Understanding Communicative Emotions from Collective External Observations

Shiro Kumano
NTT Communication Science Laboratories
3-1 Morinosato-Wakamiya
Atsugi, Kanagawa, Japan
kumano.shiro@lab.ntt.co.jp

Masafumi Matsuda
NTT Communication Science Laboratories
3-1 Morinosato-Wakamiya
Atsugi, Kanagawa, Japan
matsuda.masafumi@lab.ntt.co.jp

Kazuhiro Otsuka
NTT Communication Science Laboratories
3-1 Morinosato-Wakamiya
Atsugi, Kanagawa, Japan
otsuka.kazuhiro@lab.ntt.co.jp

Junji Yamato
NTT Communication Science Laboratories
3-1 Morinosato-Wakamiya
Atsugi, Kanagawa, Japan
yamato.junji@lab.ntt.co.jp

Dan Mikami
NTT Communication Science Laboratories
3-1 Morinosato-Wakamiya
Atsugi, Kanagawa, Japan
mikami.dan@lab.ntt.co.jp

Abstract
This paper presents a research framework for understanding communicative emotions aroused between people while interacting in conversation. Our advance is to consider how these emotions are perceived by other people, rather than what the target's internal state really is. Because such perception is subjective, we introduce the concept of using a collection of subjective external observations to objectively identify a fact. By treating the difference in perceived state as a probability distribution, we propose a computational model that describes the relationship between the perceived emotion and participants' key nonverbal behaviors, i.e. gaze and facial expressions. We also propose an evaluation method to assess the model by comparing the distributions estimated by using it with those of observers'. This paper describes initial experiments and discusses its potential.

Keywords
Empathy; antipathy; perceived emotion; distribution; observer; facial expression; gaze; Bayesian network

ACM Classification Keywords
H.1.2 [**Models and Principles**]: User/Machine Systems

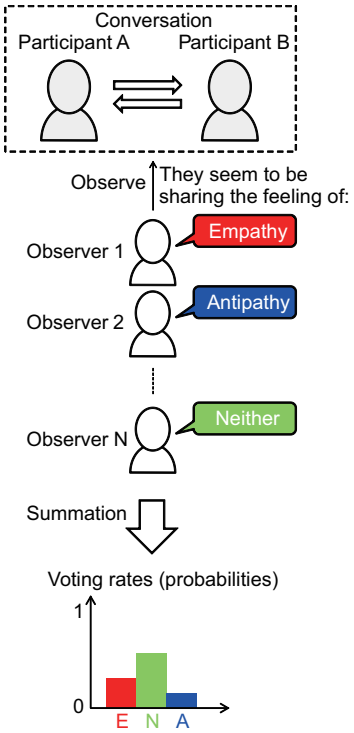

Figure 1: Empathy perception in this research, and the probabilistic representation of its distribution among observers.

Introduction

Understanding others' emotions is one of the key requirements to communicating well. Unfortunately, it's not so easy for most people, especially in remote meetings due to their poor support of nonverbal information used to express emotions, which is readily conveyed in face-to-face settings. Accordingly, a communication system that can understand user's emotions is needed for better communication.

Automatic understanding of emotion has been acknowledged as an emerging research area as introduced in [13]. However, emotion is an internal state of humans, and so researchers often face a barrier in that the ground truth is not explicitly available; this impedes the quantitative evaluation of computational models. Few effective research frameworks are known, especially for communicative emotions that are aroused between people in conversations due to the complexity of the situation; people dynamically interact and affect each other.

This paper proposes a research framework that resolves the problem. The framework consists of five components. (1) Target: We focus on empathy and antipathy, as common communicative emotions. They are expected to be the key elements for understanding multi-party conversations, because their contagion among participants affect consensus building, e.g. group pressure effect.

(2) Viewpoints: In communication, where people exchange emotions via verbal/nonverbal behaviors, an important viewpoint is how the emotions are perceived by other people, rather than what the internal state really is[1]; its diversity and uncertainty are essential attributes, or even low inter-coder agreement doesn't matter. Accordingly, we adopt the approach of using a collection of subjective external observations to objectively identify a fact.

(3) Problem setting: From these two viewpoints, we set the problem of estimating how likely external observers are to ascribe the same states of empathy or antipathy to a target pair given a set of observations, see Fig. 1. Hereinafter, we jointly call such perceptions of empathy and antipathy by observers as *empathy perceptions*.

(4) Model: We treat the difference in the perception between observers as a probability distribution, i.e. voting rates. By modeling the relationship between the distribution and participants' nonverbal behaviors with a Dynamic Bayesian Network[2], we estimate the distribution of empathy perception as a posterior distribution based on Bayesian inference; the average tendency of the observers' perceptions is expressed as a prior distribution.

(5) Evaluation: This research takes the standpoint that the model is successful if it well recreates the set of perceptions made by an adequate observer group. So, we propose a method for quantitative model verification that compares the estimated probability distribution to the distribution obtained by external observers.

The remainder of this paper first introduces related works to position this study. Next, the proposed framework is explained with a discussion of an experiment. Finally, a summary and potential for future growth are given.

[1] This viewpoint is similar to the *effect-type* description of emotion in [5]. However, [5] does not address diversity and uncertainty.

[2] The computational model was already proposed in [7]. This paper proposes a comprehensive research framework, including the approaches and viewpoints about emotions, a way to collect external observations, and the evaluation method, as well as providing an extended experiment, and so rectifies the omissions of [7].

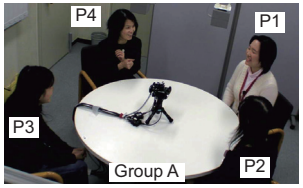

Figure 2: Example conversation scene captured by a still camera.

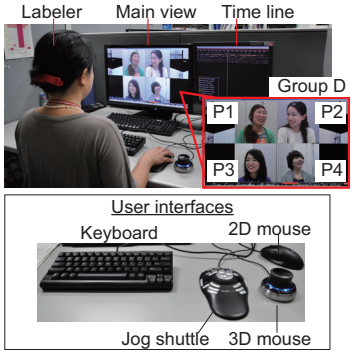

Figure 3: Labeling environment

Table 1: Frequencies of perception distribution types

Type	Example	[%]
Antipathy inferior		37
Neither dominant		27
Empathy dominant		26
Empathy inferior		0.54
Neither inferior		0.30
Flat		0.26
Antipathy dominant		0.08
Others		9.0

Probabilities:
Empathy Neither Antipathy

Related Works

There are two major approaches for building a computational model of emotion. One uses a dataset, where the ground truth is obvious, e.g. the use of acted behaviors [1] or self-reports [3]. The problems are that such behaviors often differ from the spontaneous natural ones, and it's difficult to obtain participants' self-reports in real time without altering the conversation. Many such works can be found in the excellent reviews, e.g. [13].

The other approach, our focus here, uses the judgment of multiple observers. Most works determine a single *representative value* regardless of the difference in perception between observers; the most popular technique is majority voting or averaging, e.g. as used in [9]. Some utilize interpersonal variation for evaluating the classifier [11], while others try to achieve a better approximation that improves the performance of the classifier [4][3]. After we proposed a model of empathy in [7], Meng et al. [8] proposed a multi-score learning problem, where a single sample contains multiple scores from multiple observers. Though their problem setting is similar to ours, their target (emotional posture *of a target person* during game play) and models are quite different from ours. In addition, they avoid clearly addressing why the diversity of observers' perception is important.

Among human-machine interaction studies, Huang et al. [6] used the idea of "wisdom of crowds" to propose a scheme for determining the appropriate timing of an avatar's backchannel response to a user, by using the timings obtained from multiple parasocial listeners. It differs from our research mainly in that it targets avatar behavior.

[3] Their target is to determine the dominant person in a meeting.

Definition of Empathy Perceptions

The term "empathy" (in lower-case here) used in psychology, neurophysiology etc. generally denotes an emotional state *of a target person*, though the definitions are often different between researchers as reviewed in [2]. On the other hand, we define "Empathy" ("Antipathy") (in capital letters here) as a state where an external observer imagines[4] that a pair of participants in a conversation is sharing the feeling of "empathy" ("antipathy"). Our definition differs from the common ones in two points. 1) Our target is a state *between the pair*, because communicative emotions are expected to be shared between participants via their interaction. 2) Our focus is, rather than what emotions the participants are really feeling, how they are perceived by observers. We target external, i.e. non-participant, observers to obtain explicit perception without altering the conversation.

We refrain from procedural definitions like a decision tree that almost automatically distinguishes each type of empathy perception, to avoid distorting the observer's intuitive perception, by following [10]. Accordingly, in our empathy perception, uncertainty about the definition and perception of Empathy by external observers, and participant behavior ambiguity are combined. We handle these uncertainties as probabilities in the framework of Bayesian theory by considering that these perceptions are created in a stochastic process.

Distribution of Empathy Perceptions

This section explains how to obtain a distribution of empathy perceptions made by external observers.

[4] Our definition is based on Stotland's *imagine-other* perspective [12], i.e. the observer imagines how *the target participants* are feeling, rather than *imagine-self* perspective, i.e. the observer imagines how *he/she* would feel in the participants' place.

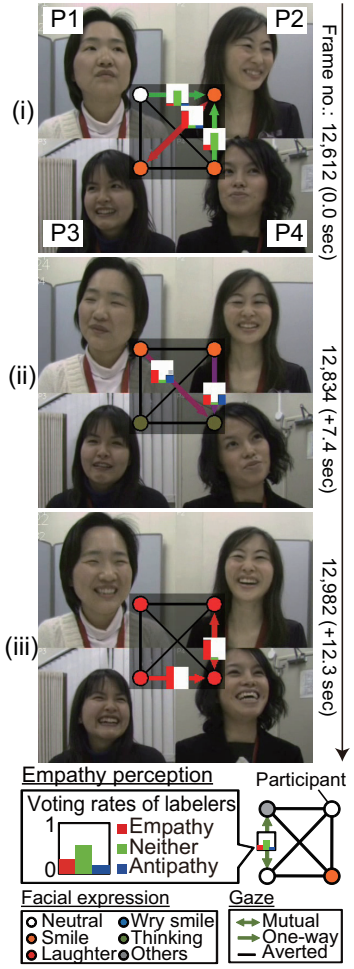

Figure 4: Example sequence of empathy perception created by labelers

Conversation Data

This paper targets four-person face-to-face conversations, as shown in Fig. 2. The participants were instructed to hold alternative-type discussions and to build consensus between the group within eight minutes on each discussion topic. The topics were "Is marriage and romantic love the same or different?" etc. The participants were 16 women (four four-person groups; G-A, G-B, G-C, and G-D) in their twenties or thirties. They had not met before the experiment. Focusing on the most lively exchanges in the participants' opinions, ten discussions, four from G-A and two from each of G-B to G-D, were picked up and analyzed. The average discussion length was 7.4 min (1.4 min S.D.). Each conversation was captured by an omnidirectional tabletop device at 30 fps, see [7].

External Observers

We employed nine external observers, hereinafter called *labelers*, who occupied the same age bracket as the participants. None of them participated in the conversations. Five of them labeled all conversations, while the remaining four processed only G-A conversations. Each labeler took about 10 minutes to process one minute of conversation for each pair.

Labeling Environment

For viewing and labeling videos, our original software, *NTT-CSL Conversation Scene Viewer*, was used, see Fig. 3; Videos could be played at normal speed or any other speed by turning a jog shuttle. The labeler could replay the video as many times as desired.

Instruction to Observers

The labelers were asked to watch the videos and to assign the label of "Empathy", "Antipathy" or "Neither" to each pair and time. We focus on emotions exchanged via visual nonverbal behaviors in this paper. So, all video sequences were labeled without access to the audio signals[5]. Also, by assuming that a participant can feel empathy or antipathy toward the other person only while looking at the person, the labeling scheme was slightly different between three gaze states existing between the pair. For mutual gaze, i.e. the pair are looking at each other, the labeler was asked to select Empathy (Antipathy) if the labeler felt they are sharing empathy (antipathy). For one-way gaze, i.e. only one of the pair is looking at the partner, labeler judged whether the gazer seemed to be feeling empathy with or antipathy to the gazee. The frames in mutually averted gaze, where neither is looking at the other, were removed as targets of labeling and analysis. These gaze states were annotated by one labeler in advance.

Labeling Results

Table 1 shows the frequencies of the eight distribution types of empathy perception. The most frequent type is Antipathy inferior, i.e. Empathy and Neither are conflicting. Other conflicting cases, though infrequent, also can be found such as Empathy inferior, Neither inferior, Flat, etc. These results reinforce the importance of treating the perceptions as distributions, instead of trying to select a single state via majority voting etc.

Case Study with Empathy Perception Labels

Figure 4 shows typical snapshots of empathy perception labels[6]. Hereinafter, the participants in these images are called P1 (upper-left), P2 (upper-right), P3 (lower-left), and P4 (lower-right), and their pairs are denoted as P1-P2 (the pair P1 and P2) etc. In (i) and (iii), the distributions for all pairs can be categorized into Empathy or Neither dominant. When positive facial expressions (FEs), i.e.

[5] We also obtained the labels made with using audio. However, they were not significantly different from those without audio.

[6] A part of video sequences are available from http://www.brl.ntt.co.jp/people/kumano/.

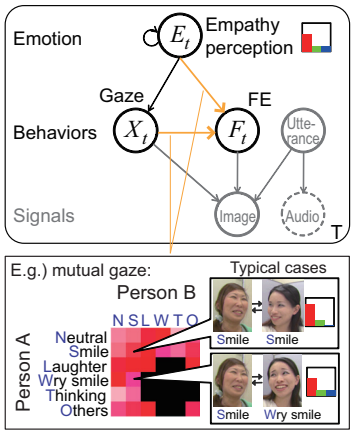

Figure 5: Graphical representation of our model [7]; the relationship between E, X, and F is modeled with the co-occurrence matrices of FEs.

Table 2: Estimation accuracy of empathy perception for each conversation group (distribution similarity S)

G-A	G-B	G-C	G-D
0.74	0.69	0.63	0.65

Table 3: Estimation accuracy of each type of empathy perception (majority-based matching rates)

Emp.	Nei.	Ant.
0.64	0.70	0.81

smiles and laughter, are strongly exhibited, labelers usually judged such interactions as Empathy. On the other hand, in (ii), none of the perception distributions have significant peaks, i.e. the perceptions of the labelers were quite different. These results suggest that the co-occurrence of FEs affects the shape and diversity of the distribution of empathy perception.

Computational Model

Our computational model describing the relationship between empathy perceptions and participants' behaviors observable for observers is based on the Dynamic Bayesian Network (DBN); the difference in perception between observers is regarded as a probability distribution. DBN can explicitly represent the structured stochastic relationships between elements, including static/dynamic dependencies and independencies. It also makes it easy to introduce other psychological findings or assumptions.

Key Participant Behaviors

We assume that, when people assign Empathy or Antipathy to a pair, their perceptions are strongly dependent on the gaze patterns and the co-occurrence of FEs between the pair. We derived this from behavioral coordination or motor mimicry in empathy; a person empathizing often adopts the expression of an observed other [2]. Among behaviors, we focus on facial expression (FE) and gaze, because FE conveys a large amount of emotional messages, while gaze is vital to inferring emotion from the FEs (monitoring), and triggers the reaction of the gazee; their combinations realize the rapid and directed transmission of emotional messages.

Dynamic Bayesian Network

We propose a simple model as the first step, see Figure 5. In Fig. 5, nodes represent random variables and edges represent dependencies between variables. The gaze between participants takes one of three states; {*mutual, one-way, and mutually averted*}. The FE state describes the co-occurrence of facial expressions between them; target FE categories for each participant are {*neutral, smile, laughter, wry smile, thinking, and others*}. Our model is characterized by the focus of the co-occurrence of facial expressions between participants via their gaze patterns. See [7] for details of the model. Moreover, by assuming that observers' perception of participants' gaze and FEs from videos is the same between observers, the labels of these behaviors were given by one labeler.

Estimation of Distribution of Empathy Perception

In the framework of Bayesian inference, we estimate joint posterior probability distributions of the sequence of empathy perception in the length of T frames, $E_{1:T}$, and model parameters, φ, given by the sequence of FE, $F_{1:T}$, and the sequence of gaze $X_{1:T}$, or $p(E_{1:T}, \varphi | F_{1:T}, X_{1:T})$. The model parameters explain the causal relationship between variables; they are time-invariant.

Evaluation Method

We quantitatively evaluate a model based on the similarity of the distributions produced by the model to the distributions of external observers.

Evaluation Measure

As the similarity measure between two distributions, we introduce the metric of the overlap area, S. For probability distributions $\boldsymbol{p}$ and $\boldsymbol{q}$, their overlap area is calculated as $\mathrm{S}(\boldsymbol{p}, \boldsymbol{q}) = \sum_i \min(p_i, q_i)$, where p_i and q_i denote the i-th component of $\boldsymbol{p}$ and $\boldsymbol{q}$, respectively. S becomes one at maximum, i.e. two distributions are exactly the same, and zero at minimum, i.e. no overlap.

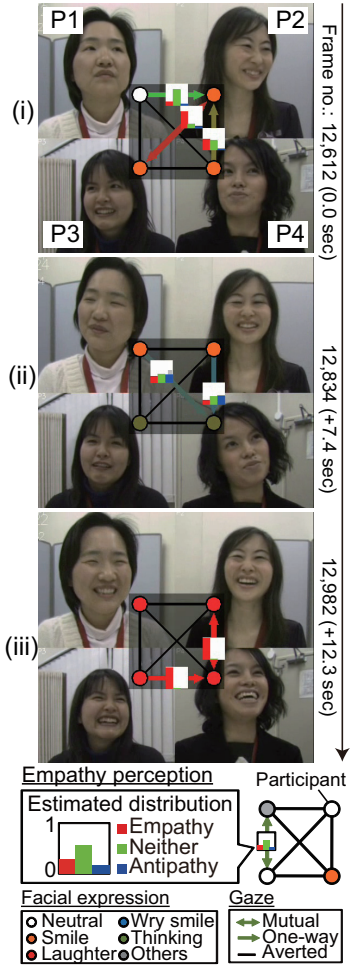

Figure 6: Example sequence of empathy perception estimated by using our model (corresponding to Fig. 4)

Evaluation Experiment and Discussion

Table 2 shows the calculated the similarity values; the mean S is 0.689 with S.D. of 0.05. This result suggests the validity of our hypothesis, i.e. FEs and gaze are key nonverbal behaviors as regards empathy perception. For further understanding, Table 3 shows majority-based agreement rates, where the estimation is considered to be correct, if the majority state(s) voted by labelers contain a state that maximizes the estimated posterior probability. The accuracies are high for all perception types.

Figure 6 shows estimation results for the scenes in Fig. 4. The estimated distributions are quite similar to those created by the labelers, when clearly exhibited positive FEs co-occur between a pair, shown as pair P2-P3 in (i) and both pairs in (iii). On the other hand, the distributions in (ii) are relatively difficult for the model to match, although the divergence of perceptions among observers can be recognized. The key to more accurate estimation is probably to introduce the variation in perception of FEs between observers and other participant behaviors, especially head gestures, as shown by P1 and P4 in (ii).

Summary

This paper presented a research framework for understanding the communicative emotion aroused during conversation. By focusing on empathy and antipathy shared between a pair of people, we introduce the viewpoint of representing them by the perceptions of external observers, and then set the problem of creating a model to recreate the perceptions. An experiment on the proposed evaluation method demonstrated that it well represents the relationship of the perceptions created from participants' gaze and facial expressions. We believe that the proposed framework makes it possible to quantitatively evaluate various phenomena, the ground truth of which is ambiguous and/or difficult to obtain like emotion.

References

[1] T. Bänziger and K. Scherer. Using actor portrayals to systematically study multimodal emotion expression: The gemep corpus. In *Proc. ACII*, volume 4738/2007, pages 476–487, 2007.

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

[3] 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*, 42(4):335–359, 2008.

[4] G. Chittaranjan, O. Aran, and D. Gatica-Perez. Exploiting observers' judgements for nonverbal group interaction analysis. In *Proc. AFGR*, pages 734–739, 2001.

[5] R. Cowie. Describing the emotional states expressed in speech. In *ITRW on Speech and Emotion*, pages 11–18, 2000.

[6] L. Huang, L.-P. Morency, and J. Gratch. Parasocial consensus sampling: combining multiple perspectives to learn virtual human behavior. In *Proc. AAMAS*, volume 1, pages 1265–1272, 2010.

[7] 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. AFGR*, pages 43–50, 2011.

[8] H. Meng, A. Kleinsmith, and N. Bianchi-Berthouze. Multi-score learning for affect recognition: the case of body postures. In *Proc. ACII*, volume 1, pages 225–234, 2011.

[9] M. Nicolaou, H. Gunes, and M. Pantic. Output-associative RVM regression for dimensional and continuous emotion prediction. In *Proc. AFGR*, pages 16–23, 2011.

[10] W. J. Potter and D. Levine - Donnerstein. Rethinking validity and reliability in content analysis. *J. Applied Communication Research*, 27(3):258–284, 1999.

[11] 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*, pages 317–320, 2005.

[12] E. Stotland. Exploratory investigations of empathy. *Advances in experimental social psychology*, 4:271–314, 1969.

[13] 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*, 31(1):39–58, 2009.