2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)

Comparing Empathy Perceived by Interlocutors in Multiparty Conversation and External Observers

Shiro Kumano, Ryo Ishii and Kazuhiro Otsuka
NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, Japan

***Abstract*—This paper investigates the basic characteristics of perceived empathy in Breithaupt's three-person model to consider a way of realizing its automatic prediction or empathy reading machines. More specifically, we report the extent to which interlocutors differ from external observers in perceiving the empathy aroused during group conversation. We also evaluate the accuracies of various frequently used models, including majority voting and multiple regression, in predicting an interlocutor's ratings from those of other interlocutors and/or those of the observers. Defining empathy as the emotional congruence between pairs of interlocutors, we studied a four-person conversation, in which previously unacquainted people held a decision-making discussion. We used a 5-point Likert scale when collecting self-reports of empathy from the interlocutors, and reports from a total of forty external observers (ten for each interlocutor) who adopted a target interlocutor's perspective. We obtained three indications. First, when no empathy ratings are available from the target interlocutor for model training, it is beneficial to ask observers to take the target interlocutor's perspective. Second, when target interlocutors' self-reports are available, it is advantageous to instruct observers not to take the target interlocutor's perspective. Third, in both scenarios, it is useful to ask interlocutors to rate the pairs excluding themselves. These findings provide some insights into good rating procedure as regards studying perceived empathy.**

1. Introduction

Empathy is the basic mechanism by which we understand and share others' thoughts and feelings [1], and shape the nature of social interactions [2], e.g. by side-taking when encountering others in conflict [3]. Unfortunately, empathy is often hard for individuals to achieve, and especially for those lack social cognition skills such as people with autism. And this can also alter their group behavior, e.g. in the form of the social pressure effect [4]. Accordingly, computer-mediated conversation support systems are expected that can enhance the quality and efficiency of communication. This will require the automatic understanding of empathy.

Empathy is a complex phenomenon with several aspects [5], [6]. Batson [1] for example used eight phenomena to summarize empathy, including cognitive empathy, imagine-self perspective, and behavioral coordination. "These phenomena are related to one another, but they are not elements, aspects, facets, or components of a single thing that is empathy, as one might say that an attitude has cognitive, affective, and behavioral components" [1]. It is therefore important to target a specific aspect if we are to make steady progress in dealing with this challenging construct. We chose emotional congruence/contagion, which represents a pairwise state where the emotion of a subject is the same as or similar to that of the target person [7]. Emotional congruence is a kind of affective/emotional empathy [1], and explains a basic aspect of empathy [8], [9]; it is based on the simulation theory, which states that empathy is realized through behavioral mimicry, feedback, and contagion [1].

When considered in terms of multi-party conversations, this definition necessitates different experimental designs and analysis approaches from those used in previous studies on individual emotion in a dyad, e.g. [10], [11]. In particular, this definition inevitably forces us to focus on perceived empathy, since the interpersonal definition precludes us from handling felt empathy, unlike intrapersonally felt emotion, because the pair consists of two members, where they know only little about the partner's actual feelings [11].

Furthermore, group analysis must pay attention to the relationship between the target pair and the perceiver; i.e. whether the perceiver is one of the pair (i.e. an insider), or not (i.e. an outsider). An insider's perception, namely what each insider in a pair believes about their relationship, can be partly explained by the two-person models of empathy considered in many theories, while an outsider's perception would require a three-person model [3]. Such a relationship impacts on their perception, although perception is also affected by various other factors, including interpersonal differences in personality, and social cognitive skills [12].

This study aims at obtaining insights for the automatic prediction of pairwise or interpersonal affective relationships. Specifically, this paper investigates the following four research questions: RQ1) How similarly do interlocutors

S. Kumano and K. Otsuka are, and R. Ishii was, with NTT Communication Science Laboratories, 3-1 Morinosato-Wakamiya, Atsugi, Kanagawa, Japan. kumano@ieee.org

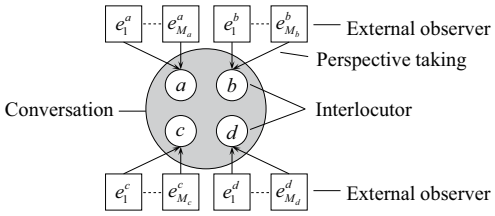

Figure 1. Empathy raters consisting of conversation interlocutors and external observers. Every person provided empathy rating for all interlocutor pairs (${}_4C_2 = 6$ pairs here) for each short time interval of the conversation.

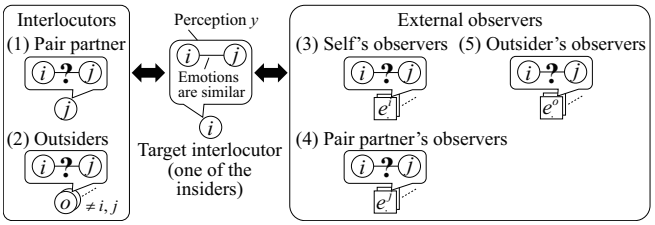

Figure 2. The investigated similarities between the ratings of interlocutor i ($\in \{a, b, c, d\}$ in Fig. 1) for the pair i-j and those of five types of raters: 1) the other pair partner (insider)(j), 2) outsiders ($o(\neq i, j)$), 3) observers who take self's (i's) perspective ($e^i_\cdot$), 4) observers who take the other insider's (j's) perspective ($e^j_\cdot$), and 5) observers who take the outsiders' (e') perspective ($e^o_\cdot$).

perceive empathy in a group as an insider and outsider of the target pair? RQ2) How differently do interlocutors participating in a conversation and external observers (hereinafter observers unless the full term is necessary) outside the conversation perceive empathy in the group? RQ3) How accurately can an interlocutor's perception be predicted by aggregating the other interlocutors' and/or observers' perceptions? RQ4) What type of raters are informative, and for what purpose? Figure 1 shows the structure of our empathy raters, and Fig. 2 describes what we investigated. RQ1-4 correspond to Fig. 2 1-2), 3-5), 1-5), and 3-5), respectively.

As computational models, we introduce various models that have been frequently used in different research areas for different tasks, including empathic accuracy (the use of a partner's rating) [11], majority voting (wisdom of crowds (WoC) [13]), and multiple regression, as introduced in [14] as standard crowd aggregation methods. Our main aim is to provide the basic characteristics of perceived empathy ratings. This is a crucial step for validating crowdsourced affective computing approach, including their computational modeling, like [15]. Note that the proposal of a new machine learning technique, such as the deep neural networks used in [16], is beyond the scope of this paper.

2. Related Work

While emotion still remains one of the main target in the affective computing research community [17], empathy is gaining attention. For example, computation models of empathy that allow conversation agents to behave empathically were proposed in [18], [19]. Xiao et al. [20] proposed a method for identifying therapists' empathy from their speech behaviors while they were interviewing simulated patients. In [21], the observers' perception of empathy, as emotional contagion, aroused in four-person conversations, was analyzed together with the interlocutors' gaze and facial expression. Furthermore, in [22], an observer's cognitive tendency as regards empathy was predicted from his/her personality traits and gender by using topic modeling. However, none of these studies tackled the prediction of the subjective, idiosyncratic empathy perception of interlocutors during multi-party conversations.

This community is keen to predict affective states perceived by observers, i.e. objective perception. To alleviate the subjectivity of perceivers, or increase the reliability of ratings, most previous work gathered the perceptions of multiple observers, and targeted representative values, e.g. the majority/peak [23] or mean [24]. They are assumed to well approximate the affect/emotion felt by the target person. However, previous studies often demonstrated that this assumption is not valid for felt/perceived emotion, e.g. [25], [26]. Furthermore, it remains unclear whether it is also invalid as regards the subjective perception of empathy, i.e. whether empathy perceived by those involved in group conversation can be predicted by aggregating the empathy perceived by observers. This paper examines this issue to determine whether or not we can progress toward the fully automatic prediction of subjective empathy perception by extending previous studies on objective perception.

Ickes et al.'s study [11] on empathic accuracy, the accuracy of social cognition, is well known. It targets dyadic interactions, where the members of the pair rate their own and their partner's thoughts and feelings. Levenson and Ruef [27] also targeted dyadic interactions, and used observer's impressions. Kenny et al.'s social relations analysis [28] aimed at measuring the effects of rater, ratee, and their interaction on the impressions in dyads. Our study differs from these studies in terms of target, i.e. pairwise emotional congruence versus individual affective states.

How individuals perceive the networks of relationships around them has long been of importance in social network research. The research has been motivated by the finding that cognition of the network determines interaction, and interaction in turn changes the network [29], [30]. Cognitive social structures in social network research are of particular note as regards the present study in that no assumptions are made about how the perceptions of one individual will relate to the perceptions of others [31], in contrast to traditional (non-cognitive) network analyses which assume that individuals' perceptions of their social structure should be correlated with perceptions of others for the same network [32]. To the best of our knowledge, cognitive social network analysis has rarely been applied to affective computing studies, and never to empathy studies.

3. Method

3.1. Subjects

We recruited four female students in their early twenties as interlocutors for a four-person group. We employed individuals who were unacquainted before the experiment to reduce idiosyncrasy and emphasize social rules and stereotypes. We recruited a further forty female students of the same ethnicity and also in their early twenties as naïve (non-expert) observers. They were unacquainted with the interlocutors.

3.2. Conversation settings

We introduced the following settings to encourage empathy in the interlocutor group. The group was asked to hold a decision-making discussion, where they had to agree on a single answer related to a given topic. To stimulate their interests in the discussion, the interlocutors performed another task before the discussion session, which was not related to empathy rating but to the discussion topic. The additional task involved playacting according to a given scenario as if behaving spontaneously. The discussion topic was "what is the most important thing while acting"[1]. The requested discussion length was 10 min, and the result was 11.5 min. The participants were seated at equal intervals in a circle with a radius of 1.4 m. This session was videotaped at 30 fps with two full-HD color cameras, each of which captured two participants. These videos were synchronized by using a clapperboard, and were combined into one video in which they are displayed vertically.

To be more exact, we first recruited 17 conversation participants (three three-person groups and two four-person groups) as a preliminary step, and asked them to perform the task mentioned above. We then selected one four-person group who yielded the highest correlation coefficients for ratings between interlocutors (corresponding to the first two rows in Table 3) as the main target of the present study, and further recruited the observers for the group.

1. The exact settings were as follows. Before the discussion session, the interlocutors were asked to take part in a self-introduction session according to the experimenters' instructions given via a document in advance. The document stated: *First, as a self-introduction, all participants, including you, will be asked to spend about 2 min introducing "the best domestic tourist spot I can recommend." However, some of the participants, rather than spontaneously recommending a tourist spot, will act according to given instructions. You were randomly selected as an actor by drawing lots. Speak as if you had been to the tourist location, and try to talk others into wanting to go there.* The tourist spots were different for different participants and randomly assigned to them. After the conversation, they completed several paper-and-pencil questionnaires, including a question that asked *whether they had really been to the spot they recommended*. They were told that none of the answers to the questionnaires would be shown to the other group members. No one answered that she had been to the spot she recommended. To avoid very negative feelings for ethical reasons, the participants were prevented from pointing out other members' faults, but were allowed to praise others' good points during the discussion. At the end of the self-introduction session, they were debriefed that they had all been playacting. After the session, they completed several paper-and-pencil psychological questionnaires, including the big five personality traits, and Davis' Interpersonal Reactivity Index [33]; although these measures were not used in the present study.

TABLE 1. RATING DESIGN FOR FOUR-PERSON GROUP

	Ratee pair					
Rater	1-2	1-3	1-4	2-3	2-4	3-4
1	I	I	I	O	O	O
2	I	O	O	I	I	O
3	O	I	O	I	O	I
4	O	O	I	O	I	I
e_1^1	I'	I'	I'	O'	O'	O'
⋮	⋮	⋮	⋮	⋮	⋮	⋮
e_M^1	I'	I'	I'	O'	O'	O'
e_1^2	I'	O'	O'	I'	I'	O'
⋮	⋮	⋮	⋮	⋮	⋮	⋮
e_M^4	O'	O'	I'	O'	I'	I'

Rater means an interlocutor or observer, while ratee pair means a pair of interlocutors who are being judged. Numbers indicate interlocutor IDs, while $e_\cdot^i$ are observers whose target is interlocutor i. Letters I and O denote the insider and outsider in the ratee pair, respectively. Primes indicate the observers' pseudo relationship in a pair, if they were participating in the conversation. Each rater judges all the ratee pairs.

3.3. Rating design

To compare ratings between interlocutors, between observers, and between interlocutors and observers, this study employed a rating design that combines round-robin [28] and one-with-many designs [28]. In the round-robin design, each interlocutor pair is rated by both the insiders and outsiders of the pair in the group. In the one-with-many design, each pair in the group is also rated by observers, who are not rated by any other people. The task of the observers was to view a group conversation and simultaneously to take the target interlocutor's role, and to rate each interlocutor pair from the target's perspective, i.e. to predict the ratings that would be given by the target interlocutor. Taking imagine-self perspective is one aspect of empathy [1], and enhances accuracies of social cognition [34]. Table 1 shows the rating design. The forty observers were randomly assigned to one of the four interlocutors. Thus, each interlocutor had $M = 10$ observers who took her perspective.

3.4. Rating procedure

All the interlocutors and observers participated in a rating task. The interlocutors completed it immediately after the discussion session. They performed this task in a different room from the conversation room. The rating room was equipped with a 24-inch monitor and had headphones on a desk for each subject. The desks were arranged so that the participants sat back to back. The interlocutors gave a report on their own emotional states and their subjective ratings regarding the empathy between each pair of participants.

Each interlocutor sequentially viewed non-overlapping 7.5-sec long video clips spaced at intervals; this yielded $T = 92$ clips. During the intervals, each person reported

her own emotion (valence-arousal dimension), and judgment regarding the empathy (emotional congruence) between each pair in the group on a 5-point Likert scale; rating $y \in \{-2, -1, 0, +1, +2\}$ (-2: Dissimilar, -1: Slightly dissimilar, 0: Neither, +1: Slightly similar, and +2: Similar). No further details of the empathy definition was given to them. We determined the clip length with reference to [27], [35]. The intervals were set at 25 sec. The emotion rating scores were not analyzed in this study.

The observers completed similar tasks. The differences between the observers' and interlocutors' tasks were as follows. First, the observers were asked to predict the ratings given by the target interlocutor. Second, the observers viewed the discussion video instead of conversing. The observers thus watched the video twice: the first time without stopping, and then including a pause for rating. The interlocutor situations, including the instructions and tasks for the target interlocutor, were given to the observers via documents. Moreover, they were separately seated in a room different from the room used for the interlocutors; there was no interaction between the observers during the experiment.

3.5. Models

The rating procedure, explained in 3.4, yielded 4 (interlocutors) + 40 (observers) = 44 ratings for each video clip and for each pair. The task was now to predict the ratings of interlocutor i to pair i-j (j is the other pair insider) in clip t, y_t^i (where the superscript indicates the rater, and ratee pair i-j is omitted for simplicity), from other interlocutors' and/or observers' ratings. Candidate sources are ratings y_t^x, where x indicates one of the insiders (i.e. interlocutor i or j), outsiders (i.e. other interlocutors), $o(\neq i, j)$, or external observers, e. External observers e^* indicates those taking the perspective of interlocutor $* \in \{i, j, o\}$. Figures 1 and 2 and Table 1 show the obtained ratings and their relationship.

This study employs a leave-one-interlocutor-out cross-validation scenario and a leave-one-out cross-validation scenario. The former is more challenging, since no rating given by the target interlocutor i is available as training samples; the only ratings available are those of all other raters. This scenario simulates a case where it is difficult to ask target interlocutors to provide ratings. In the leave-one-out cross-validation scenario, ratings of i for all clips except for target clip t are also available as training samples. This scenario limits the applications, but we can expect better prediction performance.

This study tests the following three models because of their success in various research fields [11], [13], [14]. The first two are used for the leave-one-interlocutor-out cross-validation scenario, while the last one is applicable only for the leave-one-out cross-validation scenario.

3.5.1. Prediction from another rater's rating. This model predicts interlocutor i's rating y_t^i from another rater's rating for time t. This is divided into three types. The first source is the other insider, i.e. the interlocutor j's rating, y_t^j. The second is the outsiders in the conversation, namely interlocutor o's rating, y_t^o. The last is the external observer e's rating, y_t^e. This simple model aims to measure the pairwise similarity of ratings between an interlocutor and another rater, namely inter-rater agreement.

3.5.2. Wisdom of crowds. WoC [13] is a general technique for collecting answers from a group of individuals and aggregating them to obtain a better answer than the answer given by any of the individuals. In actual applications, the mean and mode (i.e. majority) are probably the most frequently used representative values of the observers' ratings in the psychology and affective computing fields for increasing rating reliability. This study predicts y_t^i by majority voting from other raters' ratings for time t.

3.5.3. Multiple regression. Multiple regression can be applied when the target interlocutor's ratings on an interval or ratio scale are available for model training, namely in the leave-one-out cross-validation scenario. In the present study, the predictors are rating values given by some or all of the raters other than the target interlocutor. The training data for predicting y_t^i were $y_{t'}^x$, where $t' \in \{1, \cdots, t-1, t+1, \cdots, T\}$. The model was separately trained for each target interlocutor i and each ratee pair i-j.

3.6. Prediction performance measures and their comparison metrics

As performance evaluation measures, we follow [36] which recommend the use of three different measures: Pearson's correlation coefficient r, mean absolute error (MAE)[2], and a sign agreement metric (SAGR). MAE is defined as: $\sum_t |\hat{y}_t - y_t| / \sum_t 1$. SAGR [36] is obtained as the agreement level of the prediction with the target observer rating by assessing the rating as either positive (+) or negative (-): $\sum_t \delta(\text{sign}(\hat{y}_t), \text{sign}(y_t)) / \sum_t 1$, where δ is the Kronecker delta function; $\delta(\hat{y}, y)$ returns 1 if $\hat{y} = y$ or 0 otherwise. Unlike [36], our data include zero (no-sign) values, and so we calculate SAGR as the mean of the two SAGR values: the SAGR for a positive-nonpositive binary classification task and the SAGR for a negative-nonnegative binary classification task. In summary, for r and SAGR, larger is better, while for MAE, smaller is better (with the best being zero).

To test the significance of performance differences between models, we focus on the practical significance or effect sizes for the reasons given below and with reference to [37]. Due to the sequential rating procedure, the rating time series exhibited strong autocorrelation, as reported in 4.1. We were therefore forced to aggregate each of the time series for valid statistical tests. First, for each model, we obtained r, MAE, and SAGR separately for each T-length sequence. This yielded 6 (interlocutor pairs) $\times$ 2 (insiders) = 12 values for each evaluation measure. We then calculated Cohen's d [38] between pairs of models by using the twelve

2. This paper uses MAE instead of the root mean square error due to its heavy weighting for outliers, as acknowledged in [36].

samples. We consider there is at least a small effect between the models, if $d \leq 0.2$, according to Cohen's criteria [38][3].

4. Results

4.1. Basic rating characteristics

As general rating statistics, Table 2 shows the frequencies (probabilities) of each rating score for interlocutors and observers; each rater class is further divided into insiders and outsiders. A chi-square test revealed that there is a statistically significant difference between the frequency distributions of the interlocutors and observers, but it was practically trivial ($\chi^2(8, N = 24288) = 200$, $p < 0.001$, Cramer's $V = .09$). Pairwise t-tests revealed that the mean score of the interlocutors is larger (more positive) than that of the observers ($t(2206) = -13.5$, $p < 0.001$, Cohen's $d = .32$). These results match those of some previous studies. For example, [39] argued that participants rate interaction events more positively than observers due to social desirability [40]. Post hoc pairwise t-tests for the 2×2 conditions revealed that both the difference between the mean scores of the insiders and outsiders (within interlocutors) and that between those of the insiders' and outsiders' observers (within observers) were practically trivial. This suggests that both the interlocutors and observers had no or at worst a trivial bias when rating outsiders compared with rating insiders (themselves).

Table 3 shows the pairwise rating similarity between interlocutors and observers. Note that the low correlations between interlocutors suggests that it is difficult to predict an interlocutor's rating from other interlocutors' ratings, even from the rating of the partner in the pair. Our results do not conflict with the findings of previous studies showing that the self-report matching rate with the partner's report regarding valance was not statistically significantly greater than the chance level [11]. The key reason is that communication has a subjective, idiosyncratic meaning for participants [39]. There is no mutual influence between the behavioral, affective, and cognitive responses of the observer and participants, and they are not cognitively interdependent because they do not have unique subjective knowledge about on another [39].

Moreover, each rating time series $y_{\cdot}^{\cdot}$ (each of 6 interlocutor pairs $\times$ 2 insiders = 12 time series) showed a strong autocorrelation (Ljung-Box test for lag=1, $p = .0002 \pm .00009$ (standard error, S.E.)).

4.2. Prediction performance of WoC

Table 4 shows the WoC prediction performance in a leave-one-interlocutor-out cross-validation scenario. When only observers' ratings were used, the best performance was obtained when both insiders' observers were used (WoC 3+4, where the numbers in the model name correspond to the rater categories in Fig. 2, n (number of raters used) = 20: $r = .315 \pm .050$ (S.E.), MAE = 0.928 ± 0.044, and SAGR = $.708 \pm .020$). This was better than when using outsiders' observers (WoC 5, n = 20: $d = 0.29$ for r, $d < 0.2$ for MAE, and $d = 0.35$ for SAGR), and all the observers (WoC 3+4+5, n = 40: $d =< 0.2$, 0.43, and < 0.2, respectively).

The WoC 3+4 performance was further enhanced when combined with both the outsiders' ratings (namely WoC 2+3+4, n = 22: $r = .329 \pm .050$ (d = 0.24), MAE = 0.897 ± 0.046 (d = 0.57), SAGR= $.712 \pm .020$ (d = 0.22)), and with all the other interlocutors' ratings (namely WoC 1+2+3+4, n = 23: $r = .335 \pm .051$ (d = 0.28), MAE = 0.892 ± 0.046 (d = 0.55), SAGR= $.714 \pm .018$ (d = 0.27)). WoC 2+3+4 and WoC 1+2+3+4 are comparable ($d < 0.2$ for all the three measures), and both are better than WoC 1+3+4 (the pair partner + insiders' observers, n = 21; $d > 0.43$ for MAE), and WoC 1+2+3+4+5 (all other interlocutors + all observers, n = 43: $d > 0.22$ for both r and SAGR). Note that the worst model among the models that combine interlocutor(s) and observers in Table 4 was WoC 2+5 (outsiders + their observers, n = 22: $r = .239 \pm .033$, MAE = $0.949 \pm .049$, SAGR = $.685 \pm .174$), which was much worse than the two top-ranked models ($d > .42$).

From these results, it seems reasonable to conclude that WoC 3+4 and WoC 2+3+4 are the most practical choices in this scenario considering the rating cost; n = 22 (WoC 2+3+4) was more economical than n = 23 (WoC 1+2+3+4). This suggests that, when no rating data are available for target interlocutor i, it is beneficial 1) to ask *observers* to focus on the pairs *including* the interlocutor whose perspective she will take, and 2) if possible, to ask all *interlocutors* to rate pairs composed of other interlocutors. This procedure is expected to be helpful in filling the gap between the subjective perceptions of target interlocutor and other raters without using any idiosyncratic data about the target interlocutor.

TABLE 2. PROBABILITIES OF EMPATHY RATING SCORES

	Rating score					
	+2	+1	0	-1	-2	Mean
Interlocutors	.22	.37	.25	.12	.04	.62
Insiders	.23	.36	.23	.12	.06	.59
Outsiders	.21	.38	.28	.11	.02	.66
Observers	.15	.34	.23	.19	.09	.26
Insiders'	.15	.34	.21	.20	.10	.22
Outsiders'	.15	.34	.26	.17	.08	.31

4.3. Prediction performance of regression

Table 5 shows the prediction performance of regression in a leave-one-out cross-validation scenario. First, as expected, this scenario yielded much higher performance than that in the previous leave-one-interlocutor-out cross-validation scenario, thanks to the use of the self-reports of

3. The sample size of twelve was unfortunately prone to cause type-II errors, namely failure to find the true difference. A power analysis suggested that the detectable effect size is as large as 0.89 to guarantee a type-II error rate of .2 with N = 12, and the power is only 0.09 (meaning type-II error rate = 1 - .09 = .91) when N=12 and the effect size = 0.2.

TABLE 3. RATING SIMILARITIES OR INTER-CODER AGREEMENT WITH TARGET INTERLOCUTOR i

Sources of prediction (model)	#Raters used (n)	Pearson r ↑	MAE ↓	SAGR ↑
Interlocutors' ratings				
1) Pair partner ($x = j$)	1	.143 (.032)	1.174 (0.054)	.607 (.007)
2) Outsider ($x = o$)	1	.195 (.032)	1.038 (0.057)	.638 (.013)
Observers' ratings				
3) Self's observers ($x = e_i$)	1	.146 (.021)	1.180 (0.039)	.647 (.017)
4) Pair partner's observers ($x = e_j$)	1	.163 (.025)	1.172 (0.037)	.642 (.016)
5) Outsiders' observers ($x = e_o$)	1	.135 (.016)	1.149 (0.029)	.638 (.009)

The rater category numbers (1-5) correspond to those in Fig. 2. Mean and standard error in brackets.
"#Raters used" denotes the number of source raters used for prediction. "↑" and "↓" denote higher and lower performance.

TABLE 4. PREDICTION PERFORMANCE OF WOC

Sources of prediction (model)	#Raters used (n)	Pearson r ↑	MAE ↓	SAGR ↑
Interlocutors' ratings				
WoC 1+2) All interlocutors	3	.167 (.033)	1.150 (0.041)	.625 (.010)
Observers' ratings				
WoC 3) Self's observers	10	.239 (.033)	1.047 (0.061)	.694 (.020)
WoC 4) Pair partner's observers	10	.301 (.045)	1.016 (0.049)	.685 (.022)
WoC 5) Outsiders' observers	20	.265 (.030)	0.942 (0.048)	.693 (.018)
WoC 3+4) Insiders' observers	20	.315 (.050)	0.928 (0.044)	.708 (.020)
WoC 3+4+5) All observers	40	.303 (.044)	0.888 (0.052)	.707 (.018)
Both interlocutors' and observers' ratings				
WoC 1+3+4) Pair partner + insiders' observers	21	.322 (.049)	0.923 (0.044)	.709 (.021)
WoC 2+3+4) Outsiders + insiders' observers	**22**	**.329 (.050)**	**0.897 (0.046)**	**.712 (.020)**
WoC 2+5) Outsiders + outsiders' observers	22	.239 (.033)	0.949 (0.049)	.685 (.017)
WoC 1+2+3+4) All interlocutors + insiders' observers	**23**	**.335 (.051)**	**0.892 (0.046)**	**.714 (.018)**
WoC 1+3+4+5) Pair partner + all observers	41	.292 (.041)	0.908 (0.049)	.701 (.017)
WoC 2+3+4+5) Outsiders + all observers	42	.318 (.046)	0.877 (0.050)	.710 (.018)
WoC 1+2+3+4+5) All interlocutors + all observers	43	.306 (.046)	0.890 (0.052)	.706 (.017)

TABLE 5. PREDICTION PERFORMANCE OF REGRESSION

Sources of prediction (model)	#Raters used (n)	Pearson r ↑	MAE ↓	SAGR ↑
Interlocutors' ratings				
Reg 1+2) All interlocutors	3	.255 (.041)	0.853 (0.055)	.690 (.013)
Observers' ratings				
Reg 3) Self's observers	10	.306 (.053)	0.777 (0.036)	.722 (.017)
Reg 4) Pair partner's observers	10	.350 (.042)	0.769 (0.041)	.724 (.016)
Reg 5) Outsiders' observers	20	.415 (.038)	0.738 (0.049)	.745 (.016)
Reg 3+4) Insiders' observers	20	.350 (.051)	0.803 (0.040)	.709 (.016)
Reg 3+4+5) All observers	40	.378 (.047)	0.829 (0.047)	.707 (.014)
Interlocutors' ratings + observers' ratings				
Reg 1+3+4) Pair partner + insiders' observers	21	.345 (.056)	0.801 (0.036)	.708 (.018)
Reg 1+5) Pair partner + outsiders' observers	21	.395 (.041)	0.752 (0.049)	.737 (.016)
Reg 2+3+4) Outsiders + insiders' observers	22	.383 (.048)	0.773 (0.041)	.716 (.017)
Reg 2+5) Outsiders + outsiders' observers	**22**	**.450 (.046)**	**0.726 (0.055)**	**.744 (.019)**
Reg 1+2+3+4) All interlocutors + insiders' observers	23	.358 (.046)	0.795 (0.044)	.711 (.017)
Reg 1+2+5) All interlocutors + outsiders' observers	23	.439 (.045)	0.735 (0.056)	.741 (.019)
Reg 1+2+3+4+5) All interlocutors + all observers	43	.383 (.039)	0.838 (0.049)	.707 (.013)

target interlocutor i. For example, when compared with the best WoC, WoC 2+3+4, Reg 2+3+4 (outsiders + insiders' observers, n = 22: $r = .383 \pm .048$, MAE = 0.773 ± 0.041, and SAGR = $.716 \pm .017$) was better in terms of r (d = 0.34) and MAE (d = 0.84), although not in SAGR ($d < 0.2$).

Second, and more strikingly, in contrast to the previous scenario in 4.2, observers taking the insiders' perspective turned out to be counterproductive in this scenario. The best regression model was Reg 2+5 (outsiders + their observers, n = 22: $r = .450 \pm .046$, MAE = 0.726 ± 0.055, and SAGR = $.744 \pm .019$), which corresponds to the worst WoC model, WoC 2+5. The improvement was huge; 88% increase for r ($d = 1.41$), 31% decrease for MAE ($d = 1.44$), and 9% increase for SAGR ($d = 1.13$). Reg 2+5 was better than all of Reg 2+3+4 (n = 22: d = 0.31, 0.31, and 0.44 for r, MAE and SAGR, respectively), Reg 5 (outsiders' observers, n = 20: d = 0.50, 0.24, and < 0.2, respectively), and Reg 1+2+5 (all other interlocutors + outsiders' observers, n = 23: d = 0.76, 0.81, and 0.55, respectively).

These results suggest that when the rating data of target

interlocutor i for other conversational scenes are available, it looks advantageous to ask the raters, including both the interlocutors and observers, to take a perspective that is different from that of the target interlocutor (which is expected to yield a more objective perspective). More specifically, it seems effective 1) to instruct *the observers* to focus on the pairs *excluding* the interlocutor whose perspective she will take, and 2) to ask *the interlocutors* to rate pairs excluding themselves. These are reasonable results because the perspectives of both the outsiders and their observers are more objective than the insiders' subjective perspectives, and the subjective perspectives are provided by using the target interlocutor's ratings. These findings also match the results of a recent emotion study [41], which showed the complementarity of self-reports and observers' reports, although their target was individual emotional state, not pairwise empathetic state. These higher performance levels obtained with rater aggregation suggest that the raters are not completely random, although individual raters showed low performance.

5. Discussion

We have provided various results that will be helpful for the automatic prediction of perceived empathy. However, several issues still remain.

5.1. Low pairwise correlation of ratings

As shown in Table 3, the pairwise correlation coefficients were not very high compared with previous studies on individual emotions, e.g. [25], [26], [35]. However, it should be noted that many previous studies, e.g. [27], [29], targeted only scenes with high inter-coder agreement. Actually, the clips whose SD by all raters was smaller than 0.5 yielded a moderate mean absolute correlation of around 0.5. Accordingly, as in [21], this would suggest data characteristics rather than unreliable rating. We thus used all the samples to capture the big picture.

5.2. Sample size

This study only targeted a single four-person conversation, whereas we recruited forty observers. The statistical tests revealed certain practical significance, but these should be further validated with various interlocutors and conversation settings, e.g. different group sizes. However, we still believe these results can provide the affective computing community with some insights for the following two reasons. First, to reduce the effect of the above issue, from five conversation groups (three three-person groups and two four-person groups), we selected the group that showed the highest inter-coder agreement between the interlocutors. Second, we obtained reasonable results that well match those in the literature or can be explained by them, as described in Section 4.

5.3. Perceived dyadic state vs. perceived self-state

Motivated by Breithaupt's three-person model of empathy [3], this paper introduced the framework of Krackhardt's social structures analysis [31] to analyze the pairwise empathy (defined as emotional congruence in this study) *perceived* by the interlocutors. This opens several future directions.

First, this study used a pairwise rating design. However, other empathy definitions require a different design, and thus their results are comparable with ours. For example, cognitive empathy, which is the other major aspect of empathy, necessitates an individual-oriented design, where internal states felt by each interlocutor and those perceived by others should be corrected and then compared. In terms of design, this is much more relevant to previous emotion studies. The pairwise design differs from the individual-oriented design according to [42] who reported that ratings of pairwise relationships would be affected by the balance schema, as in Heider's theory [43]. The comparison of different empathy definitions is an important issue.

Second, the current definition also makes it possible to approximate the actual dyadic state by determining the similarity between emotions felt by a pair; the emotions would be those provided as self-reports or those measured via physiological responses, like galvanic skin response and heart rate, in felt-emotion studies [16], [44]. It would be necessary to compare the approximated and self-reported empathy ratings.

5.4. Rating procedure

This study employed a discrete annotation procedure for both time and empathy space. However, we acknowledge that the recent trend in the affective computing community is rating in continuous time with a continuous value, e.g. [45], [46], although observer-specific delays should be considered [47], [48]. It would also be interesting to investigate whether or not our results hold in continuous annotation scenarios.

6. Conclusion

This paper investigated the basic characteristics of perceived empathy. We reported the extent to which interlocutors differ from external observers in perceiving empathy aroused in a group conversation, and also evaluated the accuracies various frequently-used models in predicting an interlocutor's ratings from those of other interlocutors and/or those of observers. We obtained three indications. First, when no empathy ratings for the target interlocutor are available for model training, it is beneficial to ask observers to take the target interlocutor's perspective. Second, when target interlocutors' self-reports are available, it is advantageous to instruct observers not to take the target interlocutor's perspective. Third, in both scenarios, it is useful to ask interlocutors to rate the pairs excluding themselves. We believe that these findings provide insights regarding how to collect perceived empathy for its automatic prediction.

References

[1] C. D. Batson, *The Social Neuroscience of Empathy*. MIT press, 2009, ch. 1. These things called empathy: eight related but distinct phenomena, pp. 3–15.

[2] M. H. Davis, *The Social Life of Emotions*. Cambridge University Press, 2004, ch. Empathy: Negotiating the border between self and other, pp. 19–42.

[3] F. Breithaupt, "A three-person model of empathy," *Emotion Review*, vol. 4, no. 1, pp. 84–91, 2012.

[4] S. Asch, "Opinions and social pressure," *Scientific American*, vol. 193, no. 5, pp. 31–35, 1955.

[5] S. D. Preston and A. J. Hofelich, "The many faces of empathy: Parsing empathic phenomena through a proximate, dynamic-systems view of representing the other in the self," *Emotion Review*, vol. 4, no. 1, pp. 24–33, 2012.

[6] B. M. Cuff, S. J. Brown, L. Taylor, and D. J. Howat, "Empathy: A review of the concept," *Emotion Review*, vol. 8, no. 2, pp. 144–153, 2014.

[7] S. D. Preston and F. B. de Waal, "Empathy: Its ultimate and proximate bases," *Behavioral and Brain Sciences*, vol. 25, no. 1, pp. 1–20, 2002.

[8] A. Smith, "Cognitive empathy and emotional empathy in human behavior and evolution," *The Psychological Record*, vol. 56, no. 1, pp. 3–21, 2006.

[9] H. Walter, "Social cognitive neuroscience of empathy: Concepts, circuits, and genes," *Emotion Review*, vol. 4, no. 1, pp. 9–17, 2012.

[10] D. A. Kenny and L. Albright, "Accuracy in interpersonal perception: A social relations analysis." *Psychol. Bull.*, vol. 102, no. 3, pp. 390–402, 1987.

[11] W. Ickes, L. Stinson, V. Bissonnette, and S. Garcia, "Naturalistic social cognition: Empathic accuracy in mixed-sex dyads," *J. Pers. Soc. Psychol.*, vol. 59, no. 4, pp. 730–742, 1990.

[12] M. Hancock and W. Ickes, "Empathic accuracy: When does the perceiver-target relationship make a difference?" *Journal of Social and Personal Relationships*, vol. 13, no. 2, pp. 179–199, 1996.

[13] J. Surowiecki, *The Wisdom of Crowds*. New York: Anchor, 2005.

[14] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *J. Mach. Learn. Res.*, vol. 11, pp. 1297–1322, 2010.

[15] D. Ozkan and L.-P. Morency, "Modeling wisdom of crowds using latent mixture of discriminative experts." in *Proc. Assoc. Comput. Linguist.*, 2011, pp. 335–340.

[16] H. P. Martinez, Y. Bengio, and G. N. Yannakakis, "Learning deep physiological models of affect," *IEEE Comp. Intell. Magazine*, vol. 8, no. 2, pp. 20–33, 2013.

[17] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. PAMI*, vol. 31, no. 1, pp. 39–58, 2009.

[18] S. W. McQuiggan and J. C. Lester, "Modeling and evaluating empathy in embodied companion agents." *Int. J. Hum.-Comput. Stud.*, vol. 65, no. 4, pp. 348–360, 2007.

[19] H. Boukricha, I. Wachsmuth, M. N. Carminati, and P. Knoeferle, "A computational model of empathy: Empirical evaluation," in *Proc. ACII*, 2013, pp. 1–6.

[20] B. Xiao, P. G. Georgiou, Z. E. Imel, D. C. Atkins, and S. S. Narayanan, "Modeling therapist empathy and vocal entrainment in drug addiction counseling," in *Proc. Interspeech*, 2013, pp. 2861–2865.

[21] S. Kumano, K. Otsuka, D. Mikami, M. Matsuda, and J. Yamato, "Analyzing interpersonal empathy via collective impressions," *IEEE Trans. Affect. Comput.*, vol. 6, no. 4, pp. 324–336, 2015.

[22] S. Kumano, K. Otsuka, M. Matsuda, R. Ishii, and J. Yamato, "Using a probabilistic topic model to link observers' perception tendency to personality," in *Proc. ACII*, 2013, pp. 588–593.

[23] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources And Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[24] M. Nicolaou, H. Gunes, and M. Pantic, "Output-associative RVM regression for dimensional and continuous emotion prediction," in *Proc. IEEE Int'l Conf. FG'11*, 2011, pp. 186–196.

[25] C. Busso and S. S. Narayanan, "The expression and perception of emotions: Comparing assessments of self versus others," in *Proc. Interspeech*, 2008, pp. 257–260.

[26] K. Truong, D. van Leeuwen, M. Neerincx, and F. de Jong, "Arousal and valence prediction in spontaneous emotional speech: felt versus perceived emotion," in *Proc. Interspeech*, 2009, pp. 2027–2030.

[27] R. W. Levenson and A. M. Ruef, "Empathy: A physiological substrate," *J. Pers. Soc. Psychol.*, vol. 63, no. 2, pp. 234–246, 1992.

[28] D. A. Kenny, D. A. Kashy, and W. L. Cook, *Dyadic Data Analysis (Methodology in the Social Sciences)*. Guilford, 2006.

[29] K. M. Carley and D. Krackhardt, "Cognitive inconsistencies and non-symmetric friendship." *Social Networks*, vol. 18, no. 1, pp. 1–27, 1996.

[30] S. P. Borgatti and P. C. Foster, "The network paradigm in organizational research: A review and typology," *Journal of Management*, vol. 29, no. 6, pp. 991–1013, 2003.

[31] D. Krackhardt, "Cognitive social structures," *Social Networks*, vol. 9, no. 2, pp. 109–134, 1987.

[32] R. A. Brands, "Cognitive social structures in social network research: A review," *J. Organiz. Behav.*, vol. 34, pp. S82–S103, 2013.

[33] M. H. Davis, "Measuring individual differences in empathy: Evidence for a multidimensional approach," *J. Pers. Soc. Psychol.*, vol. 44, no. 1, pp. 113–126, 1983.

[34] A. D. Galinsky and G. B. Moskowitz, "Perspective-taking: decreasing stereotype expression, stereotype accessibility, and in-group favoritism," *J. Pers. Soc. Psychol.*, vol. 78, no. 4, pp. 708–724, 2000.

[35] G. Margolin, D. Hattem, R. S. John, and K. Yost, "Perceptual agreement between spouses and outside observers when coding themselves and a stranger dyad," *Behav. Assess.*, vol. 7, no. 3, pp. 235–247, 1985.

[36] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Trans. Affect. Comput.*, vol. 2, no. 2, pp. 92–105, 2011.

[37] C. Woolston, "Psychology journal bans p values. nature," *Nature*, vol. 519, p. 9, 2015.

[38] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. Routledge, 1988.

[39] C. A. Surra and C. A. Ridley, *Studying interpersonal interaction*. Guilford Press, 1991, ch. Multiple perspectives on interaction: Participants, peers, and observers, pp. 35–55.

[40] M. Sullaway and A. Christensen, "Couples and families as participant observers of their interaction," *Advances in Family Intervention, Assessment & Theory*, vol. 3, pp. 119–160, 1983.

[41] B. Zhang, G. Essl, and E. Mower Provost, "Automatic recognition of self-reported and perceived emotion: Does joint modeling help?" in *Proc. ACM ICMI*, 2016, pp. 217–224.

[42] D. Krackhardt and M. Kilduff, "Whether close or far: social distance effects on perceived balance in friendship networks," *J. Persona. Soc. Psychol.*, vol. 76, no. 5, pp. 70–782, 1999.

[43] F. Heider, *The psychology of interpersonal relations*. New York: John Wiley & Sons., 1958.

[44] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, “A multimodal database for affect recognition and implicit tagging,” *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, 2012.

[45] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, “FEELTRACE: An instrument for recording perceived emotion in real time,” in *ISCA Workshop on Speech and Emotion*, 2000, pp. 19–24.

[46] G. N. Yannakakis and H. P. Martínez, “Grounding truth via ordinal annotation,” in *Proc. ACII*, 2015, pp. 574–580.

[47] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani, “Robust continuous prediction of human emotions using multiscale dynamic cues,” in *Proc. ICMI*, 2012, pp. 501–508.

[48] S. Mariooryad and C. Busso, “Correcting time-continuous emotional labels by modeling the reaction lag of evaluators,” *IEEE Trans. Affect. Comput.*, vol. 6, no. 2, pp. 97–108, 2015.