Speaker Intimacy in Chat-Talks: Analysis and Recognition based on Verbal and Non-Verbal Information

Yuya Chiba[1], **Yoshihiro Yamazaki**[2], **Akinori Ito**[2]
[1]NTT Communication Science Laboratories, Japan
[2]Graduate School of Engineering, Tohoku University, Japan
{yuuya.chiba.ax@hco.ntt.co,yoshihiro.yamazaki.t2@dc.tohoku.ac,
aito@spcom.ecei.tohoku.ac}.jp

Abstract

Conversations based on mutual intimacy are critical for maintaining positive relationships. Conversational AIs, which are widely spread in society, are assumed to be continuously used by users in daily life. To establish long-term relationships with users, AI systems have to handle dialogues based on an awareness of user intimacy. In this study, we experimentally examined a method to estimate a speaker's intimacy to a dialogue partner in chat-talks. We used a multimodal human-human conversation corpus of 71 Japanese participants. The corpus contains metadata related to subjective intimacy score of speakers. First, we identified the effective features to estimate the speaker's intimacy by comparing the statistical parameters of the features. Then, we proposed a model to estimate the speaker's intimacy by observing the several utterances.

1 Introduction

Conversational AIs, represented by the smart speakers, are widely used in daily life. Such systems are assumed to be continuously used by users, and strategies for maintaining and developing long-term relationships with users is becoming more important. However, current dialogue systems cannot take strategy to maintain a friendly relationship with users. In this situation, the system sometimes discourages the users by responding to them disinterestedly even they talk in a friendly manner.

To establish a long-term relationship, key roles include a sense of closeness and intimacy resulting from social conversations (Bickmore et al., 2005; Cassell and Bickmore, 2003). In human-human conversations, participants express intimacy with dialogue partners by such behaviors as speaking style, facial expressions, and posture (Hornstein, 1985; Planalp, 1993). Therefore, sociable conversational agents are required to manage verbal and non-verbal behavior to build a friendly relationship with users. There is also a lot of discussion about building relationships between humans. For example, the social penetration theory (Altman and Taylor, 1973) and Knapp's relationship model (Knapp et al., 2014) explained the development of relationships as a mutual process that gradually progresses.

In human-machine conversation, a few studies have developed systems that convey intimacy to users, and effectively improve user impressions. For example, Bickmore et al. (2005) built a relational agent that introduced "immediacy" behavior (Argyle, 1988) that supports multiple interactions with users over an extended period of time. Kageyama et al. (2018) evaluated dialogue systems by changing speech styles. Similarly, Kanda et al. (2009) developed for a shopping mall guide robot that changes its behavior. Kim et al. (2013) confirmed that systems that greet the users by name are perceived as friendly. However, these above studies assumed that relationships develop in association with a particular number of accumulated interactions and that they are unilaterally replicated. To achieve dialogue management that reflects a development of relationship, a system should show intimacy to users, and simultaneously recognize intimacy from them.

In this paper, we propose a method to estimate the level of intimacy of speakers to achieve systems that engage in conversations based on mutual intimacy. We target chat-talks because such conversations play an important role in establish interpersonal relationships (Rich, 1979). In human-human dialogues, dialogue behaviors based on intimacy have been investigated by analysis based on annotation. However, it remains unclear whether such information can be extracted as features from audio-visual signals. Therefore, we first identify the features that are useful for estimating speaker intimacy by comparing statistical parameters of them. Then, we constructed a speaker intimacy estimation model using multimodal information. Our proposed model discriminates among tree levels of speaker intimacy by observing several utterances.

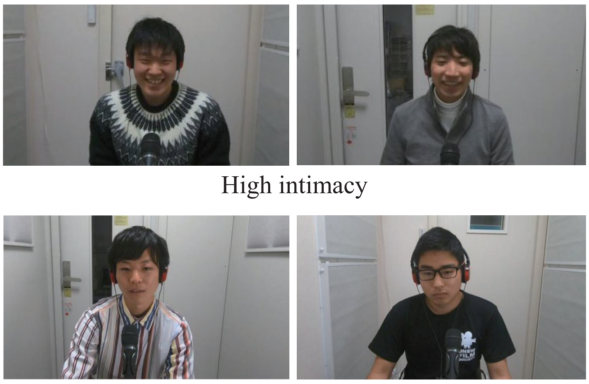
2 Related Studies

2.1 Dialogue Behavior based on Intimacy

Behaviors related to interpersonal relationships has been discussed from various perspectives. According to social penetration theory (Altman and Taylor, 1973), self-disclosure, which intentionally reveals personal information, becomes more frequent and deeper as relationships develop. Hornstein (1985) and Yamazaki et al. (2020) reported that the choice of speech intention is affected by relationships. Hall (1963) explained that the attitudes of participants change based on their interpersonal relationships. Mutual imitation is also considered to be an expression of friendship and preference. The entrainment of acoustic and prosodic features is correlated with a rapport between speakers (Lubold and Pon-Barry, 2014). The chameleon effect (Chartrand and Bargh, 1999), which is the mimicry of facial expressions and posture, is a similar phenomenon.

In addition, some studies have analyzed dialogue behaviors by focusing on such relationship stages, as friend, acquaintance, and confidant. For example, Hornstein (1985) concluded that friends use more implicit openings, raise topics, and express more responsiveness to each another by asking questions. The floor time distribution or the number of interruptions (Planalp, 1993) and various activities (Rands and Levinger, 1979) also changes according to a step of the relationships. In terms of rapport, Grahe and Bernieri (1999) reported that participants are likely to sustain longer eye contact, smile more, and lean more toward each another when building rapport.

The analysis described by these studies is based on self-reports or human annotation. It is not clear that such information can be extracted from audiovisual signals as effective features for intimacy estimation. In this paper, we investigated effectiveness of multimodal features by comparing the statistical parameters among levels of speaker's intimacy.



High intimacy

Low intimacy

Figure 1: Examples of high and low intimacy dialogue in SMOC.

2.2 Interpersonal Relationship Recognition

The estimation of interpersonal relationships has been examined by several media. Zhang et al. (2018) predicted interpersonal relationships between people in images based on facial expressions. Chu et al. (2015) proposed an immediacy prediction model using posture-based features. User profiles and sentence lengths are effective cues in SNS and e-mail interactions (Nishihara and Sunayama, 2009; Xiong et al., 2016). In human-robot conversations, Kanda and Ishiguro (2004) estimated interpersonal relationships between the participants based on interaction time. Although these studies focused on interpersonal relationship, they did not use conversational information. We are building an estimation model that is useful in various conversational situations by incorporating the aforementioned conversational information.

The most relevant research to our study is Soleymani et al. (2019), which estimated the intimacy levels of verbal self-disclosure in interview dialogues using multimodal information. In contrast, we focus on speaker intimacy in chat-talks, and do not limit the target to self-disclosure. In addition, our model is designed to incorporate such interaction between speakers as entrainment and synchrony.

3 Spontaneous Multimodal One-on-One Chat-Talk Corpus

3.1 Overview of Corpus

We used a Spontaneous Multimodal One-on-one Chat-talk (SMOC) corpus (Yamazaki et al., 2020) for the experiments. The target corpus contains the

Table 1: Summary of experimental data of this paper.

No. dialogues	345
No. recordings	690
No. pairs	69
No. speakers	71 (females: 19, males: 52)
No. utterances	23,379

audio and video of one-on-one dialogues between Japanese participants. The dialogues were conducted through by video communication between sound-proof chambers at close distance without a time lag. The audio data have no crosstalk and the video was recorded from the front of the speaker. Figure 1 shows examples of dialogue scenes between speakers with high and low intimacy.

3.2 Recording Conditions

Two participants were paired up and engaged in chat-talks. The dialogues were conducted by both acquainted and unacquainted pairs. The participants engaged in the dialogues to build a relationship with their partners. Each participant talked about five topics with two different speakers. One of the examples of topics is "My favorite foods and beverages, and the ones I don't like." Each topic lasted about 20 minutes. The speech was recorded by microphones (AT4055), and the facial expressions and gestures were recorded by video cameras (GoPro HERO7 Black) in front of the speakers. The captured video and audio data were send to the display and headphones in another sound-proof chamber through video connection for multimodal communication. The speakers talked with the partner while looking each other through the monitor. The audio data were stored with 16 kHz sampling and 16-bit quantization. The video data were recorded at 1920×1080 resolution and in a 59.94-fps MP4 format. The dialogue data have transcriptions with time-information of the beginning and ending of the utterance determined by phoneme-alignment.

We used 345 dialogues of 69 pairs recorded at an early stage of corpus construction. The total number of the dialogue recordings was 690 (345 dialogues × 2 participants). We summarized the number of the data in Table 1. The data were split into utterances based on time information.

3.3 Labels of Subjective Intimacy

The corpus has the metadata about speaker's intimacy with his/her dialogue partner. Before the conversation, each participant was asked the following questions: 1) Do you know your dialogue partner?, 2) How long have you known him/her, and 3) How close do you feel to your dialogue partner? The second and third questions were only answered by the participants who answered "yes" to the question 1). For the third question, the participants rated intimacy on a 5-grade scale, from one (not at all) to five (very much).

In this paper, we used the answer of the third question for the labels of subjective intimacy to his/her dialogue partner. The intimacy score of the participants of unacquainted pairs was set to 0. The number of dialogue recordings of score 0 was 280. Among the acquainted pairs, the numbers of the dialogue recordings rated three, four, and five were 100, 130, and 180, respectively. No participant rated less than two.

4 Analysis of Multimodal Features based on Intimacy

In this section, we analyzed the SMOC corpus based on subjective intimacy to identify the effective features for estimation. We extracted linguistic, acoustic, and visual features.

4.1 Word Frequency Distribution

First, we compared the word frequency distribution between different subjective intimacy scores. The utterances were segmented using MeCab[1] (Kudo, 2006), which is a Japanese morphological analyzer, with the NEologd dictionary[2]. We constructed Bag-of-Words (BoW) vectors for each score, and visualized the distance between them by multidimensional scaling. The result is shown in Figure 2. The figure shows the distribution of words of each score were roughly separated into three clusters: a group of scores 4 and 5, score 3, and score 0.

One reason why word frequency distribution is different between the groups is the influence of "honorifics." The target corpus's language, Japanese, has a clear honorific mechanism. The speech style changes based on the relative social position or closeness of the social distance to the dialog partners. In Japanese, the honorifics is often expressed by the auxiliary verb of the ending of the utterance. For example, the verb *taberu* (to eat) can be transformed to *tabe-masu* to express the honorifics. Here, we focused on *desu* and

[1] http://mecab.sourceforge.jp

[2] https://github.com/neologd/mecab-ipadic-neologd

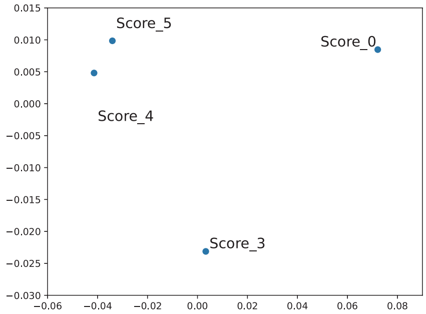


Figure 2: Visualization of distance between BOW vectors of subjective intimacy scores based on multidimensional scaling.

Table 2: Difference of average frequency of DAs relating "Question" between intimacy levels ($*p < 0.05$).

Category	Comparison	Diff.	t	p-value
Information	Low − Mid.	0.20	1.09	0.82
	Low − High	−0.30	−2.32	0.06
	Mid. − High	−0.50	−2.77	0.02*
Fact	Low−Mid.	0.54	2.36	0.06
	Low−High	0.46	2.80	0.02*
	Mid.−High	−0.09	−0.38	1.00
Experience	Low−Mid.	0.10	2.59	0.03*
	Low−High	0.12	4.07	< 0.00*
	Mid.−High	0.01	0.29	1.00
Habit	Low−Mid.	0.08	2.65	0.03*
	Low−High	0.12	5.29	< 0.00*
	Mid.−High	0.03	1.11	0.81
Desire	Low−Mid.	0.01	0.43	1.00
	Low−High	0.05	2.52	0.04*
	Mid.−High	0.03	1.37	0.51

masu, which are among the most basic auxiliary verbs, to express honorifics in Japanese. When we investigated the word usage rate of the target corpus, the order of the use of these auxiliary verbs decreased monotonically as the intimacy scores increased. Such a lexical change does not necessarily exist only in Japanese. Some languages change how to address someone to convey attitudes toward the addressee (i.e., T-forms and V-forms (Brown and Gilman, 1960)). Therefore, lexical features are effective features for estimation even in other languages.

Based on the above analysis, we separated the dialogue behavior of the participants into three classes: score 0 as "low intimacy," score 3 as "middle intimacy," and scores 4 and 5 as "high intimacy."

4.2 Dialogue Acts

Next, we compared the average frequency of the dialogue acts (DA) among levels of intimacy. DAs were extracted using Richindexer[3]. The kind of the DAs is the same with Meguro et al. (2010). First, we conducted a one-way layout ANOVA that factored the level of intimacy. Then, we conducted multiple comparison tests with Bonferroni correction for each DA that showed a significant difference by ANOVA. In this paper, we focus on DAs related "Question" due to space limitation. We summarized the results for sub-categories of "Question" in Table 2.

The table showed that such questions as the facts, experience, and habits frequently appeared in the group of low intimacy. It is adequate that these questions tend to appear at the early stage of relationships when the participants are getting to know each other. The trends of the differences varied by sub-categories, although the total number of questions decreased. These results coincide with a conventional study (Yamazaki et al., 2020), although they partially contradict another report (Hornstein, 1985). Hornstein (1985) concluded that friends were responsive to the partner by asking more questions. One possible reason of the difference is cultural differences. Although verification of the cultural difference is not a purpose of this paper, we plan to compare the DAs using other language corpora in future studies.

4.3 Entrainment of Prosody

In the analysis of acoustic features, we focused on the interaction between speakers. We extracted the prosodic features, the speaking rate, and the switching pauses from the utterances and calculated the entrainment. For the prosodic features, the maximum and the mean of the log F0 and intensity were selected based on a previous study (Kawahara et al., 2015). Here, Levitan and Hirschberg (2011) proposed a quantification method for three kinds of entrainment: proximity, convergence, and synchrony. We focused on the proximity of the acoustic features. Turn-level proximity evaluates how close the utterance's acoustic feature is to that of the preceding interlocutor's utterance. Concretely, the proximity is expressed by the absolute difference of the average of the feature between adjacent utterances. In the analysis, we compared the average proximity over dialogue among intimacy levels.

First, we conducted a one-way layout ANOVA that factored the level of the intimacy and then

[3] https://www.rd.ntt/e/research/MD0057.html

Table 3: Difference of average proximity of prosodic features: A positive difference indicates that the right group has the large entrainment ($^*p < 0.05$).

Features	Comparison	Diff.	t	p-value
Speaking rate	Low−Mid.	1.50	1.65	0.30
	Low−High	1.78	2.77	0.02*
	Mid.−High	0.28	0.32	1.00
Switching pause	Low−Mid.	0.03	0.78	1.00
	Low−High	0.04	4.38	< 0.00*
	Mid.−High	0.00	2.39	0.05
max. f0	Low−Mid.	−0.01	−1.83	0.21
	Low−High	−0.01	−2.51	0.04*
	Mid.−High	0.00	−0.05	1.00
mean f0	Low−Mid.	−0.01	−1.48	0.42
	Low−High	−0.01	−3.66	< 0.00*
	Mid.−High	−0.01	1.12	0.79

a multiple comparison test with Bonferroni correction. Table 3 shows the results of the multiple comparison tests for the features that obtained significant differences by ANOVA. The speaking rate and switching pause were significantly different between low and high intimacies, indicating that speakers who feel more intimacy to the interlocutor tend to synchronize their speaking rates and switching pauses. In terms of log F0, the entrainment was larger in the group of the low-level intimacy against our expectations. The global features calculated from the entire utterance may be too coarse to capture the entrainment. In future studies, we will calculate the entrainment using the features obtained from the beginning and ending segments of the utterance as same with the previous study (Kawahara et al., 2015).

4.4 Facial Expression Synchrony

For visual cues, the features expressed during the dialogue partner talking to are also important. Thus, we focused on the synchrony of facial expressions. Facial Action Units (AU) were extracted using OpenFace (Baltrušaitis et al.). Levitan and Hirschberg (2011) quantified the synchrony by the correlation coefficient between the features of interlocutors. Here, the intimacy scores of the target corpus differ from speaker by speaker. The correlation coefficient cannot be calculated by a three-level classification because the scores may be different between speakers of the same dialogue. Therefore, we compared the unacquainted group (U) and acquainted group (A) (i.e., score 0 and others).

Table 4 shows the results of a Welch's t test between the two groups when we compared the average synchrony over dialogue. AU02 (Outer Brow Raiser), AU06 (Cheek Raiser), and AU12 (Lip Corner Puller) were significantly larger in the acquainted group. In particular, the AU06 and AU12 features increase when the speaker expresses a smile. Therefore, it is indicated that the smile tends to co-occur in acquainted groups. In contrast, no significant differences were observed in other AUs since the other facial expressions are less likely to appear in the target dialogue.

Table 4: Difference of average synchrony of action units: U and A represent unacquainted and acquainted pairs. A negative difference indicates that the acquainted pairs have the large entrainment ($^*p < 0.05$).

Action Unit	$U - A$	t	p-value
AU01	-1.493×10^{-2}	−1.319	0.188
AU02	-3.386×10^{-2}	−2.501	0.013*
AU04	0.578×10^{-2}	0.496	0.621
AU05	-0.217×10^{-2}	−0.181	0.857
AU06	-9.453×10^{-2}	−5.884	< 0.000*
AU07	-2.107×10^{-2}	−1.560	0.120
AU09	0.058×10^{-2}	0.036	0.971
AU10	-1.275×10^{-2}	−0.830	0.407
AU12	-8.044×10^{-2}	−4.888	< 0.000*
AU14	-2.044×10^{-2}	−1.192	0.234
AU15	0.383×10^{-2}	0.336	0.737
AU17	-1.279×10^{-2}	−0.994	0.321
AU20	-0.415×10^{-2}	−0.440	0.660
AU23	-2.133×10^{-2}	−1.870	0.062
AU25	2.197×10^{-2}	1.353	0.177
AU26	-0.163×10^{-2}	−0.126	0.900
AU28	1.382×10^{-2}	1.424	0.158
AU45	0.779×10^{-2}	0.680	0.497

4.5 Gaze Activity

Finally, we investigated the gaze actions. For gaze, it is reported that the participants with high rapport are likely to sustain eye contact longer Grahe and Bernieri (1999). Therefore, we focused on gaze variations. Let the gaze angle at time t be $\mathbf{g}_t = (x_t, y_t)$, the gaze variation is represented as:

$$\Delta \mathbf{g} = \frac{1}{T-1} \sum_{t=2}^{T} ||\mathbf{g}_t - \mathbf{g}_{t-1}||. \qquad (1)$$

x_t and y_t are the gaze angles of the horizontal and vertical axes obtained using OpenFace. T is the number of frames of each dialogue. We extracted the gaze variation from every utterance, and compared the average gaze variation over dialogue.

First, we conducted a one-way layout ANOVA that factored the level of intimacy and obtained a significant difference ($p < 0.001$). Then, a multiple comparison test with Bonferroni correction was conducted. Table 5 shows the results. Significant differences were observed between "Low"

Table 5: Difference of average gaze variance: A positive difference indicates that the right group has a small variation ($^*p < 0.05$).

Comparison	Diff.	t	p-value
Low−Mid.	-0.010×10^{-3}	0.008	1.000
Low−High	1.097×10^{-3}	3.854	$< 0.001^*$
Mid.−High	1.107×10^{-3}	2.881	0.012^*

and "High," and "Mid." and "High." The gaze variation significantly decreased in the high intimacy group. This result suggests that the gaze feature is effective to estimate the speaker's intimacy.

5 Intimacy-Level Estimation Network

In the following sections, we examined an intimacy recognition method that reflects our analysis. Figure 3 shows our proposed network. As shown in the analysis, such interactions between speakers as the entrainment seem important to intimacy estimation. Therefore, the proposed intimacy recognition model takes continuous utterances as input. Here, $l_{t,n}$ and $a_{t,n}$ are the linguistic and acoustic features at time n of the t-th utterance. In addition, $v^i_{t,n}$ is the visual feature of speaker $i \in (s, p)$. s and p represent the speaker and the dialogue partner of respective utterances. N^l_t, N^a_t, and N^v_t are the length of the linguistic, acoustic, and visual feature sequences. y_t is the prediction result.

First, the network extracts the verbal and non-verbal features every utterance and encode them to the representation vectors. We employed the multi-stream attention-based BLSTM (Chiba et al., 2020) as an utterance-level encoder. In this method, the feature sequence of the respective modality is input to the individual attention-based BLSTM (Mirsamadi et al., 2017). Then, the concatenation of the representation vectors of each modality is sent to the linear layer. From these processes, the multi-stream BLSTM fuses the utterance-level multimodal information. For the visual features, it is important to represent the correlation of features (i.e., synchrony) between both speakers. Therefore, we feed the visual features of both speakers to the network. The fully-connected layer is connected after the input layers of the speaker's and the partner's visual features for dimensional reduction.

In addition, our analysis showed that the entrainment between the preceding and current utterances is important for the acoustic features. Thus, we used BLSTM for the succeeding layers of the utterance encoder to capture the relationship between utterances. The context BLSTM takes the representation vectors of continuous T utterances as input, and its output is input to the single fully-connected layer to obtain the prediction result.

5.1 Feature Extraction

The network takes the word sequence as input for the linguistic feature. Each word was converted to a 300-dimensional embedded vector using FastText (Joulin et al., 2016). As acoustic features, the eGeMAPS (Eyben et al., 2015) were extracted with 10-ms frame-shift and 20-ms frame-width. The eGeMAPS features include not only such prosodic features as pitch and loudness but also spectral features. We used the 46-dimensional features, including the Δ features. On the other hand, we used OpenFace (Baltrušaitis et al.) to extract the visual features. We used the features relating to AU, gaze direction, and face direction for the experiments. As same with the acoustic features, the Δ features were calculated for the visual features. The number of dimensions of the visual features was 80. For the visual features, the features of both speakers at the target utterance segment were extracted.

Analysis suggests that the speaking rate or gaze variance were effective for recognition. Therefore, we employed segmental features for the acoustic and visual features to enhance such information. Segmental features were obtained by calculating the statistics (e.g., mean, variance, and range) of the above features every 200 ms. Such statistics are useful to explain speaking rate and gaze variance at the local segment. We employed 12 kinds of statistics same with the Schuller et al. (2009). The number of dimensions of the definitive features were 552 for the acoustic features ($46 \times 12 = 552$) and 960 for the visual features per speaker ($80 \times 12 = 960$).

6 Experiments

6.1 Setup of Experimental Data

As in the case of analysis, we used the SMOC corpus for the recognition experiment as well. We separated the utterances of 345 dialogues into training, development, and test sets so that any of two sets do not share the same speaker. The training, development, and evaluation data were 16,314, 3,465, and 3,600 utterances, respectively.

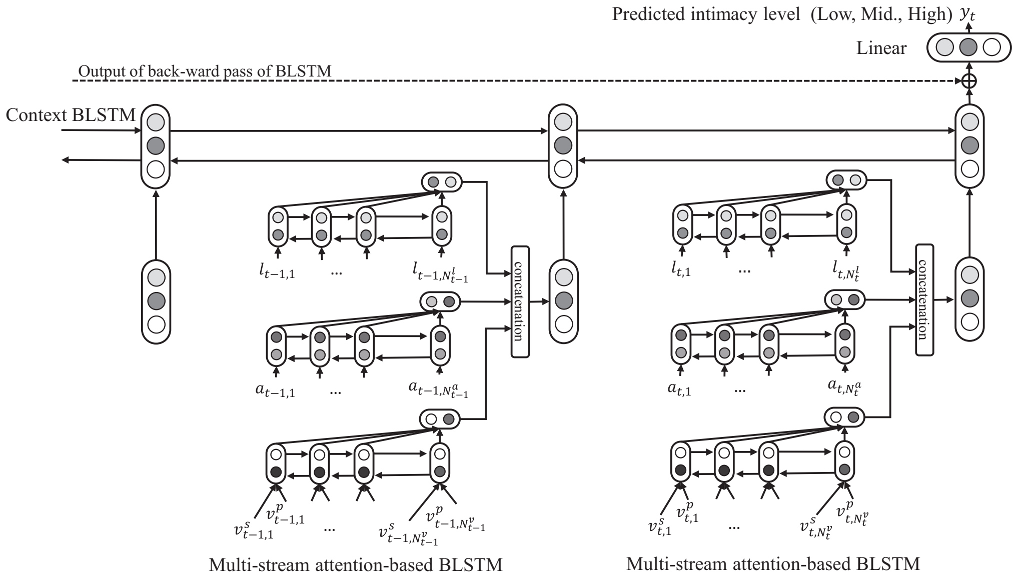

Predicted intimacy level (Low, Mid., High) y_t
Linear
Output of back-ward pass of BLSTM
Context BLSTM
concatenation
concatenation
$l_{t-1,1}$ ... $l_{t-1,N^l_{t-1}}$
$a_{t-1,1}$ ... $a_{t-1,N^a_{t-1}}$
$v^s_{t-1,1}$ $v^p_{t-1,1}$ ... $v^s_{t-1,N^p_{t-1}}$ $v^p_{t-1,N^p_{t-1}}$
$l_{t,1}$ ... l_{t,N^l_t}
$a_{t,1}$ ... a_{t,N^a_t}
$v^s_{t,1}$ $v^p_{t,1}$ ... v^s_{t,N^p_t} v^p_{t,N^p_t}
Multi-stream attention-based BLSTM
Multi-stream attention-based BLSTM

Figure 3: Network architecture for intimacy recognition: $l_{t,n}$ and $a_{t,n}$ are linguistic and acoustic features at frame n of t-th utterance. $v^i_{t,n}$ is the visual features of participant $i \in (s,p)$. s and p represent the speaker and the partner, respectively. N^l_t, N^a_t, and N^v_t are sequence length of linguistic, acoustic, and visual features. y_t is prediction result. $\oplus$ shows the summation.

The intimacy labels of the SMOC corpus were appended to the dialogues. We assigned the same label as the original dialogue to the utterances.

6.2 Conditions of Training Network

We conducted three-class discrimination among low, middle, and high levels of intimacy. The experimental data have a bias toward the distribution of intimacy levels, and we employed weighted cross-entropy loss to train the network. The losses in each class were multiplied by a weight that is proportional to the inverse of the sample size.

The numbers of hidden units were common among the multi-stream BLSTM, the context BLSTM, and fully-connected layers. We investigated the classification performance while changing the number of nodes of the hidden layers to 16, 32, 64, and 128. We used the condition that yielded the best accuracy for the validation set for the definitive evaluation. The number of layers of each BLSTM was 1. We connected the dropout layers after the output of each stream and the context BLSTM. The dropout rate was set to 0.3. The optimization method was Adam with a learning rate of 0.0005. The mini-batch size was 32 and the maximum number of epochs was 100. In the following sections, we show the recognition results for the test set.

7 Experimental Results of Intimacy Recognition

First, we evaluated the effectiveness of the multimodal features. In this experiments, we used continuous four utterances for the classification (i.e., $T = 4$). Table 6 shows the recognition results. A, V, and L denote the acoustic, visual, and linguistic features, respectively. Rec., Pre., and F1. represent the recall, the precision, and the F1-score. Chance shows the results when all test samples are classified to high-level intimacy, which is the most frequent class.

As shown in the table, the results of the proposed models surpassed the chance-level results. The results indicated that the model was adequately trained to estimate the level of intimacy from verbal and non-verbal cues. Comparison of the single modality showed that a higher F1-score was obtained with linguistic information. This result suggests that the utterance styles and the choices of the DA were captured using linguistic features. The combination of audio, visual, and linguistic features improved the performance, and we obtained an F1-score of 0.594. Therefore, the non-verbal information employed in this study was an effective feature to enhance verbal information.

However, the performance of the acoustic and visual information alone did not surpass the lin-

Table 6: Intimacy Recognition Results: A, V, and L denote acoustic, visual, and linguistic features, respectively. Rec., Pre., and F1. represent the recall, precision, and F1-score. Bold fonts are the best performance between modalities. Chance shows results when all test samples are classified as high-level intimacy, which is the most frequent label.

	Low			Middle			High			Macro Average		
Modality	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
A	0.313	0.739	0.439	0.399	0.354	0.375	0.565	0.257	0.353	0.425	0.450	0.389
V	0.262	0.554	0.356	0.165	0.032	0.053	0.568	0.495	0.529	0.332	0.360	0.313
L	**0.857**	0.770	**0.811**	0.272	0.196	0.228	0.652	0.762	0.703	**0.594**	0.576	0.581
A+V	0.275	0.709	0.397	0.499	0.338	0.403	0.587	0.266	0.367	0.454	0.438	0.389
A+L	0.627	**0.835**	0.716	0.469	0.314	0.376	0.671	0.672	0.672	0.589	0.607	0.588
V+L	0.759	0.791	0.775	0.258	0.115	0.159	0.652	**0.801**	**0.719**	0.557	0.569	0.551
A+V+L	0.567	0.811	0.667	**0.506**	**0.432**	**0.466**	**0.693**	0.608	0.648	0.589	**0.617**	**0.594**
Chance	–	–	–	–	–	–	–	–	–	0.177	0.333	0.231

Table 7: Influence of length of dialogue context: Table shows F1-score of classification.

No. utterances N	Low	Mid.	High	Average
2	0.539	0.370	0.556	0.488
4	0.668	0.466	0.648	0.594
6	**0.736**	**0.467**	**0.730**	**0.645**
8	0.716	0.416	0.715	0.616

guistic feature. In particular, the visual information had the lowest estimation performance. This result reflected that the features extracted from OpenFace are insufficient to comprehensively represent non-verbal behavior. Posture and gesture are cues that predict rapport (Grahe and Bernieri, 1999), and we will examine the effectiveness of them for intimacy-level estimation in future studies.

Next, we examined the influence of context length. Table 7 shows F1-scores when changing the length of the dialogue context. As shown in the table, performance improved with a longer dialogue context, and we obtained the best performance at $N = 6$. It is confirmed that our proposed model can estimate the speaker's intimacy to some extent by observing three utterance interchanges. Since the labels were originally assigned to each dialogue, it is considered to be appropriate that the long dialogue context is effective to estimate the speaker's intimacy. On the other hand, performance decreased when the number of interchanges exceeded three (i.e., $T = 8$). One reason for this result is the dialogue data is insufficient. In particular, the data size of the middle-level intimacy was relatively small, and the F1-score did not improve even the network observes the longer context. Therefore, the dialogue data of acquainted pairs that are not close friends should be collected in future studies.

8 Summary and Future Studies

In this paper, we examined the recognition method of speaker intimacy in chat-talks. First, we identified the effective verbal and non-verbal features to estimate subjective intimacy-levels. Then, we developed an intimacy-level estimation model that reflected the analysis results. Our proposed model discriminated user intimacy among low, middle, and high levels. From experiments, we obtained the best F1-score of 0.645 when using the acoustic, visual, and linguistic features. However, some remaining issues must be solved to apply our proposed method to actual dialogue systems.

First, the data used in this study are human-human dialogues, and the behavior of participants might be different in human-machine dialogues. One possible solution is model adaptation. Our proposed network can be adapted to the human-machine dialogues by fine-tuning. In near future, we plan to collect human-machine dialogues based on the wizard-of-Oz basis. Besides, there is a class imbalance problem. In the target dataset, the data size of middle-level intimacy is relatively small, and the performance of this class did not improve. Therefore, collection of dialogue between acquaintance speakers is needed to improve the overall performance of the base model.

In addition, it is crucial that how the system behaves to recognized user intimacy to achieve a dialogue system based on mutual intimacy. Therefore, we next plan to examine a dialogue generation method combining the intimacy-estimation network with a recent response generation model (e.g., (Smith et al., 2020)).

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP20K19903.

References

Irwin Altman and Dalmas Taylor. 1973. *Social penetration: The development of interpersonal relationships*, volume 212. Holt, Rinehart & Winston.

Michael Argyle. 1988. *Bodily Communication*. New York: Methuen & Co, Ltd.

Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. OpenFace: An open source facial behavior analysis toolkit. In *Proc. IEEE Winter Conference on Applications of Computer Vision*, pages 1–10.

Timothy Bickmore, Lisa Caruso, and Kerri Clough-Gorr. 2005. Acceptance and usability of a relational agent interface by urban older adults. In *Proc. CHI*, pages 1212–1215.

Roger Brown and Albert Gilman. 1960. *The pronouns of power and solidarity*. Bobbs-Merrill.

Justine Cassell and Timothy Bickmore. 2003. Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modeling and User-Adapted Interaction*, 13(1):89–132.

Tanya Chartrand and John Bargh. 1999. The chameleon effect: The perception–behavior link and social interaction. *J. Pers. Soc. Psychol.*, 76(6):893–910.

Yuya Chiba, Takashi Nose, and Akinori Ito. 2020. Multi-stream attention-based BLSTM with feature segmentation for speech emotion recognition. In *Proc. INTERSPEECH*, pages 3301–3305.

Xiao Chu, Wanli Ouyang, Wei Yang, and Xiaogang Wang. 2015. Multi-task recurrent neural network for immediacy prediction. In *Proc. ICCV*, pages 3352–3360.

Florian Eyben, Klaus Scherer, Björn Schuller, et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.

Jon Grahe and Frank Bernieri. 1999. The importance of nonverbal cues in judging rapport. *Journal of Nonverbal Behavior*, 23(4):253–269.

Edward Hall. 1963. A system for the notation of proxemic behavior 1. *American Anthropologist*, 65(5):1003–1026.

Gail Hornstein. 1985. Intimacy in conversational style as a function of the degree of closeness between members of a dyad. *J. Pers. Soc. Psychol.*, 49(3):671–681.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. FastText. zip: Compressing text classification models. *arXiv preprint:1612.03651*, pages 1–13.

Yukiko Kageyama, Yuya Chiba, Takashi Nose, and Akinori Ito. 2018. Improving user impression in spoken dialog system with gradual speech form control. In *Proc. SIGDIAL*, pages 235–240.

Takayuki Kanda and Hiroshi Ishiguro. 2004. Friendship estimation model for social robots to understand human relationships. In *Proc. ROMAN*, pages 539–544.

Takayuki Kanda, Masahiro Shiomi, Zenta Miyashita, Hiroshi Ishiguro, and Norihiro Hagita. 2009. An affective guide robot in a shopping mall. In *Proc. HRI*, pages 173–180.

Tatsuya Kawahara, Takashi Yamaguchi, Miki Uesato, Koichiro Yoshino, and Katsuya Takanashi. 2015. Synchrony in prosodic and linguistic features between backchannels and preceding utterances in attentive listening. In *Proc. APSIPA-ASC*, pages 392–395.

Yunkyung Kim, Sonya Kwak, and Myung-Suk Kim. 2013. Am I acceptable to you? Effect of a robot's verbal language forms on people's social distance from robots. *Comput. Human Behav.*, 29(3):1091–1101.

Mark Knapp, Anita Vangelisti, and John Caughlin. 2014. *Interpersonal communication and human relationships*. Pearson.

Taku Kudo. 2006. Mecab: Yet another part-of-speech and morphological analyzer. *http://mecab.sourceforge.jp*.

Rivka Levitan and Julia Hirschberg. 2011. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Proc. INTERSPEECH*, pages 3081–3084.

Nichola Lubold and Heather Pon-Barry. 2014. Acoustic-prosodic entrainment and rapport in collaborative learning dialogues. In *Proc. the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, pages 5–12.

Toyomi Meguro, Ryuichiro Higashinaka, Yasuhiro Minami, and Kohji Dohsaka. 2010. Controlling listening-oriented dialogue using partially observable Markov decision processes. In *Proc. COLING*, pages 761–769.

Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang. 2017. Automatic speech emotion recognition using recurrent neural networks with local attention. In *Proc. ICASSP*, pages 2227–2231.

Yoko Nishihara and Wataru Sunayama. 2009. Estimation of friendship and hierarchy from conversation records. *Inf. Sci.*, 179(11):1592–1598.

Sally Planalp. 1993. Friends' and acquaintances' conversations II: Coded differences. *J. Soc. Pers. Relat.*, 10(3):339–354.

Marylyn Rands and George Levinger. 1979. Implicit theories of relationship: An intergenerational study. *J. Pers. Soc. Psychol*, 37(5):645–661.

Elaine Rich. 1979. User modeling via stereotypes. *Cognitive Science*, 3(4):329–354.

Björn Schuller, Stefan Steidl, and Anton Batliner. 2009. The INTERSPEECH 2009 emotion challenge. In *Proc. INTERSPEECH*, pages 312–315.

Eric Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. *arXiv preprint arXiv:2004.08449*, pages 1–10.

Mohammad Soleymani, Kalin Stefanov, Sin-Hwa Kang, Jan Ondras, and Jonathan Gratch. 2019. Multimodal analysis and estimation of intimate self-disclosure. In *Proc. ICMI*, pages 59–68.

Liyan Xiong, Yin Lei, Weichun Huang, Xiaohui Huang, and Maosheng Zhong. 2016. An estimation model for social relationship strength based on users' profiles, co-occurrence and interaction activities. *Neurocomputing*, 214:927–934.

Yoshihiro Yamazaki, Yuya Chiba, Takashi Nose, and Akinori Ito. 2020. Construction and analysis of a multimodal chat-talk corpus for dialog systems considering interpersonal closeness. In *Proc. LREC*, pages 443–448.

Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2018. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, 126(5):550–569.