Analyzing Interpersonal Empathy
via Collective Impressions

Shiro Kumano, *Member, IEEE,* Kazuhiro Otsuka, *Member, IEEE,* Dan Mikami, *Member, IEEE,*
Masafumi Matsuda, and Junji Yamato, *Senior Member, IEEE*

Abstract—This paper presents a research framework for understanding the empathy that arises between people while they are conversing. By focusing on the process by which empathy is perceived by other people, this paper aims to develop a computational model that automatically infers perceived empathy from participant behavior. To describe such perceived empathy objectively, we introduce the idea of using the collective impressions of external observers. In particular, we focus on the fact that the perception of other's empathy varies from person to person, and take the standpoint that this individual difference itself is an essential attribute of human communication for building, for example, successful human relationships and consensus. This paper describes a probabilistic model of the process that we built based on the Bayesian network, and that relates the empathy perceived by observers to how the gaze and facial expressions of participants co-occur between a pair. In this model, the probability distribution represents the diversity of observers' impression, which reflects the individual differences in the schema when perceiving others' empathy from their behaviors, and the ambiguity of the behaviors. Comprehensive experiments demonstrate that the inferred distributions are similar to those made by observers.

Index Terms—Empathy, perception, cognition, collective impressions, subjectivity, objectivity, voting rates, observer, facial expression, gaze, probabilistic modeling, Bayesian network

✦

1 INTRODUCTION

F ACE-TO-FACE conversation is the primary way of sharing information, understanding others' empathy, and making decisions in social life. Unfortunately, it is not very easy for people to fully understand what others are feeling during a conversation, or for participants to agree completely about a controversial topic. To improve the quality and efficiency of communication, it would be helpful if our conversations could be supported by information technology such as computer-mediated visual telecommunication or conversational agents/robots. To realize such applications, it is essential to understand automatically both human behavior and participant empathy, which evolve over the course of the interaction and affect the conversation.

Against this background, automatic meeting analysis has recently been acknowledged as an emerging research area [1], [2]. Most previous studies offer only preliminary steps toward the recognition of the behaviors of conversation participants individually or interactions among people: e.g. speaker detection,

- *S. Kumano, K. Otsuka, D. Mikami, M. Matsuda, and J. Yamato are with Nippon Telegraph and Telephone Corporation, Kanagawa, Japan. E-mail: kumano@ieee.org*

speech recognition, and the recognition of nonverbal behaviors such as gaze, facial expressions, gestures and postures, and their interaction, e.g. who is talking to whom, and turn taking. The emotional aspect is only now being addressed, e.g. in [3], [4].

Many existing psychological studies support the importance of empathy; e.g. empathy will play a significant role in shaping the nature of human social interactions [5]. In terms of the eight phenomena of empathy [6], many studies define empathy as emotional contagion or emotional empathy based on the simulation theory, which tries to explain that empathy is realized through behavioral mimicry, feedback, and contagion [6]. Recent neuroscientific evidence for the mechanism is summarized in [7].

To understand empathy in conversation, it is important to shed light on the communication process by which empathy is expressed and perceived between people via their interactions [8], [9]. A typical conversation scene is one where the empathy of a participant (first person) is mainly expressed by his/her nonverbal behaviors, and it is oriented to the other participant (second person) by casting and modifying the gaze. The second person perceives the empathy from the first person's behavior, and this perception is likely to affect the second person's empathy. The second's empathy also could affect the first's empathy via the second person's behavior. This expression-perception loop is a basic element of face-to-face interaction; the continuous repetition and accumulation of loops yield shared feelings, consensus, and relationships between people [8], [10], [11].
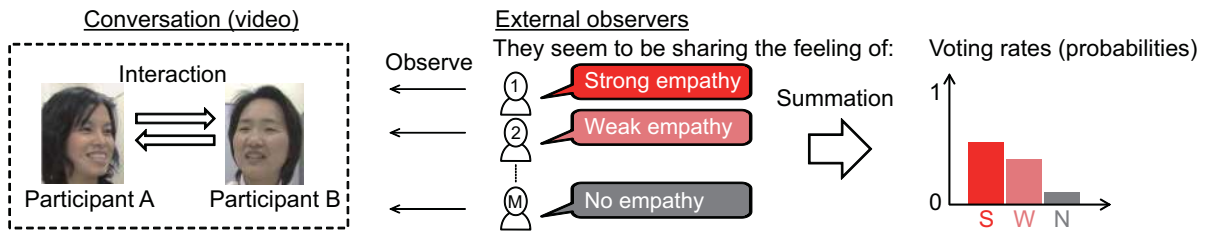
Fig. 1. Perceived empathy in this research: A collection of subjective impressions from multiple external observers is used to objectively describe empathy shared between a conversing couple. The impressions are aggregated as a probability distribution, i.e. voting rates.

The expression-perception loop inevitably involves various ambiguities. These ambiguities arise from the fact that humans cannot directly determine another's actual empathy, but instead have to guess it from the person's behaviors according to their own perception schema, e.g. theory of mind (ToM) [12]. For example, people might express empathy with different behaviors in terms of intensities, types and/or modalities. In addition, even when observing the same expression, different observers might perceive empathy differently. These ambiguities can lead to failures in communication, such as a misunderstanding of attitude, opinion, and personality.

In multi-party conversations, the perceptions of empathy are formed by both the interacting pair and the side participants. Each side participant perceives the empathy between a pair by focusing on their behaviors, i.e. how the pair are interacting, rather than individually. The perception of the side participant could change his/her internal state, such as emotion, feeling, and thought, which affects his/her next interaction with others. Since side participants frequently switch roles to become interaction participants and vice versa, the perceptions of side participants are a significant factor driving the conversation. For example, if someone who does not agree with us receives empathy from a third person, we might change our mind as a result of the third person's attitude. A good example of this is the social pressure effect [13]. Thus, it is important to analyze how the empathy of participants is perceived by others via their behaviors from the viewpoint of side participants.

The aim of the present study is to propose a framework for objectively understanding empathy between a pair of participants as perceived by others, to build a computational model, and to infer automatically the perceptions by using the model to support the deep analysis of conversations. The present study defines empathy as instantaneous "emotional contagion" between a pair of participants, and defines perceived empathy (perceived counter-empathy) as "the impression of an observer who judges that a pair of participants engaged in a multi-party conversation is in an empathic (counter-empathic) state at that moment, where the pair are exhibiting a one-way or

mutual gaze, and the observer is watching the conversation scene". The perceived empathy is an atomic element of the empathic expression-perception loop. By expressing and perceiving positive/neutral/negative feelings or attitudes toward each other via behaviors over a short term, the communicating pair are seen to be sharing feelings of empathy.

As an objective description of perceived empathy, we introduce the concept of collective impressions, i.e. a collection of perceptions made by multiple observers outside the conversation. Each observer perceives participant empathy from his/her own subjective viewpoint. We hypothesize that collecting the impressions of multiple observers suppresses the subjectivity of each individual's assessment. We set external observers to watch a video of conversation to emulate the perception process of the side participants.

To handle the ambiguities present in the impressions of observers effectively, this paper aims to build a probabilistic model, called a Bayesian Network [14], which relates the empathy perceived by observers to the nonverbal behaviors of the participants. In our model, the probability distribution represents the individual differences in the impressions of the observers, i.e. it provides voting rates showing how many observers voted or will vote for each empathy category. For example, if the numbers of votes for three categories cast by nine observers are (3, 5, 1), the probability distribution is $(1/3, 5/9, 1/9)$. The distribution reflects the inter-observer difference in the perception schema and the ambiguity of participant behaviors. More unfocused voting means that the interaction has more significant ambiguities in terms of impression, or vice versa. We consider diversity and ambiguity to be essential attributes, and they must be well handled to better support conversations and encourage feelings of satisfaction. Figure 1 summarizes the target and its objective description in this paper.

In addition to the novelty of the inference target, our model is also characterized by the input since it focuses on how participant behaviors co-occur between a pair. We study the relationship of facial expression (FE) and gaze with empathy perception, especially in their combination between a pair of participants. FE plays a major role in conveying empathic

messages [15]; gaze is vital for monitoring the FEs and triggers a gaze shift and a reaction from the gazee. We here provide typical examples. Two participants are looking at each other and smiling. Many people observing this event would perceive that the pair are sharing the feeling of empathy. On the other hand, if one is smiling but the other is not, the number of observers perceiving empathy is likely to fall. We probabilistically model such tendencies in terms of how likely observers are to perceive the empathy of the pair when their FEs co-occur.

This research considers that the model is successful if it well recreates the distributions of impressions received by an adequate observer group. Accordingly, we propose a quantitative model verification scheme that compares the inferred probability distribution with the distribution obtained by a number of external observers. Most previous studies that utilize the judgment of multiple coders target the approximation of the emotional *state*, such as the six basic emotion categories [16] and valence/arousal [17], of a person as the ground truth as determined by, for example, majority voting [18] or averaging [19]. They thus evaluate their models with regard to the rate at which they can correctly recognize/identify the *state*. We call this technique the consensus approach.

We perform an experiment in which we use four-party conversation data from four groups, and ask between five and nine observers to provide their impressions about the empathy between each pair of participants. The results demonstrate that the inferred distributions are quite similar to those made by the observers. We further compare our proposed approach with baseline methods, including the consensus approach. Of particular note is that the results show that the proposed approach is superior to the consensus approach, especially for scenes where coders disagree.

The remainder of this paper is organized as follows. We introduce work related to this study in 2. The collective impressions of empathy are described and assessed in 3. The proposed model is described in 4, and evaluated in 5. A discussion and the potential for future growth are provided in 6. And we summarize this study in 7.

2 RELATED WORK

This section positions this study by comparison with related work.

We first focus on research on the automatic recognition of human emotions, which are relevant to and studied more than empathy in the engineering research area. Even though continuous efforts have been made over the last three decades, most research has aimed at estimating the perceived emotion for the target [20]. We, on the other hand, focus on how empathic states between a pair who are conversing

are perceived by others. The difference between the concepts of empathy necessitates a different research framework.

General research frameworks on the automatic inference of a target phenomenon, not only the emotional state but also other states, involve the critical issue of how to determine the ground truth. There are two major approaches to this issue [20], [21]. One manually sets the ground truth in advance, and then captures observation data, e.g. recruits actors and asks them to behave as if they were feeling a specified emotion [22]. However, it is difficult to control a conversation in such a way that each interaction between participants creates the voting rates that the researchers expected.

The other approach first captures observation data, and then sets the ground truth. This approach often uses self-reports [18], [23] or external coders' reports [18], [19], [24], [25]. Self-reports are unfortunately inappropriate for interactive situations such as conversations. First, real-time reporting severely alters the conversation due to its heavy mental loads. Previous studies, e.g. [21], [26], [27], used real-time tools for self-reporting, but the subject simply received a stimulus, e.g. watched an emotion-eliciting video [27] or his/her own conversation video [23], or listened to music [26]. Second, it is considered to be difficult for the general public to correctly report their moderately elicited emotions at every moment in depth after a conversation, although people can report several categories of emotion at certain moments if the emotion was intense [28].

The judgment of coders has often been used in previous studies to approximate real emotion; the assumption is that the coders are good at inferring other's emotional states [29]. Example datasets annotated with individual emotions by multiple external observers include SEMAINE [30] and IEMOCAP [18]. Multiple coders are used to ensure the reliability, or more accurately reproducibility [31], of the coding, and the target emotional state is determined by techniques such as majority voting (e.g. [18]), averaging (e.g. [19]), and consensus (e.g. [24], [25]). In other words, the traditional viewpoint basically considers the inter-coder difference in interpretation to be a noise that degrades the reliability.

Recently, researchers are gradually starting to recognize that the inter-coder difference is of prime importance and cannot be ignored, but most have not attempted to automatically infer the inter-coder difference. For example, Steidl et al. [32] utilize inter-coder variation to evaluate traditional emotion classifiers for recognizing the approximated real emotional state of a person. There have also been some attempts to extract a better ground truth label from the disparate labels of multiple coders [33], [34], [35], [36]. Their key idea is to introduce several parameters such as the level of expertise of coders and the difficulty of coding.

Their work has demonstrated that their approaches can infer the ground truth label more accurately than naïve majority voting. On the other hand, we focus on the label distribution. We consider that the real empathy of target people and its impression on the observer should be distinguished, and there is no clear reason to weight the impression of each individual in communication. Thus, we do not assume a unique ground truth label.

More recently, Meng et al. [37] and Scherer et al. [38] discussed the task of estimating the distribution of emotional labels made by multiple coders; this point is similar to ours. However, their targets, their ways of viewing and approaching the target, and their computational models are quite different from ours. 1) Targets: They focus on the impression of the *emotional states* of *individuals*: his/her emotional postures during game play [37], and voice quality [38]. On the other hand, we target the impression of the *pair-wise empathic states*. 2) Features: The different targets necessitate different features. They employ the features of a single person, while we model the co-occurrence of behaviors between the pair. 3) Models: They directly model a discriminative function with SVMs, regression etc., while we choose probabilistic modeling to explain the target label distribution by combining conditional probabilities.

Our approach of asking a group of people for help is a kind of crowdsourcing, which is now attracting a great deal of attention in various research communities. Among the four types of crowdsourcing [39], our approach is most like offline *crowd voting* in that it recognizes the value of a full set of judgments made by a crowd. To the best of our knowledge, this is the first study to associate observers' voting rates with impressions about empathy. A related type of crowdsourcing is the wisdom of crowds [40], which seeks to aggregate anonymously sourced data by techniques such as averaging or majority voting. For example, Soleymani and Larson [41] explored how significantly the average level of observer-reported boredom differs with different numbers of observers obtained by crowdsourcing. Biel et al. [42] tried to infer automatically the average crowdsourced impression of multiple observers of video bloggers. Some of the abovementioned studies [33], [34], [35], [36] are also of this type, although they attempt to model the annotator's characteristics.

The importance of perceived empathy is supported by some social psychological studies. Ickes et al. [43] and Levenson and Ruef [44] defined empathy as the ability to accurately perceive how another person is feeling, and studied dyadic conversations. They reported that it is difficult for the interaction partner [43] and external observers [44] to infer accurately the valence or emotional tone of the target person. These results support the importance of discriminating the real empathy between target people and the empathy perceived by others. Barsade [45] demonstrated the effect of emotional contagion between participants on conversation. She focused on long-term change in group mood, while we focus on instantaneous pair-wise emotional contagion/conflict. Moreover, Cowie and Cornelius [29] showed, in pioneering work on the effect-type description of emotion, the rough concept of the impact of emotional speech on the listener. However, none of them ([29], [43], [44], [45]) focused on the inter-observer difference.

We have already proposed our research framework and model in [46] and [47], respectively. This paper provides a substantially extended survey of related work, the use of more reliable FE labels that were obtained by three coders, an evaluation of the annotations based on traditional inter-coder agreement tests, discussions about the number of observers and the stability of the distribution of the perception, and an experiment on a variety of baseline models, including comparison with the consensus approach, which attempts to identify the majority class. To focus on the modeling of the relationship between observers' perceived empathy and participants' behaviors, the present study utilizes participants' behavior as identified by human observer; our previous work described in [47] tried to automatically recognize facial expressions from images.

3 VARIABILITY IN HUMAN JUDGMENT

This section explains our data, which includes conversation videos, perceived empathy obtained from external observers, and the annotation of interlocutor behaviors. The properties of the perceived empathy is then analyzed from various statistical aspects.

3.1 Subjects: external observers

We employed nine external observers. They were Japanese females in the same age bracket as the participants who were chosen because the similarity in age and gender increases empathy [48]. They had met neither each other nor conversation participants, explained in 3.2, before the experiment.

3.2 Materials: conversation data

This paper targets four-person face-to-face conversations, as shown in Fig. 2 (a). The participants were 16 Japanese women (four four-person groups: $\mathcal{G}_\mathcal{A}$, $\mathcal{G}_\mathcal{B}$, $\mathcal{G}_\mathcal{C}$, and $\mathcal{G}_\mathcal{D}$) in their twenties or thirties. They had not met before the experiment.

They were first asked to have a short chat that included a self-introduction, and then to engage in about seven alternative-type discussions with intervals between them. Each group's discussions took place on a single day. The participants were instructed to build consensus as a group, i.e. derive a single answer, within eight minutes for each discussion
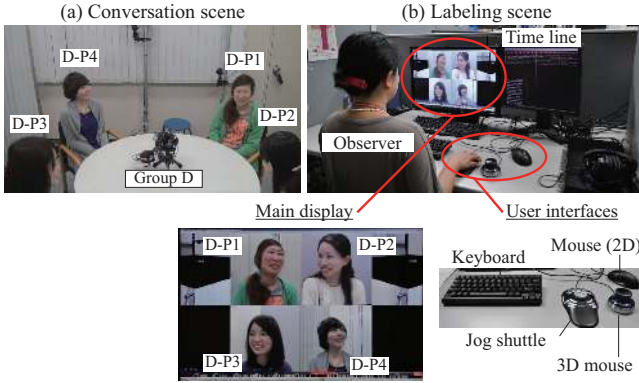
Fig. 2. (a) Conversation scene, and (b) labeling environment.

topic. The discussion topics included "Who is more beneficial, men or women?", "Are marriage and romantic love the same or different?" The discussion topics were assigned to the groups on the basis of the participants' opinions expressed in previously completed questionnaires and designed to cause frequent agreement and disagreement. Focusing on the liveliest exchanges, this paper selects up and analyzes ten discussions: four from $\mathcal{G}_\mathcal{A}$ and two each from $\mathcal{G}_\mathcal{B}$ to $\mathcal{G}_\mathcal{D}$. The average discussion length was 7.4 min (1.4 min S.D.).

Conversations were captured at 30 fps by using an IEEE1394 camera for each participant ($\mathcal{G}_\mathcal{B}$ and $\mathcal{G}_\mathcal{C}$) or a tabletop device for round-table meetings [49] ($\mathcal{G}_\mathcal{A}$ and $\mathcal{G}_\mathcal{D}$). The sizes of the captured color images were 640×480 pixels (bust shot) and 2448×1024 pixels (omnidirectional), respectively.

3.3 Procedure and apparatus

The observers were asked to watch the conversation videos and to assign the label from the following lists that most closely represented their impression of each pair of participants each time: (1) "Strong empathy", (2) "Weak empathy", (3) "Neither empathy nor counter-empathy", (4) "Weak counter-empathy", and (5) "Strong counter-empathy".

We try to capture the perceived empathy as clearly as possible without giving the observers any instruction that could distort their intuitive perceptions. Specifically, with regard to the definition, empathy has the characteristic that "most people understand and share a common meaning but for which it is impossible to provide an adequate definition"; such a concept is called projective content [50]. By following the guideline in [50], our instructions to the observers contained neither technical terms nor procedural definitions such as a long list of detailed rules whose use would almost automatically distinguish the type of perceived empathy from participant behaviors.

Five of the observers labeled all the conversations, while the remaining four processed only the $\mathcal{G}_\mathcal{A}$

conversations. That is, the number of observers, M, for each conversation is five or nine. The observer could replay the video as many times as desired. We asked each observer to finish the labeling of one conversation within one day (7.5 hours), and most observers succeeded in doing so. All labeling was done in isolation.

The observers labeled the video sequences without accessing the audio signals to focus on empathy exchanged via visual nonverbal behavior. When the target is to infer the *internal states* of a person, many studies have reported that multimodality, e.g. audio-visual fusion, is actually advantageous [20]. However, it makes it difficult to explore the impact of visual behaviors on *observer perception*, as demonstrated in [51], when targeting rapport perception.

The labeling was not frame-by-frame but region-by-region. That is, the frames at which the observer's impression changed were extracted, and then the sequence of frames between two labels was assigned the label of the head frame of the sequence. By considering that a pair of participants are interacting only if at least one of them is looking at the other, the present study excludes averted gaze states when labeling targets. This is because we target visual communication, so it is reasonable to assume that no visual message is directly sent/received between the pair in averted gaze states. It is a practical way of realizing annotation at an acceptable cost, because the averted gaze accounts for over a half of our dataset, as described in 3.4.1.

Our original software, *NTT-CSL Conversation Scene Viewer*, was used for viewing and labeling videos via interfaces, as shown in Fig. 2 (b). The videos could be played at normal speed or the speed could be changed by turning a jog shuttle. A 26-inch and a 16-inch monitor were used; the larger one was used to display a movie that showed all the participants at quarter-size, while the smaller one was used to display the timeline of a sequence of given labels.

3.4 Statistics of perceived empathy

The statistics of the annotated empathy labels by our observers were: Empathy (strong: .11 and weak: .39), Neither (.48), and Counter-empathy (weak: .025 and strong: .0017). The frequencies of Empathy and Neither are comparable, while Counter-empathy was very infrequent. The imbalance makes the five-point distributions created by the five or nine labelers too sparse for an analysis of distribution types. Accordingly, this paper simply distinguishes Empathy or not[1]; i.e. categories (3-5) are grouped as the "No

1. We also conducted an evaluation of the inference performance of the proposed model in 5.3 with a different grouping, where we ignore label strength, i.e. groups of (1)&(2), (3), and (4)&(5), although the samples of the last group are small, and using the original labels in the five-point scale. Consequently, comparable results were obtained in both cases.

empathy" class. That is, there are three categories of perceived empathy, $C = 3$, in the present study: "Strong empathy", "Weak empathy", and "No empathy". In this case, the average distribution of perceived empathy, $\bar{\boldsymbol{p}}$, was $\bar{\boldsymbol{p}} = (.11, .39, .50)$.

The constraint of the probability distribution, i.e. the summation of elements must be one, means that the degree of freedom (df) is $C - 1$. So, three is the minimum number of states that is meaningful to this study. If $C = 2$, the problem is similar to common ones, i.e. only a single real value is inferred, e.g. the inference of FE intensity [52]. Moreover, the present study ignores the order of these categories, i.e. it handles perceived empathy labels as nominal data. If a researcher needs to handle them as ordinal data, the validity should be carefully assessed. But, this exceeds the scope of this paper.

### 3.4.1 Inter-coder agreement of annotation

Three measures are often used to evaluate the reliability of annotation: reproducibility, stability, and accuracy [31]. Each, in short, measures the degree of matching between different coders (i.e. inter-coder agreement), different times (i.e. intra-coder difference), and a target annotation and gold standard decided by a theory or expert, respectively. This section targets reproducibility. Stability is discussed in 3.4.2, and accuracy is basically meaningless in our case, because we consider the inter-observer difference to be essential.

We first computed three inter-coder agreement measures for perceived empathy: Intraclass Correlation Coefficients (ICCs)[2], which is well summarized in [53], Rosenthal's effective reliability coefficient R_{SB} [54], and Conger's Kappa coefficient κ [55][3]. The results yield two summaries. First is the agreement of mean labels among coders; ICC(1, k) = .846/.709, ICC(A, k)=.853/.732, ICC(C, k)=.894/.794, and R_{SB}=.897/.823, where the left and right parts separated by slashes denote the coefficients of $\mathcal{G}_A$ (k=9) and the other groups (k=5), respectively. This suggests that the mean labels output by our labelers can be considered reliable [53]; so we compare our proposed approach with the consensus approach described in 5.4. Second, and more importantly, our data shows a low average correlation between a pair of coders: ICC(1,1) = .379/.328, ICC(A,1) = .392/.353, ICC(C,1) = .485/.436, and mean pair-wise correlation r = .481-.492, and κ = .249. These data support our basic claim, namely that the labelers often disagree.

TABLE 1
Frequencies of impression distribution types

	Type	Ex.	[%]		Type	Ex.	[%]
X-dominant	Strong emp.-dominant		2.3	X-inferior	Strong emp.-inferior		31.4
	Weak emp.-dominant		17.5		Weak emp.-inferior		0.9
	No emp.-dominant		36.8		No emp.-inferior		10.4
	Flat		0.6		Probabilities		

Strong emp. ■ Weak emp. ■ No emp.

To clarify how the perception differs among the observers, Table 1 categorizes the perception *distribution* into the following seven *distribution types*; "X-dominants", "X-inferiors", and "Flat"; where "X" means a perception of Strong empathy, Weak empathy, and No empathy[4]. These types are classified based on the standard deviation and skewness of the target distribution: If the standard deviation (SD) is equal to or greater than $1/C^2$, type = "Flat". Otherwise, the skewness is positive, type = "X-dominant", or type = "X-inferior". Note that when $C = 3$, the sign of skewness means the number of elements whose probabilities exceed the mean (i.e. $1/C$). As a result, conflicting cases, i.e. X-inferior and Flat types, accounting for as much as 43% can be found. These results reinforce the importance of treating the impressions as distributions, instead of trying to select a single state via majority voting.

3.4.2 Stability of perceived empathy distribution

Next, the stability of the perceived empathy distribution is investigated via a test-retest procedure. One observer reassessed all the conversations after an interval of about eight months. The ratio of frames given the same label to all target frames, i.e. the overall percent agreement, p_a, was .729. Some readers might consider that the ratio is not very high, but stability is the weakest form of reliability [31]. Our perceived empathy is probably affected by factors such as mood, as found with the recognition of FEs in [56]. However, more importantly, the fluctuations in a distribution of perceptions made by multiple observers tend to cancel each other out since we can assume that each observer will independently change his/her label on any two given occasions. Substituting p_a into the Spearman-Brown formula yields the stability of *the collective impressions* of the observer group, i.e. the stability of the distribution. The results are .931 with $M = 5$, and .960 with $M = 9$. Compared with the inference accuracy described in 5, these values are sufficiently high for intra-observer variation over time to be ignored.

3.4.3 Validity of number of observers

We discuss the validity of the number of observers by estimating its impact on the evaluation of inference performance. For simplicity, we make the following three assumptions: 1) M observers are randomly sampled from a population observer group. 2) The target is an instantaneous scene where a pair of participants are interacting. 3) Each observer has a predefined label for this scene, and the observer population for each category is proportional to the average distribution of perceived empathy, $\bar{p}$, described in 3.4.1. In this case, we evaluate the expectation of a distribution similarity $\mathcal{S}$ between $\bar{p}$ and the possible distributions made by these M observers.

In theory, the probability of each possible distribution occurring follows a multinomial distribution $Mult(\boldsymbol{m}|\bar{\boldsymbol{p}}, M)$, where $\boldsymbol{m} = (m_1, \cdots, m_e, \cdots, m_C)$ and m_e denotes the number of observers who vote for category e. Accordingly, the expected distribution similarity is calculated as $\sum_{\boldsymbol{m}} Mult(\boldsymbol{m}|\bar{\boldsymbol{p}}, M) \cdot \mathcal{S}(\boldsymbol{p_m}, \bar{\boldsymbol{p}})$, where $\boldsymbol{p_m}$ is the probability distribution for a set of votes $\boldsymbol{m}$. Among the various similarity measures $\mathcal{S}$ possible for use in comparing two probability distributions [57], the present study selects the overlap area (OA), because it is a widely used form of similarity and is strongly related to the Bayesian framework [58][5]. OA is calculated as $\mathcal{S}(\boldsymbol{p}, \hat{\boldsymbol{p}}) := \sum_{i=1}^{C} \min(p_e, \hat{p}_e)$, where p_e and $\hat{p}_e$ denote the e-th component of $\boldsymbol{p}$ and $\hat{\boldsymbol{p}}$, respectively. OA has a maximum value of one, i.e. two distributions are exactly the same, and a minimum value of zero, i.e. no overlap at all.

The results with $M = 5$ and $M = 9$ were .759 and .829, respectively. They are much higher than the inference accuracy, described in 5. Thus, we consider that the number of observers is sufficient for the evaluation described in this paper. Moreover, the SD of the probability of each component, i.e. m_e/M, in the possible distributions is proportional to $1/\sqrt{M}$. So, for example, to reduce the SD by half requires a total of $4M$ observers. Likewise, to further increase the expected distribution similarity, a similar number of additional observers would be required.

3.5 Key participant behaviors and their coding

Our assumption that the observer's impressions are driven by the combined behaviors of a participant pair was derived from the following two facts. First, the only cues to the pair's empathic states for the observer are the pair's behaviors; they did not have any other knowledge about the pair's characteristics. Second, the pair's behaviors often coincided with regard to empathy, i.e. "behavioral coordination" [6].

Of the behaviors, we focus on FE and gaze. FE plays a major role in conveying empathic messages [15]. The consistency and inconsistency of FEs between a pair is closely related to their empathy [59], [60]. In addition, gaze is vital when inferring empathy from the FEs (monitoring). Furthermore, gaze triggers a gaze shift and a reaction from the gazee, namely the person being looked at, toward the gazer, namely the person who is looking. As a result, the behaviors of a pair in a mutual gaze state are likely to correspond [60]. Accordingly, whether consciously or unconsciously, observers can be expected to perceive others' empathy mainly from these cues. Moreover, head gestures are often highly correlated with FEs in our data, e.g. nodding with smiling, or tilting without smiling, and so we decided to focus on FE and gaze in this paper.

These behaviors were manually annotated in this paper. One of the observers was asked to assign gaze behavior. The type of gaze pattern between a pair was either {"mutual", "one-way", or "averted"}. The frequencies of the gaze labels are .09, .29, and .62, respectively.

FEs were annotated by the coder and two additional coders and their majority label was used as the FE label to ensure the reliability of the annotation. Each of the coders grouped FE into seventeen categories in this study: neutral, smiling, laughter, chuckle, thinking, surprised, embarrassed, wry smile, disgust, bored, provoked, puzzled, sad, angry, afraid/fearful, disbelieving, and other expressions. These categories were empirically selected due to their strong expected relationship with perceived empathy with the reference to FACS [16] and mind reading manual [61]. In the coding, one sample image for each of most of the categories was given to the coders. The images were selected in advance from the conversation videos by the authors. We did not give any additional verbal instruction for any category.

Because the 14 labels other than neutral, smiling and thinking were very infrequent ($< 1\%$)[6], the present study groups the 17 categories into four categories: "neutral", "smiling", "thinking", and "others." Labels of laughter and the remaining infrequent 13 categories were first grouped with the smiling and "others" categories, respectively; then their majority label was set as the final label. Moreover, if there was no majority label, e.g. when the labels of the three coders were completely different, it was considered to fall into the "others" category[7]. The frequencies of the majority FE labels are as follows: neutral = .55,

---

5. For example, OA is theoretically the same as the Bayes minimum misclassification (or error) probability [58]. But, it is difficult to say that OA is the best measure. Actually, we found that the other measures yielded similar results to OA. So, we consider that these measures are comparable.

6. A possible reason for the unbalanced results, i.e. the infrequent labels of counter-empathy and negative FEs, is that participants tried to establish good relationships by often showing friendly expressions, i.e. smiling, rather than trying to get their own way, even though the participants had quite different opinions as indicated by their pre-conversation questionnaires.

7. We have already tried to give such no-majority cases a fifth label, i.e. a new label, but similar inference performance, as described in 5, was obtained.
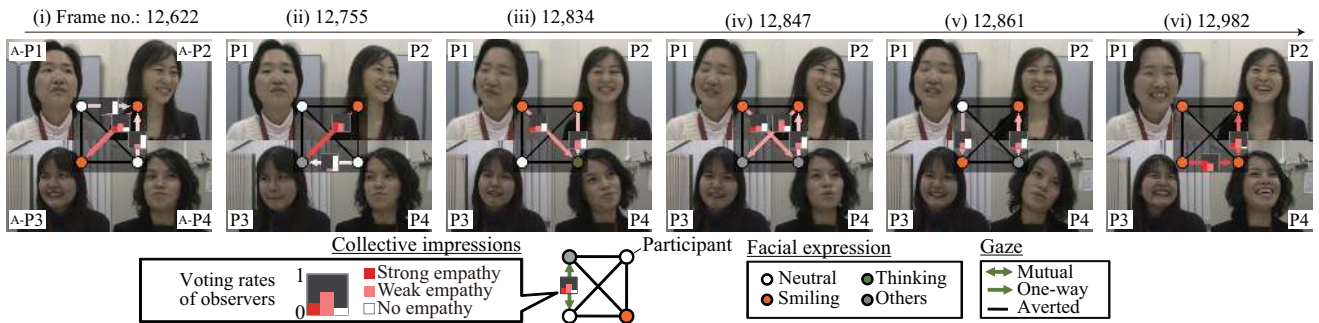
Fig. 3. Example sequence of perceived empathy created by observers for $\mathcal{G}_A$. Each node indicates a participant, and the bar chart on an edge between the participants indicates the perceived empathy distribution for the pair.

smiling = .35, thinking = .02, and others = .08.

To validate the gaze annotations by following the inter-coder agreement test, two additional female coders in their twenties or thirties labeled gaze in one conversation session; i.e. the total number of gaze coders in this evaluation was also three. The resulting Conger's Kappa coefficients are $\kappa = .887$ for gaze and .500 for FE (the four categories). According to the benchmarks in [53], the gaze annotation is judged to be excellent, while the FE annotation is judged moderate.

3.6 Case study with perceived empathy labels

Figure 3 shows snapshots of a short scene, where observers demonstrate widely different types of perceived empathy[8]. Hereinafter, the participants in Fig. 3 are called P1 (upper-left), P2 (upper-right), P3 (lower-left), and P4 (lower-right). In this scene, the conversation proceeded as follows. First, P2 attempted to persuade P3 by smiling in (i), then P3 was nodding and smiling in (ii); i.e. the pair consisting of P2 and P3 (hereinafter denoted as P2-P3) were looking at each other and smiling. Almost all the observers assigned empathy although of different intensity, i.e. Strong or Weak empathy, to their interaction (shown as red or pink bars). On the other hand, pairs P1-P2, P2-P4, and P3-P4, where P1 and P4, in their one-way gaze at P2 or P3, did not clearly smile, and were labeled by most observers as No empathy (shown as tall white bars).

In (iii), P2 asked for others' responses while looking around, but in (iv) and (v) everyone looked confused. None of the corresponding perceived empathy distributions have significant peaks, that is the impressions were quite different among the observers. This incoherent distribution may be due to both the varied combination of FEs between participants and the ambiguity created by the FEs. In this scene, only P2 is clearly smiling, while the other participants are responding with partial smiles or other delicate FEs with heads tilted (P1 and P4). In (vi), finally, P4

reacted cheerfully to P2 and then everyone laughed and seemed to relax. Almost all the observers assigned Empathy to pairs P2-P4 and P3-P4, although the perceived strength was different among the observers. As in this case, when both participants in each pair clearly exhibited positive FEs, namely smiling or laughter, the observers usually judged such interactions as Empathy. Note that, P1 is laughing in (vi), but she has her eyes' closed. So, P1-P4 in the averted gaze state was not annotated.

4 COMPUTATIONAL MODEL

Our aim is to develop a computational model that relates observers' perceived empathy to the participants' gaze and FEs. To handle the expected variety of ambiguities, this paper uses probabilistic modeling, namely a Bayesian network (BN) [14]. Our probabilistic model explicitly represents the structural relationship between the elements, including static dependencies and independencies. By using the trained model, the perceived empathy distribution of each pair on each occasion can be inferred from their gaze and FEs at that time.

4.1 Bayesian network

We evaluate simple BNs to confirm the fundamental validity of the proposed framework, although many models of perceived empathy can be considered. BNs make it easy to add other modalities subsequently, e.g. vocal cues, and/or other psychological findings or assumptions.

Although perceived empathy is, in practice, expected to be dependent between pairs of participants, this paper assumes independence to simplify the mathematics. Figure 4 shows the BN for a pair (i, j), i.e. the pair of participants i and j. This figure draws a time-slice at time t, and the structure is time-invariant. Nodes represent random variables. Edges represent dependencies between variables, and are modeled with parameters φ. The parameters are time-invariant, but may be different between participant pairs. The average tendency as regards how observers perceive

---

8. Parts of the video sequences are available at http://www.brl.ntt.co.jp/people/kumano/research/empathy.htm.
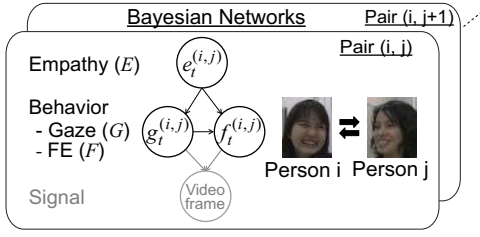
Fig. 4. Graphical representation of the proposed model that describes the static relationship between the perceived empathy of external observers for a pair (i, j) and the pair's gaze and facial expressions (FEs).

empathy from participant behaviors is expressed as a prior distribution of the parameters with hyperparameters θ as prior knowledge.

To focus on exploring the relationship between perceived empathy, gaze and FEs, this paper assumes that these behaviors are already known. To realize the automatic inference from audio-visual signals, we have already proposed automatic gaze detection [62] and facial expression recognition [63]. For a preliminary experiment on automatic FE recognition for the inference of perceived empathy, see [47].

4.2 Proposed model

The inference target is the joint probability distribution of a sequence of the perceived empathy for pair (i, j), $E^{(i,j)}$, and model parameters $\varphi^{(i,j)}$, given by a participants' gaze sequence, $G^{(i,j)}$, and FE sequence, $F^{(i,j)}$, and hyperparameters θ; i.e. $P(E^{(i,j)}, \varphi^{(i,j)}, |G^{(i,j)}, F^{(i,j)}, \theta)$. Hereafter, φ, θ, and the index of the pair (i, j) are omitted unless necessary.

By following Bayes' rule, the conditional probability is proportional to the joint probability, namely $P(E|G, F) \propto P(E, G, F)$. It is decomposed to yield

$$P(E, G, F) := P(E)P(G|E)P(F|G, E). \qquad (1)$$

$P(E)$ is the marginal probability of perceived empathy. $P(G|E)$ is the likelihood of perceived empathy for observed gaze. $P(F|G, E)$ is the likelihood of perceived empathy for the combination of observed gaze and FEs.

The only term that explains the frequency of perceived empathy is $P(E)$. This term indicates that the more frequent category in the training data has higher probability in the joint probability distribution, and vice-versa. However, our preliminary experiment demonstrated that this term often degrades the inference performance. This would be because this term is overtrained due to the imbalance in perceived empathy in our conversation data, as described in 3.4.1. Thus, this study uses a uniform distribution, where each type of perceived empathy is assumed to occur with the same probability.

The likelihood $P(G|E)$ describes the tendency for empathy impressions to be caused by mutual gaze

and one-way gaze. Even when the gaze of a pair is averted, our algorithm continues to infer the joint probability from the first frame to the last frame. The likelihood $P(G|E)$ is taken to be uniform for averted gaze. However, frames in averted gaze are ignored when calculating the inference performance in 5.3.

4.3 FE co-occurrence probability tables

Our model is characterized by how likely observers are to ascribe empathy to a pair given the co-occurrence of facial expressions between them, $P(F|G, E)$. We refer to the model, although it is simply a probability table, as the FE co-occurrence probability table. FE co-occurrence probability tables are prepared separately for mutual gaze and one-way gaze states.

The likelihood $P(F|G, E)$ is decomposed to yield

$$P(F|G, E) = \prod_{t=1}^{T} P(f_t|g_t, e_t), \qquad (2)$$

where T denotes the sequence length. The right term denotes the likelihood of perceived empathy for a pair at time t, e_t, when the gaze state between the pair is g_t and the co-occurrence of their FEs is f_t. As described in 3.3, perceived empathy $e \in \{1, \cdots, C\}$. The gaze state g is in {"mutual", "one-way", "averted"}. The FE co-occurrence f is described by an element of a 4×4 table, where the FE of each person is in one of the four categories explained in 3.4.1.

Figure 5 shows FE co-occurrence probability tables formed entirely from conversation data. For example, the co-occurrence of smile, $(2, 2)$ in both gaze states, indicates that most observers judged such scenes as exhibiting Strong or Weak empathy. The differences between the bar charts of the FE co-occurrence and gaze states suggest that they are the key cues for perceived empathy.

4.4 Inference of joint distribution

We employ a Bayesian approach to infer the joint distribution of all unknown variables for given measurements. In Bayesian analysis, a priori knowledge about the model parameters φ are represented as prior distributions, $p(\varphi|\theta)$, where θ denotes the a set of parameters of this distribution (i.e. hyperparameters). This distribution is omitted in 4.2. This paper employs natural conjugate prior distributions [64] for mathematical convenience. In particular, we use independent Dirichlet distributions which are commonly used as the prior of discrete random variables. Hyperparameters θ are set to be proportional to the frequencies of target events in the training data. The hyperparameters are constant among the participant pairs, whereas the parameters differ. The joint distribution is unfortunately hard to calculate precisely due to its complexity, so we utilized the Gibbs sampler [65], a
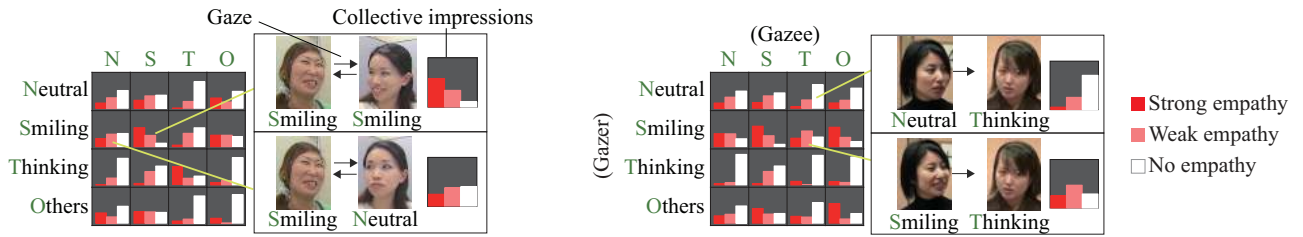
Fig. 5. FE co-occurrence probability tables $P(F|E,G)$ for mutual gaze (left) and one-way gaze (right).

variant of the Markov Chain Monte Carlo (MCMC) method. See [47] for more details of the definition of the probability distribution and the inference by using the Gibbs sampler.

5 INFERENCE PERFORMANCE EVALUATION

We propose a method for quantitatively evaluating a model of perceived empathy that uses the similarity of the joint distributions produced by the model to the distributions made by external observers.

This paper employs the leave-one-conversation-group-out cross validation. This evaluates how well perceived empathy distributions created by an unseen conversation group can be replicated by the model, the hyperparameters of which are trained by using data except those for the target conversation group. For example, when the impressions for conversation group $\mathcal{G}_\mathcal{A}$ are inferred, the hyperparameters are validated by using data of $\mathcal{G}_\mathcal{B}$, $\mathcal{G}_\mathcal{C}$, and $\mathcal{G}_\mathcal{D}$.

Among the various similarity measures between two probability distributions [57], this paper considers two common metrics [58]: overlap area (OA) for measuring the accuracy, as described in 3.4.3, and root mean square error (RMSE), $\mathcal{S}(\boldsymbol{p}, \boldsymbol{p}') := \sqrt{\sum_{e=1}^{C}(p_e - p'_e)^2/C}$, for measuring the inference error[9]. In summary, for OA, larger is better, while for RMSE, smaller is better (the best is zero).

5.1 Baseline models

Eight baseline models of three types were prepared. The first type uses only gaze or FE. The gaze-only model returns $P(E, G) = P(E)P(G|E)$. The FE-only model outputs $P(E, F) = P(E)P(F|E)$, where $P(F|E)$ is equivalent to $P(F|G, E)$, explained in 4.3, by ignoring the gaze state.

The second type consists of three simple models: Baselines 1, 2, and 3. Baseline 1 continually outputs a flat distribution $(1/3, 1/3, 1/3)$. Baseline 2 always returns the average distribution of our data $\bar{\boldsymbol{p}}$. Baseline 3 returns a distribution where the probability of the majority class is one, and those of the rest are zero; for example, if the numbers of votes is $(3, 5, 1)$, this

9. Given the difficulty of deciding which measure is the best for the present study we also tried other similar measures, including Bhattacharyya coefficients, cosine similarity and ranking loss [57], and found that they yield comparable results to OA and RMSE.

model outputs $(0, 1, 0)$. Note that it uses the ground truth distribution.

The last type is based on a traditional consensus approach. The proposed approach differs from it only in terms of training and performance evaluation; both the compared approaches use the same model, which is the proposed BN described in 4. The consensus approach trains models only by using *majority labels* that are made for each sample, while the proposed approach uses *all the labels* obtained from annotators. In testing, the consensus approach evaluates the models based on accuracy, i.e. the frequency with which they can find the *majority labels*. On the other hand, the proposed approach uses how precisely the models can reproduce the label *distribution*.

Accordingly, two types of experiments are performed for comparing these approaches. The first experiment follows the evaluation of the consensus approach, while the second follows the evaluation of the proposed approach. The first experiment treats their output as single labels, whereby the proposed approach is forced to output a single label with maximum probability. The second experiment compares their outputs as distributions. That is, these two methods are compared on how accurately they can infer joint probability distributions.

Note that the first experiment uses only samples where the coders reached a consensus (such as X-dominant types), because such an agreement in annotation is inherently assumed in the consensus approach. The condition of the successful consensus is that the voting rate for a perceived empathy type exceeds a threshold, $\tau = .5/.75/1.0$. Although seldom required in the consensus approach, the threshold is introduced to approximately assess the impact of the extent of inter-coder agreement on the performance.

5.2 Hypothesis testing

The proposed model is compared with $K - 1$ baseline models ($K = 9$) by using the Friedman test followed by the Nemenyi test, as recommended in [66]. The null hypothesis is that all the models are equivalent and so their average performance ranks should be equal. To eliminate the temporal dependency of our perceived empathy data, we use the mean ranks of the four conversation groups (group-level tests), and the mean ranks in 40-sec non-overlapping thin slices (temporal

windows) of the data (block-level tests)[10]. The group-level tests aim at evaluating the generalizability of the models to brand-new four-person groups. However, because we have few groups, we also perform the block-level tests to simulate a situation with larger sample size ($N = 4$ for group-level tests, and 215 for block-level tests). To verify the statistical impact of conversation groups, sessions and pairs on the test statistics, bootstrapping [67] is employed.

The test proceeded as follows: first, K models were ranked in each frame according to their accuracies, and the average ranks were calculated in each group or each block; this yielded N original test samples for each model. Second, the overall rank difference between K models was assessed by the Friedman test. Consequently, the null hypothesis was rejected at $p < .05$ for all the tests. Thus, in the post-hoc pairwise comparison, N rank differences between the models were then obtained for the original samples and 100,000 bootstrap samples[11]. Finally, the p-value was obtained as the rate of the bootstrap samples that yielded a larger absolute difference in mean rank than the original absolute rank difference. In the post-hoc test, the critical values were corrected by dividing by $\sqrt{K(K+1)/6N}$.

5.3 Comparison with baseline models

Table 2 shows the inference accuracies obtained with the proposed and baseline models. The proposed model statistically outperforms most baseline models to a significant degree. Although the baseline with the average distribution yields non-significant differences for both similarity measures at the group level ($p > .05$), the corresponding effect sizes ($r = .69$ for both), calculated as $r = z/\sqrt{2N}$ with reference to [68], are judged as "large" according to Cohen's criteria [69]. At the block level, all the models show statistical significance ($p < .005$). These results suggest the validity of our hypothesis, namely that modeling variability in annotators is important, and FEs and gaze are key nonverbal behaviors as regards perceived empathy.

Table 3 shows the inference accuracy (OA) for each perceived empathy distribution type. This helps us to understand the characteristics of the proposed model in more detail, although further analysis is required to explore the impact of each distribution type on the conversation. Table 3 suggests that gaze and FE contribute to different distribution types; specifically, FE is needed to infer the distributions that include

10. We first merged all frames into a single sequence (vector), then divided it by a fixed length. The block size was determined so that the autocorrelation of the merged sequence converges to a sufficiently low value. Consequently, the autocorrelation is 0.008 with block size of 1,200 frames.

11. Although a larger bootstrap sample size is more desirable, the statistics sufficiently converged with this size in our experiment.

TABLE 2
Average inference accuracy of perceived empathy

Model	OA↑ (rank↓)	RMSE↓ (rank↓)
Proposed		
Gaze + FE	.721 (4.14/4.01)	.208 (4.10/3.97)
Gaze only	.692 (4.66*/4.69***)	.227 (4.62*/4.64***)
FE only	.635 (5.72**/5.71***)	.260 (5.63**/5.63***)
Baseline		
Flat dist.	.605 (6.28***/6.27***)	.287 (6.21***/6.21***)
Avg. dist.	.684 (4.74/4.88***)	.230 (4.67/4.77***)
Majority	.701 (4.97/5.23***)	.232 (5.06/5.33***)
Consensus		
$\tau = .5$	.719 (4.25/4.16*)	.214 (4.28/4.19***)
$\tau = .75$	.703 (4.76/4.66***)	.228 (4.80/4.72***)
$\tau = 1.0$	.668 (5.49**/5.39***)	.257 (5.63**/5.52***)

"↑" and "↓" denote higher and lower performance. OA and RMSE values are the averages for target frames and participant pairs. Left and right hand sides in each bracket indicate the means of group- and block-level ranks, respectively. Symbols "*", "**", and "***" denote $p < .05$ $p < .01$, and $p < .005$, respectively, for the difference in the mean ranks from the proposed gaze + FE.

the perception of Strong empathy, such as Strong-Empathy-dominant distributions, and No empathy-inferior distributions.

5.4 Comparison with consensus approach

The aforementioned two types of experiments demonstrated that the proposed approach statistically significantly outperforms the consensus approach.

In the first experiment, the correct recognition rates, which are the hit rate of majority labels, of the proposed and consensus approaches are .818 and .752 (frame average), respectively, and $p < .005$ (the Wilcoxon signed-ranks test through a similar procedure in 5.2, where the number of test samples was 214). The above recognition rates were obtained with full agreement ($\tau = 1.0$). Similar results were obtained for different τ values between .5 and 1.0.

Table 2 summarizes the results of the second experiment. The significance decreases according to the decrease in τ. For example, the consensus approach with $\tau = .5$ yields the most comparable results for the proposed model; $p > .05$, $r = .57$ (large effect size [69]) at the group level, and $p < .05$, $r = .12$ (small effect size [69]) at the block level. These results well summarize the characteristics of our model. The generalizability of the proposed method to a new conversation group is greater than that of the consensus approach to a practically significant degree (at least a small effect size), while to obtain the statistical significance the conversation length should be roughly equal to or longer than 7.4 min (the average conversation length in our data); such a case is common in particular for multi-party conversations.

Moreover, as expected, the results show that the proposed approach is superior, especially for scenes where coders disagree marginally. The inference accuracies of the proposed and consensus approaches with $\tau = 1.0$ are .721 and .668 in total; .614 and

TABLE 3
Inference accuracy for each distribution type of perceived empathy with regard to the distribution similarity OA

Model	Frame avg.	Type avg.	Type 1 (SE-dom)	Type 2 (WE-dom)	Type 3 (NE-dom)	Type 4 (SE-inf)	Type 5 (WE-inf)	Type 6 (NE-inf)	Type 7 (Flat)
Proposed model (gaze + FE)	.721	.665	.464	.630	.747	.809	.582	.647	.775
Gaze only	.692	.636	.397	.595	.697	.841	.590	.551	.779
FE only	.635	.694	.751	.563	.601	.700	.698	.734	.807

SE, WE, and NE mean Strong empathy, Weak empathy, and No empathy, respectively.

.564 for X-dominant distributions, and .703 and .556 for other distributions, respectively. These results are supported by a simulation similar to that described in 3.4.3, where expected distributions are modeled with a multinomial distribution. If the inter-coder agreement is marginal, the consensus method somewhat ignores the tails of distribution. For example, if the event probability distribution of the multinomial distribution is equal to the actual mean distribution $\bar{p} = (.11, .39, .50)$ and the number of observers is nine, then the probability that each class becomes the majority is $(.008, .3, .6)$ by omitting ties; the resulting distribution is biased towards the majority.

Moreover, the proposed and the consensus methods are identical in extreme cases where numerous coders fully randomly or consistently give labels to samples. Of particular note is the former case, where the event probability is $(1/3, 1/3, 1/3)$, and the probability that each class becomes the majority is equal to the event probability. This means that both methods train the model by using a flat distribution. However, most consensus approached do not accept this case where the majority changes at random among the samples.

6 DISCUSSION

We consider that the validity of the proposed framework in analyzing the process of perceived empathy was basically confirmed by the quantitative evaluation. However, this research is still in the development phase, and further advancement is possible.

First, the categorical treatment of FEs in the present study did not show excellent inter-coder agreement. To increase the agreement of the FE annotation, physical-motion-based descriptions would be more appropriate, for example Ekman and Friesen's FACS coding of FEs [16]. Such a description would also be helpful in realizing a person-independent FE recognizer. Other possible choice is to handle the FE labels of multiple coders as distributions, as with empathy perception labels. However, this does not appear to be very effective, because the inter-coder agreement of FE ($\kappa = .500$) is much higher than that of empathy perception ($\kappa = .249$). Moreover, despite the employment of the fewer coders, the reliability statistics appear to be higher than those in some previous studies, e.g. [70]. Possible reasons for this are that, unlike the previous studies, the sex, age, ethnicity, and culture of the interlocutors and observers are similar or the same, and the observers viewed the videos without accessing the audio signals.

Next, the proposed models only describe instantaneous relationships between participant gaze and FEs and observer impressions. However, observers would also focus on other behavioral cues such as head gestures and proximity, as well as prosody and verbal messages with recourse to audio. In addition, the time lag of behaviors between a pair (action and response) would inherently affects the perceived empathy. Furthermore, although this paper assumes the independence of pairs, social pressure effects [13], for example, suggest that this is not always the case. It would be very interesting to relax this assumption.

To further explore the perception process, more samples of perceived counter-empathy are required from more conflicting situations such as debates and marriage counseling/discussion [44]. In addition, although the present study employed a forced choice using a 5-point scale, continuous metrics, such as [23], would be more suitable. Furthermore, the perceptions would differ with the observation environment. First, a limitation of the present study is that the gaze patterns were given to the observers in advance. The independent labeling of the gaze patterns would make it possible to investigate what features the observers focused on. Second, the observers in the present study were generally able to continue watching both participants at the same time on the regular-sized monitor. However, participants in face-to-face conversation necessarily miss some of the others' visual behaviors due to the limited field of view of the human visual system.

Another interesting topic would be to explore the way in which the observer is affected by perceiving the empathy between a pair. According to the perception-action model [48], perceiving empathy between a pair could automatically trigger behavioral coordination and a feeling of empathy in the observer to the pair. If the observer is a real side participant, his/her empathic response will newly arouse other side participants to perceive empathy between him/her and the pair.

7 CONCLUSION

This paper presented a research framework for understanding the empathy aroused during conversa-

tion. By focusing on the empathy shared between pairs of people, we introduced the idea of objectively describing observers' impressions as a collection of impressions of external observers, and then set the problem of creating a model that could use participant behaviors to generate valid distributions of impressions. An experiment employing the proposed evaluation method demonstrated that the proposed computational model well recreates the distribution of observers' impressions from the co-occurrence of participants' gaze and facial expressions. This research is still in its developmental phase, and various possible enhancements were discussed.

REFERENCES

[1] D. Gatica-Perez, "Analyzing group interactions in conversations: a review," in *Proc. IEEE Int'l Conf. Multisensor Fusion and Integration for Intelligent Systems*, 2006, pp. 41–46.

[2] K. Otsuka, "Conversation scene analysis," in *IEEE Signal Processing Magazine*, vol. 28, 2011, pp. 127–131.

[3] M. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer, "The first facial expression recognition and analysis challenge," in *Proc. Workshop AVEC*, 2011.

[4] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "AVEC 2011 - the first international audio/visual emotion challenge," in *Int'l Audio/Visual Emotion Challenge and Workshop (AVEC 2011)*, 2011.

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

[6] 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.

[7] P. M. Niedenthal, "Embodying emotion," *Science*, vol. 316, pp. 1002–1005, 2007.

[8] E. Brunswik, *Perception and the representative design of psychological experiments*. Berkeley: University of California Press, 1956.

[9] N. Chovil, *The psychology of facial expression*. Camhridge UK: Camhridge University Press, 1997, ch. Facing others: A social communicative perspective on facial displays, pp. 321–333.

[10] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Commun.*, vol. 40, no. 1-2, pp. 227–256, 2003.

[11] M. Knapp and J. Hall, *Nonverbal Communication in Human Interaction (7th edition)*. Belmont, CA: Wadsworth, 2010.

[12] S. Baron-Cohen, A. Leslie, and U. Frith, "Does the autistic child have a 'theory of mind'?" *Cognition*, vol. 21, pp. 37–46, 1985.

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

[14] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann, 1988.

[15] N. Chovil, "Discourse-oriented facial displays in conversation," *Res. on Lang. and Social Int.*, vol. 25, pp. 163–194, 1991.

[16] P. Ekman and W. V. Friesen, *The Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, 1978.

[17] R. E. Thayer, *The biopsychology of mood and arousal*. New York: Oxford University Press, 1989.

[18] 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," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.

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

[20] 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.

[21] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[22] T. Bänziger and K. Scherer, "Using actor portrayals to systematically study multimodal emotion expression: The GEMEP corpus," in *Proc. Int'l Conf. ACII*, 2007, pp. 476–487.

[23] A. M. Ruef and R. W. Levenson, *Handbook of emotion elicitation and assessment. Series in affective science.* New York, NY, US: Oxford University Press, 2007, ch. Continuous measurement of emotion: The affect rating dial, pp. 286–297.

[24] D. Reidsma and R. A. op den, "Exploiting 'subjective' annotations," in *Workshop on Human Judgements in Computational Linguistics, Coling*, 2008, pp. 8–16.

[25] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, vol. 18, no. 4, pp. 407–422, 2005.

[26] J. Broekens, A. Pronker, and M. Neuteboom, "Real time labeling of affect in music using the AffectButton," in *Proc. int'l workshop AFFINE*, 2010, pp. 21–26.

[27] G. Laurans, P. M. A. Desmet, and P. Hekkert, "The emotion slider: A self-report device for the continuous measurement of emotion," in *Proc. Int'l Conf. ACII*, 2009, pp. 1–6.

[28] E. L. Rosenberg and P. Ekman, "Coherence between expressive and experiential systems in emotion," *Cognition and Emotion*, vol. 8, no. 3, pp. 201–229, 1994.

[29] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication*, vol. 40, no. 1-2, pp. 5–32, 2003.

[30] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic, "The SEMAINE corpus of emotionally coloured character interactions," in *Proc. ICME*, 2010, pp. 1079–1084.

[31] K. Krippendorff, *Content analysis: An Introduction to Its Methodology*. Sage, 1980.

[32] S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann, "'Of all things the measure is man': Automatic classification of emotions and inter-labeler consistency," in *Proc. ICASSP*, 2005, pp. 317–320.

[33] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. R. Movellan, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," in *Proc. Advances in Neural Information Processing Systems*, 2009, pp. 2035–2043.

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

[35] P. Welinder, S. Branson, S. Belongie, and P. Perona, "The multidimensional wisdom of crowds," in *Neural Information Processing Systems Conference (NIPS)*, 2010, pp. 2424–2432.

[36] G. Chittaranjan, O. Aran, and D. Gatica-Perez, "Exploiting observers' judgements for nonverbal group interaction analysis," in *Proc. IEEE Conf. FG*, 2011, pp. 734–739.

[37] H. Meng, A. Kleinsmith, and N. Bianchi-Berthouze, "Multi-score learning for affect recognition: the case of body postures," in *Proc. Int'l Conf. ACII*, vol. 1, 2011, pp. 225–234.

[38] S. Scherer, J. Kane, C. Gobl, and F. Schwenker, "Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification," *Computer Speech & Language*, vol. 27, no. 1, pp. 263–287, 2013.

[39] J. Howe, *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. New York: Crown Publishing, 2008.

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

[41] M. Soleymani and M. Larson, "Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus," in *Proc. the ACM SIGIR 2010 workshop on crowdsourcing for search evaluation*, 2010, pp. 4–8.

[42] J.-I. Biel and D. Gatica-Perez, "The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs," *IEEE Trans. Multimedia*, vol. 15, no. 1, pp. 41–55, 2013.

[43] 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.

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

[45] S. G. Barsade, "The ripple effect: Emotional contagion and its influence on group behavior," *Administrative Science Quarterly*, vol. 47, no. 4, pp. 644–675, 2002.

[46] S. Kumano, K. Otsuka, D. Mikami, M. Matsuda, and J. Yamato, "Understanding communicative emotions from collective external observations," in *Proc. CHI '12 extended abstracts on Human factors in computing systems*, 2012, pp. 2201–2206.

[47] S. Kumano, K. Otsuka, D. Mikami, and J. Yamato, "Analyzing empathetic interactions based on the probabilistic modeling of the co-occurrence patterns of facial expressions in group meetings," in *Proc. IEEE Int'l Conf. FG*, 2011, pp. 43–50.

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

[49] K. Otsuka, S. Araki, K. Ishizuka, M. Fujimoto, M. Heinrich, and J. Yamato, "A realtime multimodal system for analyzing group meetings by combining face pose tracking and speaker diarization," in *Proc. ICMI*, 2008, pp. 257–264.

[50] W. J. Potter and D. Levine-Donnerstein, "Rethinking validity and reliability in content analysis," *J. Applied Communication Research*, vol. 27, no. 3, pp. 258–284, 1999.

[51] 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, 1999.

[52] M. Pantic and M. Bartlett, *Face Recognition*. Vienna, Austria: I-Tech. Educ. Publ., 2007, ch. Machine analysis of facial expressions, pp. 377–416.

[53] K. L. Gwet, *Handbook of Inter-Rater Reliability (3rd Edition)*. Gaithersburg, MD: Advanced Analytics, LLC, 2012.

[54] R. Rosenthal, *The new handbook of methods in nonverbal behavior research*. Oxford, UK: Oxford University Press, 2005, ch. Conducting judgment studies: Some methodological issues.

[55] A. J. Conger, "Integration and generalization of kappas for multiple raters," *Psychol. Bull.*, vol. 88, pp. 233–328, 1980.

[56] A. L. Bouhuys, G. Bloem, and T. Groothuis, "Induction of depressed and elated mood by music influences the perception of facial emotional expressions in healthy subjects," *J. Affective Disorders*, vol. 33, no. 4, pp. 215–226, 1996.

[57] S. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *Int'l J. Math. Mod. Meth. Appl. S.*, vol. 1, no. 4, pp. 300–307, 2007.

[58] S. Cha and S. N. Srihari, "On measuring the distance between histograms," *Pattern Recognition*, vol. 35, pp. 1355–1370, 2002.

[59] T. Chartrand and J. Bargh, "The chameleon effect: the perception-behavior link and social interaction," *J. Pers. Soc. Psychol.*, vol. 76, no. 6, pp. 893–910, 1999.

[60] J. B. Bavelas, A. Black, C. R. Lemery, and J. Mullett, ""I show how you feel": Motor mimicry as a communicative act," *J. Pers. Soc. Psychol.*, vol. 50, pp. 322–329, 1986.

[61] S. Baron-Cohen, *Mind Reading: The Interactive Guide to Emotions*. London: Jessica Kingsley Publishers, 2004.

[62] S. Gorga and K. Otsuka, "Conversation scene analysis based on dynamic Bayesian network and image-based gaze detection," in *Proc. ICMI-MLMI*, 2010.

[63] S. Kumano, K. Otsuka, J. Yamato, E. Maeda, and Y. Sato, "Pose-invariant facial expression recognition using variable-intensity templates," *IJCV*, vol. 83, pp. 178–194, 2009.

[64] J. M. Bernardo and A. F. M. Smith., *Bayesian Theory*. John Wiley & Sons, Ltd., 1994.

[65] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. PAMI*, vol. 6, no. 1, pp. 721–741, 1984.

[66] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.

[67] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*. New York: Chapman and Hall, 1993.

[68] R. Rosenthal, *Meta-Analytic Procedures for Social Research*. Beverly Hills, CA: SAGE Publications, 1991.

[69] J. Cohen, "A power primer," *Psychological Bulletin*, vol. 112, no. 1, pp. 155–159, 1992.

[70] G. J. McKeown and I. Sneddon, "Modeling continuous self-report measures of perceived emotion using generalized additive mixed models," *Psychol. Methods*, vol. 19, no. 1, pp. 155–174, 2014.

Shiro Kumano Shiro Kumano received a PhD degree in Information Science and Technology from the University of Tokyo in 2009. He is currently a research scientist at NTT Communication Science Laboratories. His research interests include computer vision, and affective computing. He received the ACCV 2007 Honorable Mention Award. He has served as an organizing committee member of IAPR International Conference on Machine Vision Applications. He is a member of the IEEE, IEICE, and IPSJ.

Kazuhiro Otsuka Kazuhiro Otsuka received his B.E. and M.E. degrees in electrical and computer engineering from Yokohama National University in 1993 and 1995, respectively. He joined the NTT Human Interface Laboratories, Nippon Telegraph and Telephone Corporation in 1995. He received his Ph.D. in information science from Nagoya University in 2007. He was a distinguished invited researcher at Idiap Research Institute in 2010. He is now a senior research scientist in the NTT Communication Science Laboratories and is entitled as a distinguished researcher in NTT. His current research interests include communication science, multimodal interactions, and computer vision. He was awarded the Best Paper Award of IPSJ National Convention in 1998, the IAPR Int. Conf. on Image Analysis and Processing Best Paper Award in 1999, the ACM Int. Conf. on Multimodal Interfaces 2007 Outstanding Paper Award, the Meeting on Image Recognition and Understanding (MIRU) 2009 Excellent Paper Award, the IEICE Best Paper Award 2010, the IEICE KIYASU-Zen'iti Award 2010, and the MIRU2011 Interactive Session Award. He is a member of the IEEE, the IEICE and the IPSJ.

Dan Mikami Dan Mikami received his BEng and MEng degree from Keio University, Kanagawa, Japan, in 2000 and 2002, respectively. He has been working for Nippon Telegraph and Telephone Corporation, NTT from 2002. He received his Ph.D. in engineering from Tsukuba University in 2012. His current research activities are mainly focused on robust visual object tracking. He received the Meeting on Image Recognition and Understanding (MIRU) 2009 Excellent Paper Award, the IEICE Best Paper Award 2010, the IEICE KIYASU-Zen'iti Award 2010, and the MIRU2011 Interactive Session Award.

Masafumi Matsuda Masafumi Matsuda received the B.A., M.A., and Ph.D. degrees from Hokkaido University, Hokkaido, Japan, in 1998, 2000, 2004, respectively. He joined NTT Communication Science Laboratories, Kyoto, Japan in 2003. He has been engaged in research on human interactions and human cognition from a social psychological perspective. Dr. Matsuda received the Best Paper Award from Japanese Psychological Association in 2002 and Human Communication Award from IEICE in 2010 and 2006. He is a member of IEICE.

Junji Yamato Junji Yamato is the Executive Manager of Media Information Laboratory, NTT Communication Science Laboratories. He received the B.E., M.E., and Ph.D. degrees from the University of Tokyo in 1988, 1990, and 2000, respectively, and the S.M. degree in electrical engineering and computer science from Massachusetts Institute of Technology in 1998. His areas of expertise are computer vision, pattern recognition, human-robot interaction, and multiparty conversation analysis.

He is a visiting professor of Hokkaido University and Tokyo DENKI University and a lecturer of Waseda University. He is a senior member of IEEE, and the Association for Computing Machinery.