal labels ($c_A \succ c_B$ or $c_A \prec c_B$). The reason we used AP rather than ranking metrics is described in Section V-E.

For regression, we used RMSE as the evaluation metric for hyperparameter optimization. The reason we used RMSE is that the goal of comparison between models is to compare models trained via the PL framework with models trained via the general regression framework. As a general practice in training regression models, RMSE is used as the evaluation metric.

E. MARGIN THRESHOLD

Lotfian and Busso [8] showed that the difference among emotion levels of a paired sample improves the reliability of the training set. We define the *margin* as the absolute value of the difference among rapport scores: margin $m = |y_A - y_B|$, where y_A and y_B are rapport scores. If the margin m is greater than a given threshold, we used the paired sample as the input of the PL model for training.

We hypothesize that a margin threshold increases the reliability of the paired samples because the threshold reduces the uncertainty in an ordinal relation of a paired sample. Even the rapport score that is self-reported is slightly noisy. Metallinou and Narayanan [24], for example, reported that raters modify

TABLE 2. Ranking performances for PL models with the threshold set at 5 and regression models: A+L, acoustic and linguistic features (multimodal); A, acoustic features; L, linguistic features. The random baseline is the average performance over 100 trials.

		KTCC	A@H	A@L
Model	Modality			
PL	A	0.09	45.33	34.67
	L	0.06	32.67	32.00
	A+L	**0.18**	**46.00**	**40.00**
Regression	A	0.06	33.33	37.33
	L	-0.05	26.00	28.67
	A+L	-0.00	32.67	30.67
Random (100 trial)		-0.00	31.52	31.92

their ratings when experimenters ask them to annotate once more. This report suggested that the ordinal relations of the paired sample with close rapport scores may vary due to intrapersonal variability. In contrast, we can consider that the ordinal relations of the paired sample with a large margin are reliable and valid. The larger margin, however, reduces the number of paired samples in the training set because fewer paired samples satisfy the threshold.

To reduce uncertainty in the validation set, we also applied the margin threshold to the validation set. Then, we used AP as the evaluation metric for hyperparameter optimization because we cannot calculate ranking performances for a subset that consists of paired samples satisfying the threshold.

VI. RESULTS AND DISCUSSION

We first compare the preference learning (PL) model with the regression model to validate our proposed method (Section VI-A). We then investigate the contribution of each modality for RACOP on both PL and regression (Section VI-B). Finally, we examine how the stage of rapport development impacts ranking performance (Section VI-C).

A. COMPARISON OF PREFERENCE LEARNING AND REGRESSION

We show that PL is a more suitable approach for RACOP than regression. Then, we demonstrate that the margin threshold improves the reliability of the training and validation sets.

First, we compare the multimodal PL model with the best regression model. The 6-8 lines of Table 2 show the ranking performance of regression models when using various modalities. The best regression model is the unimodal model trained on acoustic features (KTCC, 0.06; A@H, 33.33; A@L, 37.33). For the PL model, we evaluated the ranking performances in a range of margin thresholds from 0 to 7. The reason for the range is that the number of paired samples in the validation set is not enough in some folds when the threshold is higher than 7. If we set the threshold as 8, the number of paired samples in the validation set is less than or equal to three pairs in some folds.

Figure 3 shows the ranking performance of the PL model for each margin threshold (orange marker) and the best regression model (dotted line). As the figure shows, the multimodal PL model outperforms the best regression model for all metrics as long as a sufficient threshold is set. For KTCC, the multimodal PL model outperforms the best regression model except for $m = 1$; for A@H, the multimodal PL model outperforms the best regression model for every threshold. Although the accuracy of the two models is similar for A@L, the multimodal PL model is slightly better as long as the threshold is more than 1. The results show that PL is a more suitable approach for RACOP than regression. One explanation for the results is that PL is less affected by two biases—nonlinear scale and subjectivity of ratings [7].

Second, we investigate the relationship between the margin threshold and the ranking performance of the PL model. Figure 3-(a) shows that KTCC improves with the increasing threshold in the 1 to 5 range. The results suggest that a margin threshold improves the reliability of the training and validation sets. KTCC, however, drops when the margin is greater than 6 because the large margin reduces the number of paired samples that can be used for training. The green bar indicates the number of paired samples that satisfy the threshold out of all paired samples.

B. ANALYSIS OF EFFECTIVE MODALITIES

We investigate the contribution of each modality to RACOP. First, we compare unimodal models trained on acoustic features (A) with models trained on linguistic features (L) on both PL and regression. In this experiment, we set the margin threshold as 5 for PL. Table 2 shows that the PL model (A) outperforms the PL model (L) for all ranking metrics; the regression model (A) also outperforms the regression model (L). We can therefore conclude that acoustic features are more effective for RACOP than linguistic features. The results agree with other researchers who reported that nonverbal cues are more reliable than verbal cues because nonverbal behavior occurs unconsciously [45]. Furthermore, the ranking performance of the regression model (L) is lower than that of the random baseline. In our datasets, linguistic features impair the ranking performance of the regression model. The results suggest that extracting linguistic cues to predict exact values of rapport ratings is more difficult than extracting linguistic cues to predict ordinal relations of them.

Second, the table shows the effectiveness of multimodal features for PL. Among all models, the multimodal PL model achieves the best performance for all metrics. The results suggest that multimodal features by early fusion lead the PL model to capture cues for the rapport levels that the unimodal model does not capture. The performance of the multimodal regression model, however, is lower than that of the unimodal regression model (A) for all metrics.

C. THE STAGE OF RAPPORT DEVELOPMENT

We analyze the relationship between the stage of rapport development and the ranking performance of PL models. In our datasets, participants communicated with each other based on three topics. Pairs of participants gradually built

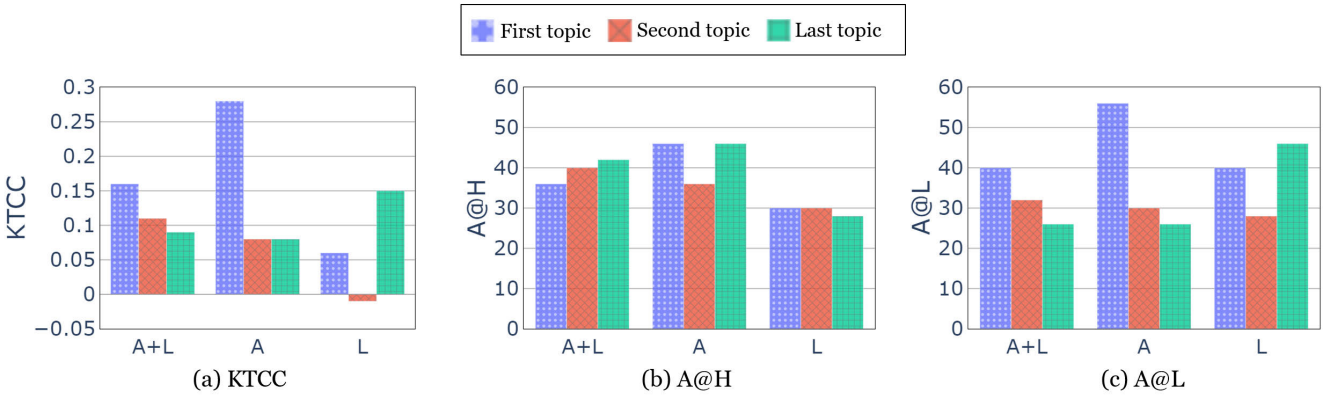


FIGURE 4. Ranking performances for each subset that consisted of interactions with one topic (color). Models were also trained on only each subset, and there are three models as follows: the multimodal model trained on acoustic and linguistic features (A+L), the unimodal model trained on acoustic features (A), and the unimodal model trained on linguistic features (L).

rapport as the number of interactions increased (see III-C). We divided all data into three subsets according to topics. Figure 4 shows the evaluation for each subset, and the models were trained by only one subset. The experimental settings are the same as previous experiments except that the dataset is a subset. In this experiment, we are able to use only one-third of the interactions for training; accordingly, we set the margin threshold as 0 to use as many interactions as possible.

First, we focus on multimodal PL models (A+L). As Figure 4 shows, the performance of KTCC and A@L for the first topic is the highest, and the performance decreases as the number of interactions increases. In contrast, the performance of A@H for the first topic is the lowest, and the performance increases as the number of interactions increases.

For KTCC, the results show that it becomes more difficult for our model to predict the order of rapport levels as pairs of participants gradually build rapport. We considered that there are two ways to interpret the results—assimilation and the difficulty of capturing coordination cues.

One interpretation of the results is that the differences in participants' behavior according to rapport levels decrease because rapport levels that participants rate for their partner converge at a certain level as the number of interactions increases. This convergence is called assimilation [46]. To validate this interpretation, we examined whether there are significant differences in the mean margin of rapport score between the paired sample among three topics. The metrics indicate the extent to which a participant rates their conversation partners in the same way. The mean margin between paired samples for the first topic is 15.69 (SD = 13.18), the second topic is 15.19 (SD = 12.36), and the last topic is 14.54 (SD = 13.76). The result shows that the mean margin between paired samples decreases as the number of interactions increases. The results of the t test with Bonferroni correction (the significance level is $p < 0.001$), however, showed that no significant differences are observed between topics (the first topic–the second topic: $t = 0.64$, $p = 0.53$, $df = 206$, the second topic–the last topic: $t = 0.92$, $p = 0.36$, $df = 206$, the first topic–the last topic: $t = 1.27$, $p = 0.21$, $df = 206$). Assimilation, therefore, is inadequate to explain the decreasing performance of KTCC.

Another interpretation of the results is that our models cannot capture cues of coordination in late interactions. Tickle-Degnen and Rosenthal [10] suggested that the importance of three behavioral components—mutual attentiveness, positivity, and coordination—for building rapport differs according to the stage of rapport development. The presence of positivity, for example, plays a more important role in developing rapport during early interactions (first-time meeting), and the degree of coordination plays a more important role during late interactions [10]. Cues indicating coordination, for example, are interactional synchrony and mirroring. Meta-analyses reported that the relations between cues indicating coordination and positive social outcomes (e.g., rapport) are robust during both verbal and nonverbal behavior [47], [48]; furthermore, Natale [49] examined levels of vocal intensity synchrony in three interactions for each pair of participants. The results showed that levels of vocal intensity synchrony are greater as the number of interactions increases. These studies suggest that behavior related to coordination is observed more frequently as rapport levels increase. Cues indicating coordination may be difficult to encode in our models because our model treats the sequence of two participants in an interaction separately. On the other hand, positivity—feelings of happiness and friendliness—may be encoded more easily than coordination; therefore, the KTCC of our models in early interactions is higher than the KTCC in late interactions. From Figure 4-(a), we can infer that cues indicating positivity are more clearly observed in acoustic features than in linguistic features.

For A@H, the results show that the multimodal PL model can determine the highest-rapport conversational partner in late interactions more accurately than in early

interactions. Even with an overall increase in the rapport levels with conversation partners, there may be a clear difference between participants' behaviors in interactions with the highest-rapport partner and those with the other partners. In contrast, for A@L, the multimodal PL model can determine the low-rapport conversational partner in early interactions more accurately than in late interactions; furthermore, we can observe similar changes in the unimodal PL model (A). From this result, we can infer that cues indicating low rapport in early interactions are more clearly observed in acoustic features than in linguistic features.

For all ranking metrics of the first topic, the unimodal PL model (A) outperforms the multimodal PL model. Our interpretation of the results based on social penetration theory [32] and our observations of some videos is as follows. On the first topic (first-time meeting), the verbal content of utterances may not only be ineffective for predicting rapport levels but also be noise because participants share simple and safe information according to social norms. On the other hand, for intimate topics (e.g., the introduction of self-shortcomings), the verbal content of utterances may be effective for predicting rapport levels because participants share more intimate information with their high-rapport partners and do not share it with their low-rapport partners.

D. LIMITATIONS

As we have seen, the ranking performance of our model in late interactions is less than the performance in early interactions. One explanation for the results is that our models cannot capture cues of coordination that are important for building rapport in late interactions. To capture cues of coordination, we need to consider interspeaker influences in interactions. To use interspeaker influences, researchers in emotion recognition in conversation (ERC) developed models that use neural network architectures, such as recurrent networks [50] and graph convolutional networks [51], [52]. Although these models achieve state-of-the-art performance in multiple datasets for utterance-level emotion recognition, the models cannot be applied to conversation-level rapport recognition without alterations. Further studies of the model architecture, therefore, are required to capture cues indicating coordination.

We have not conducted a detailed analysis of the behavioral patterns for each participant according to their conversation partners with different rapport levels because it is beyond the scope of our study. However, the findings from such analyses are important not only for social signal processing but also for social psychology. A recent study [53] showed that the relationship between behavior and rapport levels is nonlinear and complex. Tickle-Degnen [54] suggested that "optimal" levels of expressivity and coordination should bring pairs of participants high levels of rapport. Although there are many studies on levels of rapport and behavior patterns (e.g., [10]), there is room for further investigation into how the same participants change their behavior according to their conversation partners with different rapport levels.

VII. CONCLUSION

This study addressed the novel task of ranking conversation partners based on self-reported rapport levels (RACOP). Furthermore, we created a new dataset for RACOP. First, we evaluated the ranking model trained via the preference learning (PL) framework. The results showed that PL is a more suitable approach for RACOP than regression. The results also suggested that a margin threshold improves the reliability of the training and validation sets. Second, we investigated the effect of modality on RACOP. The results indicated that acoustic features are more effective than linguistic features in RACOP. Moreover, multimodal features are most effective for PL models. Finally, we reported that the PL model predicts ordered lists more accurately in early interactions than in late interactions. The results suggested that further studies of the model architecture are required to encode cues of coordination in late interactions.

REFERENCES

[1] L. Huang, L.-P. Morency, and J. Gratch, "Virtual rapport 2.0," in *Intelligent Virtual Agents*. Berlin, Germany: Springer, 2011, pp. 68–79.

[2] T. Sinha and J. Cassell, "We click, we align, we learn: Impact of influence and convergence processes on student learning and rapport building," in *Proc. 1st Workshop Modeling Interpersonal Synchrony Influence*. ACM, Nov. 2015, pp. 13–20.

[3] J. Nadler, "Rapport: Rapport in negotiation and conflict resolution," *Marquette Law Rev.*, vol. 87, no. 4, p. 25, 2004.

[4] J. L. Hagad, R. Legaspi, M. Numao, and M. Suarez, "Predicting levels of rapport in dyadic interactions through automatic detection of posture and posture congruence," in *Proc. IEEE 3rd Int. Conf. Privacy, Secur., Risk Trust IEEE 3rd Int. Conf. Social Comput.*, Oct. 2011, pp. 613–616.

[5] A. Cerekovic, O. Aran, and D. Gatica-Perez, "Rapport with virtual agents: What do human social cues and personality explain?" *IEEE Trans. Affect. Comput.*, vol. 8, no. 3, pp. 382–395, Jul. 2017.

[6] Y. Matsuyama, A. Bhardwaj, R. Zhao, O. Romeo, S. Akoju, and J. Cassell, "Socially-aware animated intelligent personal assistant agent," in *Proc. 17th Annu. Meeting Special Interest Group Discourse Dialogue*, 2016, pp. 224–227.

[7] H. P. Martínez, G. N. Yannakakis, and J. Hallam, "Don't classify ratings of affect; rank them!" *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 314–326, Jul. 2014.

[8] R. Lotfian and C. Busso, "Practical considerations on the use of preference learning for ranking emotional speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5205–5209.

[9] S. Parthasarathy, R. Lotfian, and C. Busso, "Ranking emotional attributes with deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 4995–4999.

[10] L. Tickle-Degnen and R. Rosenthal, "The nature of rapport and its nonverbal correlates," *Psychol. Inquiry*, vol. 1, no. 4, pp. 285–293, Oct. 1990.

[11] F. J. Bernieri, J. S. Gillis, J. M. Davis, and J. E. Grahe, "Dyad rapport and the accuracy of its judgment across situations: A lens model analysis," *J. Personality Social Psychol.*, vol. 71, no. 1, pp. 110–129, Jul. 1996.

[12] J. E. Grahe and F. J. Bernieri, "The importance of nonverbal cues in judging rapport," *J. Nonverbal Behav.*, vol. 23, no. 4, pp. 253–269, Dec. 1999.

[13] N. Wang and J. Gratch, "Rapport and facial expression," in *Proc. 3rd Int. Conf. Affect. Comput. Intell. Interact. Workshops*, Sep. 2009, pp. 1–6.

[14] P. Müller, M. X. Huang, and A. Bulling, "Detecting low rapport during natural interactions in small groups from non-verbal behaviour," in *Proc. 23rd Int. Conf. Intell. User Interfaces*. ACM, Mar. 2018, pp. 153–164.

[15] R. Zhao, T. Sinha, A. W. Black, and J. Cassell, "Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior," in *Intelligent Virtual Agents* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, 2016, pp. 218–233.

[16] M. A. Madaio, R. Lasko, J. Cassell, and A. Ogan, "Using temporal association rule mining to predict dyadic rapport in peer tutoring," *Educ. Data Mining*, vol. 10, pp. 1–4, Jan. 2017.

[17] Y. Raphalen, C. Clavel, and J. Cassell, "'You might think about slightly revising the title': Identifying hedges in peer-tutoring interactions," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 2160–2174.

[18] J. Gratch, A. Okhmatovskaia, F. Lamothe, S. Marsella, M. Morales, R. J. van der Werf, and L.-P. Morency, "Virtual rapport," in *Intelligent Virtual Agents*. Berlin, Germany: Springer, 2006, pp. 14–27.

[19] A. Abulimiti, J. Cassell, and J. Ginzburg, "by the way do you like spider man? Towards a social planning model for rapport," in *Proc. 25th Workshop Semantics Pragmatics Dialogue*, Sep. 2021. [Online]. Available: http://semdial.org/anthology/papers/Z/Z21/Z21-3004/

[20] R. Cowie, E. Douglas-Cowie, S. Savvidou, and E. McMahon, "FEELTRACE: An instrument for recording perceived emotion in real time," in *Proc. ISCA Workshop Speech Emotion*, Sep. 2000, pp. 19–24.

[21] R. Likert, "A technique for the measurement of attitudes," *Arch. Psychol.*, vol. 22, p. 140, Jan. 1932.

[22] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *J. Behav. Therapy Experim. Psychiatry*, vol. 25, no. 1, pp. 49–59, Mar. 1994.

[23] H. Yang and C. Lee, "Annotation matters: A comprehensive study on recognizing intended, self-reported, and observed emotion labels using physiology," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2019, pp. 1–7.

[24] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–8.

[25] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions: An emerging approach," *IEEE Trans. Affect. Comput.*, vol. 12, no. 1, pp. 16–35, Jan. 2021.

[26] Y.-H. Yang and H. H. Chen, "Ranking-based emotion recognition for music organization and retrieval," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 762–774, May 2011.

[27] P. Lopes, A. Liapis, and G. N. Yannakakis, "Modelling affect for horror soundscapes," *IEEE Trans. Affect. Comput.*, vol. 10, no. 2, pp. 209–222, Apr. 2019.

[28] K. Kondo, T. Nakamura, Y. Nakamura, and S. Satoh, "Siamese-structure deep neural network recognizing changes in facial expression according to the degree of smiling," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 4605–4612.

[29] D. Melhart, K. Sfikas, G. Giannakakis, and G. A. Liapis, "A study on affect model validity: Nominal vs ordinal labels," in *Proc. 2nd Workshop Artif. Intell. Affect. Comput.*, vol. 86, W. Hsu and H. Yates, Eds., Jul. 2020, pp. 27–34.

[30] G. Zoumpourlis and I. Patras, "Pairwise ranking network for affect recognition," in *Proc. 9th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2021, pp. 1–8.

[31] A. Carpenter and K. Greene, "Social penetration theory," in *The International Encyclopedia of Interpersonal Communication*. Hoboken, NJ, USA: Wiley, Dec. 2015.

[32] I. Altman and D. A. Taylor, *Social Penetration: The Development of Interpersonal Relationships*. New York, NY, USA: Holt, Rinehart & Winston, 1973.

[33] L. A. Baxter and W. W. Wilmot, "Taboo topics in close relationships," *J. Social Pers. Relationships*, vol. 2, no. 3, pp. 253–269, Sep. 1985.

[34] M. Kimura, M. Yogo, and I. Daibo, "Expressivity halo effect in the conversation about emotional episodes," *Jpn. J. Res. Emotions*, vol. 12, no. 1, pp. 12–23, 2005.

[35] R. B. Zajonc, "Attitudinal effects of mere exposure," *J. Pers. Soc. Psychol.*, vol. 9, no. 2, pp. 1–27, Jun. 1968.

[36] K. L. Zink, M. Perry, K. London, O. Floto, B. Bassin, J. Burkhardt, and S. A. Santen, "'Let me tell you about my…' Provider self-disclosure in the emergency department builds patient rapport," *Western J. Emerg. Med.*, vol. 18, no. 1, pp. 43–49, Jan. 2017.

[37] J. J. Cameron, J. G. Holmes, and J. D. Vorauer, "When self-disclosure goes awry: Negative consequences of revealing personal failures for lower self-esteem individuals," *J. Experim. Social Psychol.*, vol. 45, no. 1, pp. 217–222, Jan. 2009.

[38] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *Proc. 22nd Int. Conf. Mach. Learn.* ACM, 2005, pp. 89–96.

[39] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSmile: The Munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM Int. Conf. Multimedia*. ACM, Oct. 2010, pp. 1459–1462.

[40] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Apr. 2016.

[41] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186.

[42] W. Wei, S. Li, S. Okada, and K. Komatani, "Multimodal user satisfaction recognition for non-task oriented dialogue systems," in *Proc. Int. Conf. Multimodal Interact.* ACM, Oct. 2021, pp. 586–594.

[43] S. Katada, S. Okada, and K. Komatani, "Effects of physiological signals in different types of multimodal sentiment estimation," *IEEE Trans. Affect. Comput.*, early access, Mar. 3, 2022, doi: 10.1109/TAFFC.2022.3155604.

[44] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. ACM, Jul. 2019, pp. 2623–2631.

[45] A. Pentland, *Honest Signals: How They Shape Our World*. Cambridge, MA, USA: MIT Press, 2008.

[46] D. A. Kenny, *Interpersonal Perception: The Foundation of Social Relationships*. New York, NY, USA: Guilford Publications, Nov. 2019.

[47] R. Mogan, R. Fischer, and J. A. Bulbulia, "To be in synchrony or not? A meta-analysis of synchrony's effects on behavior, perception, cognition and affect," *J. Experim. Social Psychol.*, vol. 72, pp. 13–20, Sep. 2017.

[48] I. M. Vicaria and L. Dickens, "Meta-analyses of the intra- and interpersonal outcomes of interpersonal coordination," *J. Nonverbal Behav.*, vol. 40, no. 4, pp. 335–361, Dec. 2016.

[49] M. Natale, "Convergence of mean vocal intensity in dyadic communication as a function of social desirability.," *J. Personality Social Psychol.*, vol. 32, no. 5, pp. 790–804, Nov. 1975.

[50] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: An attentive RNN for emotion detection in conversations," in *Proc. 33rs AAAI Conf. Artif. Intell. 31st Innov. Appl. Artif. Intell. Conf. 9th AAAI Symp. Educ. Adv. Artif. Intell.*, Jan. 2019, pp. 6818–6825.

[51] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 154–164.

[52] T. Ishiwatari, Y. Yasuda, T. Miyazaki, and J. Goto, "Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 7360–7370.

[53] A. A. Nelson, J. E. Grahe, and F. Ramseyer, "Interacting in flow: An analysis of rapport-based behavior as optimal experience," *SAGE Open*, vol. 6, no. 4, Oct. 2016, Art. no. 2158244016684173.

[54] L. Tickle-Degnen, "Nonverbal behavior and its functions in the ecosystem of rapport," in *The SAGE Handbook of Nonverbal Communication*, vol. 587, V. Manusov, Ed. Thousand Oaks, CA, USA: SAGE Publications, 2006, pp. 381–399.

TAKATO HAYASHI received the B.S. degree in agriculture from Tottori University, in 2021. He is currently pursuing the M.S. degree with the Japan Advanced Institute of Science and Technology (JAIST). His research interests include social signal processing, multimodal interaction, and machine learning.

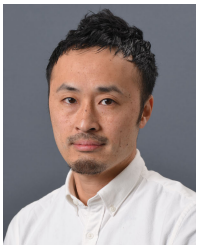
CANDY OLIVIA MAWALIM received the B.S. degree in computer science from Institut Teknologi Bandung (ITB), Bandung, Indonesia, and the M.S. and Ph.D. degrees from the School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), in 2019 and 2022, respectively. She was selected as a Research Fellow for young scientists DC1 (JSPS), from 2020 to 2022. Since April 2022, she has been an Assistant Professor with the Social Signal Interaction Group, School of Information Science, JAIST. Her main research interests include speech signal processing, hearing perception, voice privacy, and machine learning.

RYO ISHII received the M.S. degree in engineering from the Tokyo University of Agriculture and Technology, Japan, and the Ph.D. degree in informatics from Kyoto University, in 2013. In 2008, he joined NTT Corporation, where he is currently a Distinguished Researcher with Human Informatics Laboratories. His research interests include multimodal interaction and social signal processing. He is a member of IEICE, JSAI, and HIS.

AKIRA MORIKAWA received the M.S. degree in engineering from Kobe University. In 2011, he joined NTT Corporation, where he is currently a Researcher Engineer with Human Informatics Laboratories. His research interests include multimodal interaction and security.

ATSUSHI FUKAYAMA graduated from the Graduate School of Informatics, Kyoto University, in 1999. Then, he joined Nippon Telegraph and Telephone Corporation. After working on media recognition technology, research and development of human–computer interaction technology, and practical application development of network services, he has been leading another the Me Research Group, NTT Digital Twin Computing Research Center, since 2021.

TAKAO NAKAMURA received the B.S. degree in mathematics from Waseda University, Tokyo, in 1994, and the Ph.D. degree in informatics from the Graduate University for Advanced Studies, Kanagawa, in 2008. In 1994, he joined Human Informatics Laboratories, NTT Corporation, and studied media processing technologies, content distribution systems, artificial intelligence, and their applications. He is currently the Director of the NTT Digital Twin Computing Research Center, and involved in research and development of the concepts and technologies for digital twin computing.



SHOGO OKADA (Member, IEEE) received the Ph.D. degree from the Tokyo Institute of Technology, Japan, in 2008. In 2008 and 2011, he joined the Tokyo Institute of Technology, Kyoto University, as an Assistant Professor. In 2014, he joined the Idiap Research Institute, Switzerland, as a Visiting Faculty Member. He directs the Social Signal and Interaction Group, Japan Advanced Institute of Science and Technology (JAIST), Japan, where he is currently an Associate Professor. His research interests include social signal processing, human dynamics, multimodal interaction, and machine learning. He is a member of the ACM.