IEICE TRANS. INF. & SYST., VOL.E97–D, NO.8 AUGUST 2014

PAPER Special Section on Cyberworlds

Analyzing Perceived Empathy Based on Reaction Time in Behavioral Mimicry

Shiro KUMANO[†a)], **Kazuhiro OTSUKA**[†], **Masafumi MATSUDA**[†], ***and*** **Junji YAMATO**[†], ***Members***

SUMMARY This study analyzes emotions established between people while interacting in face-to-face conversation. By focusing on empathy and antipathy, especially the process by which they are perceived by external observers, this paper aims to elucidate the tendency of their perception and from it develop a computational model that realizes the automatic inference of perceived empathy/antipathy. This paper makes two main contributions. First, an experiment demonstrates that an observer's perception of an interacting pair is affected by the time lags found in their actions and reactions in facial expressions and by whether their expressions are congruent or not. For example, a congruent but delayed reaction is unlikely to be perceived as empathy. Based on our findings, we propose a probabilistic model that relates the perceived empathy/antipathy of external observers to the actions and reactions of conversation participants. An experiment is conducted on ten conversations performed by 16 women in which the perceptions of nine external observers are gathered. The results demonstrate that timing cues are useful in improving the inference performance, especially for perceived antipathy.

key words: *empathy, emotional contagion, perception, mimicry, facial expression, response time, time lag, Bayes model*

1. Introduction

Face-to-face conversation is the primary way of sharing information, understanding others' emotion, and making decisions in social life. Unfortunately, it's not so easy for people to fully understand what the others are feeling in a conversation, or reach full agreement about a controversial topic. The quality and efficiency of communication can be enhanced by applying information technologies to conversation support systems, such as in real-time computer-mediated visual telecommunication, conversational agents/robots, and counseling of autistic communicators. To realize such applications, it's required to automatically understand not only human behavior but also the participants' emotions which temporally evolve in the course of the interaction and impacts the conversation. Accordingly, the main target of automatic meeting analysis is now shifting from behavior to emotion [1]–[5].

To understand emotion in conversation, it is important to shed light on the communication process by which emotion is expressed, perceived, and shared between people via their interactions. Most previous works on the automatic recognition of human emotion mainly focus on the basic emotions of people in isolation, i.e. not interacting with others, and try to infer what type of basic emotion a target person is really feeling. Recently, from a more communication-directed viewpoint, some researchers proposed to infer how emotions aroused between a pair of people in multi-party conversations are perceived by external observers. For example, the targets of [6] are empathy as emotional contagion, and antipathy or counter-empathy as emotional conflict.

When people imagine the emotional states of others in conversation, they are thought to utilize two kinds of cues: dynamic cues and static cues. Dynamic cues are timing and/or the order of behaviors between the pair. The dynamic cues are intrinsically critical in communication sessions, where participants are rhythmically and interchangeably displaying their behaviors as indicators of their emotions. Numerous literatures have explored such cues in dyadic interactions, e.g. [7]–[9]. Static cues that are obtained from a snippet or still image of the conversation can also explain a part of the perceived emotions. Such static cues include how strongly expressions are displayed and/or what kinds of behaviors co-occur between a pair. As an example of the use of the static cues, the previous studies [6], [10] focus only on the instantaneous co-occurrence of facial expressions and gaze between a target pair. However, to the best of our knowledge, there is no computational model that describes the differences between external observers' subjective perceptions of empathy/antipathy between a pair, when the dynamics of the pair's interaction is changed.

This paper aims to investigate the relationship between the dynamics of participants' behaviors and subjective perception of external observers about empathy/antipathy between the pair. The present paper has two key contributions: First, it hypothesizes and demonstrates that the observer's perception of an interacting pair is formed by congruence and time lag between sender's action and receiver's reaction. Second, from the findings, it proposes a computational model for estimating perceived empathy/antipathy*. The following paragraphs describe the details.

We derive the hypothesis by assuming that the tendencies of observer's perception about empathy/antipathy are

Manuscript received November 19, 2013.
Manuscript revised March 4, 2014.

[†]The authors are with NTT Communication Science Laboratories, NTT Corporation, Atsugi-shi, 243–0198 Japan.
a) E-mail: kumano.shiro@lab.ntt.co.jp

DOI: 10.1587/transinf.E97.D.2008

*We have already presented the preliminary hypothesis testing and the model in [11]. This paper substantially enhances the reliability of the hypothesis testing and inference performance by adding two coders for the annotation of interlocutor's facial expressions. Furthermore, this paper determines more exact perception boundary of the time lag of the receiver's facial reaction by testing a variety of boundaries.

Table 1 Key hypothesis of the present study: expected dominant observer perception.

Behavioral congruence between a pair	Reaction time	
	Rapid (0–400 ms)	Delayed (500–1,000 ms)
Congruent	Empathy (H1)	Antipathy (H2)
Incongruent	Antipathy (H3)	Empathy (H4)

similar to those of human physical reactions to emotion-eliciting stimuli. Previous psychological studies, e.g. [12]–[15], have studied and documented the characteristics of imitated facial reactions when exposed to another human's facial expression. In short, as described in detail in Sect. 2, both the reaction time and the interpersonal relationship between the pair change the congruence/incongruence of the interaction. Table 1 summarizes our hypothesized characteristics of perceived empathy and antipathy. The present study focuses only on the timing of facial expressions between a pair as key behaviors, though they might not be the most crucial factors, because it's currently intractable to fully consider all possible combinations of single- and cross-channel behavioral congruence between a pair. However, a test with our face-to-face multi-party conversation dataset yields promising results in support of the hypotheses.

The model presented herein consists of two sub-models. One describes the relationship between perceived empathy/antipathy and the time lag between action and reaction; we call it the *timing model*. The other, a static model, is based on the instantaneous co-occurrence of their behaviors. In addition to facial expression and gaze, the focus of in [6], [10], the present study also considers head gestures and utterances. As expected from the present psychological study on perceived empathy/antipathy, as explained in Sect. 4, our experiment demonstrates that the timing model is helpful in improving the inference performance, especially for antipathy, and so is superior to the use of the static model only.

Following [6], this paper describes the perceptions of multiple observers as a distribution. Here, distribution means a probability distribution that represents voting rates expressing how many observers voted for or will vote for each category. It is considered in [6] that the set of perception is a kind of collective perception, and hence it is a practical description for objectively and deeply understanding conversation scenes. It also has an engineering merit. This expression can be put into the probabilistic inference framework, so a variety of analytical techniques for estimating model parameters can be applied. As to applications, a distribution showing such rich information of collective perception would be beneficial for the automatic recording of meeting minutes, or the automatic support for appraisers, facilitators and some researchers, such as social psychologists, sociologists, and psychotherapists.

The remainder of this paper first introduces related works to position this study and to derive the present hypothesis in Sect. 2. Next, our definition of perceived empathy/antipathy is explained in Sect. 3. A psychological study that assesses the impact of timing and congruence between facial action-reaction is detailed in Sect. 4. A probabilistic model for estimating perception from interlocutor behaviors is described and evaluated in Sect. 5. Finally, we summarize this study in Sect. 6.

2. Related Works

This section positions this study in a comparison of related works.

One target that has been well studied in psychology and neuropsychology is *the human as a reactor* to emotion-eliciting stimuli, including other's facial expressions. For example, when observing other's emotional face, people involuntarily and rapidly mimic the presented facial expression, even for negative emotional displays [12], e.g. to smile at a happy face or frown at an angry face. Many previous studies, e.g. [12]–[15], reported that the response time in facial electromyographic (EMG) reactions is around 300–400 ms, if the subject is prepared to mimic the presented facial expression. On the contrary, if the subject tries to show an opposite reaction, e.g. to frown at a happy face or smile at an angry face, the response delay increases to 500–1,000 ms [12], [13]. These mimicking patterns, whether congruent or not, depend strongly on context, e.g. relationship between the subject and the persons issuing the target facial expression [16]–[18]. For example, congruent reaction, i.e. a reaction the same as or similar to the target facial expression, is likely to be produced for a cooperative partner, while an incongruent reaction is expected from a competitive partner [16].

In pioneering works on perceived empathy, Ickes et al. [19] and Levenson et al. [20] define empathy as empathetic accuracy, which means the ability to perceive accurately how another person is feeling. They investigated the accuracy of the interaction partner [19] or external observers [20]. Although Levenson et al. [20] demonstrated physiological linkage between the target and the external observers, neither of them focused on behavioral congruence between a pair or its time lag, nor proposed any computational models of perceived empathy that could realize automatic inference.

Numerous computational models for inferring emotion can be found in the excellent reviews published to date, e.g. [3]. There are two major approaches for building such models, though they have often been confused in the engineering community [21]. One focuses on emotion that the target person is really feeling, and the other focuses on the impression of observers. The main difference between these approaches is whether the ground truth is obvious or not. It is usually essential, because many machine learning techniques require a ground truth.

When the focus is the actual emotions of the target person, it can be explicitly obtained by using self-reports [2] or acted behaviors [22]. However, when observer perception is used, the problem is how to determine the ground truth given the different subjective judgments expected. Most pre-

vious 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 [4].

On the other hand, some researchers recently tried to handle the inter-observer difference as a distribution, e.g. voting rates. For example, [6] posed the problem setting of estimating the distribution among target categories (empathy, antipathy or neither). This expression is advantageous not only to handle data that does not follow normal distribution, e.g. multi-modal non-Gaussian distributions, but also to keep its variation. They modeled the static tendency of how the co-occurrence of facial expressions in a pair of conversation participants impacted the empathy as perceived by the observers. However, this reference considers neither other behavioral channels, e.g. head gesture and utterance, nor dynamic cues.

Some researchers attempted to automatically detect behavioral mimicry from wearable motion sensors [23] or audio-visual features [24]. Furthermore, the dynamics of action and reaction have been utilized for generating the reactions of a conversational avatar to a user. For example, Morency et al. [25] proposed a probabilistic model for predicting the timing of listener's head gesture by analyzing the time lag of the listener's backchannel against speaker's utterance in human-human dyad conversations. However, none of them did not address how such interactions are perceived by observers.

3. Definition of Perceived Empathy/Antipathy

Following [6], we focus on the communicative aspect of empathy/antipathy. The definition of perceived empathy in [6] differs slightly from the traditional. Traditional definitions of empathy basically address the emotional phenomena *actually arising in a subject*, e.g. empathetic accuracy [19] described in Sect. 2. On the other hand, this study targets the pair-wise emotional state aroused *between a pair of people* while interacting with each other, and how it is *perceived by others* from a communicative viewpoint. The definition of empathy/antipathy in the present study is "an instantaneous state where a pair of interlocutors under a multi-party conversation is in the same/similar or conflicting emotional state." Among the eight distinct phenomena of empathy summarized in [26]†, our definition is most strongly associated with "emotional contagion" and "imagine-other perspective."

"Emotional contagion" represents the state where the emotion of a subject is the same as or similar to that of the target person [27]. We employ the definition of empathy as emotional contagion, and antipathy as emotional conflict. Emotional contagion well matches our concept that communicative emotions are shared between participants via their interaction. To understand this phenomenon, pair-wise treatment of communicative emotions is necessary. Moreover, among the eight phenomena, emotional contagion is expected to be the most common definition, and we believe that it is the most important aspect for understanding the flow, quality and performance of any conversation.

"Imagine-other perspective" [28] means the way of viewing others; the observer imagines how the target participants are feeling. On the other hand, "imagine-self perspective" [28] means how the observer imagines how he/she would feel in the participants' place. We employ the imagine-other perspective for a more objective description of perceived emotions than the imagine-self perspective, because the imagine-other perspective creates more other(target person)-oriented emotional responses [28]. Moreover, because the objects in our case are a pair of participants, the perspective of this study can be called "imagine-pair perspective".

Furthermore, "behavioral coordination" explains the tendency by which a person empathizing often adopts the behavior of the observed other [26]. The proposed model is based on behavioral coordination in empathy. Hereinafter, perceived empathy and perceived antipathy are jointly called just perceived empathy without discrimination, unless necessary.

The term empathy is difficult to rigorously define, but most people share a common understanding of empathy. Such a concept is called projective content [29]. So, to obtain the intuitive perception of observers, by following the guideline in [29], the instructions to the observers in the present study contained neither technical terms nor procedural definitions like a long list of detailed rules; the usage of which would almost automatically distinguish the type of perceived empathy from participant behaviors.

In this case, the observers could be expected to arrive at their own detailed definition of empathy. Accordingly, our empathy perception covers both the variation in the definition and the real perception. We consider that a set of these labels can be an objective description, because it is a collection of subjective data with regard to the intuitive perception of empathy. We believe that this is a practical description for users in understanding complex conversation situations. It is difficult to separate the variation of perception and definition at this moment. However, the aim of this study is not to clarify all characteristics of empathy but to find and utilize effective properties for automatic meeting analysis.

4. Timing Analysis: A Psychological Study

This section analyzes the relationship between perceived empathy of observers and facial behaviors of conversation participants. The main focus is to elucidate how significantly the time lag and congruence between facial action and reaction in a pair affects the observer's perception of the interaction. As a result, most of our hypotheses are basically supported.

†The eight distinct phenomena of empathy summarized in [26] are: "cognitive empathy", "behavioral coordination" ("motor mimicry"), "emotional contagion", "aesthetic empathy", "imagine-other perspective", "imagine-self perspective", "personal distress", and "empathic concern".

4.1 Hypotheses

From existing findings on the characteristics of human reaction to emotion-eliciting stimuli described in Sect. 2, we explore the possibility that time lags and congruence can be used to infer observers' perceived empathy. As an important nonverbal behavior for observers to judge emotions of conversation participants, facial expressions, i.e. facial action and reaction, are focused here; Facial expression is the primary channel for the transmission of emotion [30].

Our basic hypothesis is that the external observer consciously or unconsciously utilizes the characteristics of human as a reactor when perceiving empathy/antipathy for target interaction. First, the time lag is probably critical in judging whether a receiver's reaction is spontaneous or intentional. Second, given that the interpersonal relationship between a pair determines the congruence of the interaction [16]–[18], we consider the empathetic/antipathetic relationship. This is reasonable because one aspect of empathy is emotional contagion, i.e. the interpersonal relationship of emotions between a pair is the same or similar [26], [27]; another aspect is behavioral congruence or mimicry [26].

The hypothesis is decomposed into the following four sub hypotheses in terms of the congruence and delay between action and reaction: (H1) Observers are likely to perceive empathy when the facial expressions of the target pair are *congruent* with lag of *0–400* ms. (H2) Observers are *unlikely* to perceive empathy when the facial expressions of the target pair are *congruent* with lag of *500–1,000* ms. (H3) Observers are likely to perceive antipathy when the facial expressions of the target pair are *incongruent* with lag of *0–400* ms. (H4) Observers are *unlikely* to perceive antipathy when the facial expressions of the target pair are *incongruent* with lag of *500–1,000* ms. Table 1 summarizes these hypotheses.

Now consider a typical example wherein a speaker is smiling at a listener in an attempt to gain his/her agreement, and an observer is looking at the pair. If the listener rapidly (lag of 0–400 ms) returns a smile to the speaker, the observer would perceive that their interaction is spontaneous and their emotional states are empathetic, i.e. the same or at least similar (H1). If an incongruent response is rapidly returned, e.g. negative facial expression to smile, the observer would perceive spontaneously displayed antipathy (H3). If the listener displayed a smile but after some delay (lag of 500–1,000 ms), the observer would perceive it as intentional as opposed to spontaneous, and therefore contains an element of deception and it is this that leads to the antipathy (H2).

4.2 Subjects: External Observers

Nine observers were employed. They were Japanese females in their twenties or thirties. They had met neither each other nor the conversation participants before the experiment.

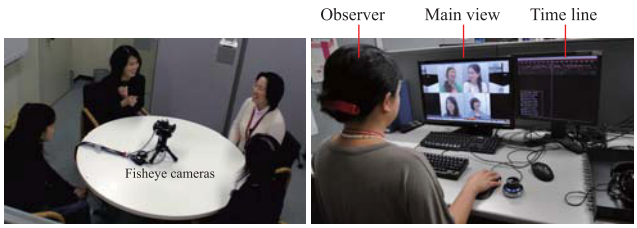

Fig. 1 Snapshots of conversation (left) and labeling (right) scenes.

4.3 Stimuli: Conversation Data

This paper targets four-person face-to-face conversations, as shown in the left part of Fig. 1. The participants were instructed to hold alternative-type discussions and to build consensus as a group, i.e. agree on a single answer, on each discussion topic within eight minutes. The participants were 16 Japanese women (four four-person groups: G_A, G_B, G_C, and G_D) in their twenties or thirties. They were chosen to occupy the same gender and age bracket to raise the probability of empathy [27]. They were asked to first hold a short chat with self-introduction, then hold about seven discussions with some intervals between them. All conversations were captured at 30 fps by IEEE1394 color cameras.

This study collected interlocutors who had not met before the experiment for analytical simplicity. Unacquainted individuals would be more likely to behave by following social rules [31] rather than interpersonal relationship, e.g. dominance or hierarchy. For example, as introduced in Sect. 2, in-group and out-group people tend to show different facial mimicry patterns [18]. The social rules, which are basically shared among a large community, would make it simpler for observers to understand the interaction than local rules that are shared among a specific small groups like families and friends.

Most discussion topics were assigned to the groups on the basis of the participants' opinions found in prequestionnaires so as to cause frequent concordance and disagreement. The discussions were held on a single day for each group. This paper picks up and analyzes ten discussions that are expected to include lively exchanges and a variety of empathy scenes, including empathetic, antipathetic, and ambiguous interactions. We dropped the rest of discussions due to the high cost of annotation, as described later in Sect. 4.4. The selected discussion topics were "Is marriage and romantic love the same or different?", "Who are more beneficial, men or women?", "Is marriage is necessary for life?", "Are blood types and personality highly related?", "Do there exist beneficial or detrimental blood types?", "Should preferential treatment for full-time housewives be introduced?", "Should smoking in public space be fully prohibited by law?", and "Should euthanasia be legally recognized in Japan?." Only the first topic was selected three times, because it yielded the most lively discussions. Four of the ten conversations were selected from G_A and two from each of G_B to G_D. The average discussion length was

7.4 min (1.4 min S.D.).

One female annotator in advance annotated facial expression, gaze (person receiving visual focus of attention), head gesture, and utterance of each participant in these conversations frame-by-frame. To enhance the reliability of the annotation of facial expression, i.e. the nonverbal behavior mainly focused in this paper, two other female coders additionally annotated interlocutor's facial expressions.

The label sets for facial expression are: positive class) smile, laughter, chuckle; neutral class) neutral, thinking, surprised, embarrassed, other neutral expressions; negative class) wry smile, disgust, bored, provoking, puzzled, sad, angry, afraid/fear, disbelieving, and other negative expressions. These 18 categories were prepared with reference to Facial Action Coding System (FACS) [32], and Mind Reading guideline [33]. The label sets for head gesture and utterance are {no gesture, nod (3 levels), shake (3 lv.), tilt (3 lv.), their combination}, and {speaking, silence}, respectively. Moreover, the coders were also allowed to select hard-to-judge label, if necessary.

Top five frequently used facial expression categories by the three coders, except for hard-to-judge (1.8%), are neutral (52.0%), smile (36.3%), thinking (3.5%), wry smile (2.1%), and laughter (1.6%). Each of the remaining 13 categories were selected less than 1%. Consequently, the frequencies of valence classes are 38.0% (positive), 56.5% (neutral), and 3.7% (negative). The resulting Conger's Kappa coefficients (inter-coder agreement) of facial expression annotation are $\kappa = .459$ for the 18 categories, and $\kappa = .533$ for positive/neutral/negative classes. According to the benchmarks in [34], this annotation is judged as moderate.

4.4 Procedure: Labeling of Perceived Empathy

Videos were viewed and labeled using our original software [5], [6]. The right part of Fig. 1 shows an example labeling scene. Two monitors, 26-inch and 16-inch, were used, the larger one was for displaying a movie that showed all participants at quarter-size, while the smaller one was for displaying a timeline of a sequence of perceived empathy labels given by the observer so far. Videos could be played at normal speed or any other speed by turning a jog shuttle. The observer could replay the video as many times as desired. All labeling was done in isolation.

Five of the observers labeled all conversations, while the remaining four processed only G_A conversations. Each observer was asked to finish the labeling of one conversation within one day (7.5 h), and most observers succeeded in doing so. Observers labeled all video sequences without recourse to the audio signals to focus on emotions exchanged by visual nonverbal behaviors. The labeling was region-by-region. That is, the frames at which the observer's perception changed were extracted, and then the sequence of frames between two labels was assigned the label of the head frame of the sequence. So, the temporal resolution of labeling was the same as the video rate, i.e. 30 fps.

The observers were asked to watch the conversation videos and to assign one of the following bipolar labels, the one closest to their perception, to each pair and at each time in each video sequence: "Strong Empathy" (+2), "Weak Empathy" (+1), "Neither Empathy nor Antipathy" (0), "Weak Antipathy" (−1), and "Strong Antipathy" (−2). The instruction was to "judge whether a pair of participants in a multi-party conversation is in the same/similar (empathetic) or conflicting (antipathetic) emotional state at that moment." Moreover, the observers judged empathy after well grasping the spatial arrangement of the interlocutors. This was done by gaze annotation for one session as pre-training with a birds-eye-view and the bust-shot modes, as shown in [5] and Fig. 1, respectively.

Because five-point distributions created by the five or nine labelers are too sparse for analysis on their distribution types, the present study ignores label strength; +2&+1, 0, and −1&−2 are called "Empathy", "Neither", and "Antipathy", respectively. This study considers that a pair of participants are interacting only if at least one of them is looking at the other. Other frames, i.e. those of mutually averted gaze, were removed as targets of labeling and analysis. See [6] for more details.

To demonstrate the degree of perception variation, the inter-coder agreement of empathy perception was calculated. Conger's kappa is $\kappa = .291$. It is much smaller than that of facial expression annotation ($\kappa = .533$ for three classes), and is judged as poor according to [34]. Furthermore, following [6], the number of samples that have a remarkable single majority, i.e. samples assumed in most previous studies, is counted. Consequently, the frequency is as high as 47.1%. Although these are not the only measures that can be used to show the degree of data variation, they reinforce the importance of treating the perception as distributions, instead of trying to select a single state via majority voting etc.

4.5 Analysis

This analysis aims to investigate whether the perceived empathy of observers is really affected by both the time lag and congruence between action and reaction of a target pair holding a conversation. We determine how likely each perceived empathy is to be labeled for each time lag and interaction congruence.

The frequency of each type of perceived empathy, e, for a pair of people is counted only at the start of the reaction. This is because most changes in perceived empathy are produced by this timing, as demonstrated in Sect. 5. The frequency is separately calculated for each time lag, dt, and each congruence/incongruence state, c, between action and reaction. Time lag dt is grouped into $0 - \tau$ ms (rapid) and $\tau - 1{,}000$ ms (delayed) in this section, where τ means an expected perceptual boundary of the time lag of facial expressions. Coordination state c is a binary state describing whether the action and reaction were the same or not. In judging congruence, the 18 categories of facial expression were grouped into three emotional tones; positive, neutral,

Table 2 Original frequency of perceived empathy.

(a) Congruent facial interactions

Reaction time	Empathy	Neither	Antipathy
Rapid (0–433 ms)	570	222	18
Delayed (433–1,000 ms)	921	432	42

(b) Incongruent facial interactions

Reaction time	Empathy	Neither	Antipathy
Rapid (0–433 ms)	120	115	24
Delayed (433–1,000 ms)	440	262	27

and negative. The frequency of perceived empathy is expressed as $N_{dt,c}(e)$. Set $\{N_{\cdot,\cdot}(e)\}_{e=1}^{3}$ means a frequency distribution of perceived empathy e on one of four (2×2) conditions of congruence and time lag. Moreover, actions that were not looked at by the receivers upon emergence, i.e. action start, were dropped, because such actions are not expected to trigger reactions.

The frequency distributions of these four conditions are compared by using a chi-square test. In addition, for qualitative comparison, each of the original frequencies is normalized with regard to dt, i.e. $N'_{dt,c}(e) = N_{dt,c}(e) / \sum_{dt} N_{dt,c}(e)$. This normalization emphasizes the trend of each perceived empathy type by offsetting the imbalance in sample size between the types; empathy labels were about 100 times more frequent than antipathy labels in the data [6]. If the normalized frequency distributions are different both for each dt and c, it suggests that both time lag and congruence affect observers' perceived empathy. Note that the original, i.e. unnormalized, frequencies are used in the chi-square test.

4.6 Results

Table 2 shows the unnormalized frequencies of perceived empathy. Here, we first calculated the frequencies separately by using each of the facial expression labels of three coders, and then aggregated them. Figure 2 shows the distributions of normalized frequency[†], N', for each condition. Statistically significant differences were found among these conditions ($\chi^2 = 71.2$, $df = 6$, $p < 0.001$). As expected, these results basically support the validity of our hypotheses; observer perception, especially antipathy, strongly depends on both the time lag and congruence between action and reaction.

To determine the threshold of the time lag, or the perceptual boundary between rapid and delayed response, we calculated the p-values with every threshold in the range of 100 ms to 1,000 ms at 33 ms intervals, i.e. the video frame rate (30 fps). Consequently, thresholding at 433 ms showed the lowest p-value, i.e. the most significant difference between the distributions of perceived empathy. So, we conclude that the perceptual boundary of the time lag of facial expressions is around 400 ms. This boundary

[†]For easier understanding of the differences in perceived empathy e, the normalized frequencies are further normalized with regard to e in Fig. 2. This additional normalization is not essential. The important point is the difference in the frequency of perceived empathy e.

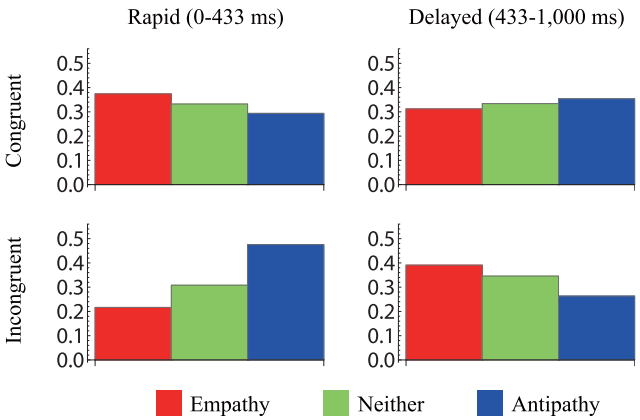

Fig. 2 Normalized frequencies of perceived empathy for each congruence (row) and reaction time (column), i.e. $N'_{dt,c}(e)$: Vertical axes denote normalized frequencies. A chi-square test shows the significant differences among these conditions ($\chi^2 = 71.2$, $df = 6$, $p < 0.001$).

Table 3 Summary of resulting dominant tendency in observer perception.

Facial congruence between a pair	Reaction time: Rapid (0–433 ms)	Reaction time: Delayed (433–1,000 ms)
Congruent	Empathy	Antipathy
Incongruent	Antipathy	Empathy

well matches the boundary of electromyographic (EMG) response time between spontaneous (0–400 ms) and deliberate (500–1,000 ms) facial expressions [12]–[15]. Some researchers have reported that spontaneous facial mimicry is often perceptible or visible as late as the latency of 800 ms, e.g. [35]. But, this early boundary would be possible, because, as pointed out in [36], time of onset of facial mimicry may be under the influence of several factors, including social context and attitude; actually, the response time varies across studies [36]. Further study is required, but face-to-face interaction is expected to make the boundary rapider than 800 ms, because observer would know from his/her own experiences that receiver in conversation can predict the timing of sender's facial action from the sender's other behavioral cues, such as prosody and gaze, hence would be ready to react to it.

Figure 2 demonstrates that when facial interactions are congruent and rapid (upper left distribution), makes their perception more likely to be Empathy. So, H1 can be accepted. In contrast, congruent and delayed interactions (upper right distribution) are hard to perceive as Empathy, so it's reasonable to accept H2. Incongruent and rapid facial expressions (lower left distribution) are the most likely to be recognized as Antipathy, so it is reasonable to accept H3. In incongruent and delayed responses (lower right distribution), Antipathy is infrequent on this condition. Thus, H4 is also acceptable. Table 3 summarizes these results. Antipathy shows clearer characteristics than Empathy and Neither. This suggests that people impose severe timing constraints on negative reactions, i.e. rapid facial incongruence and delayed facial congruence, unlike the other reactions.

Although the targets are different, these tendencies are consistent with those obtained in previous works, e.g. [20].

We also analyzed other single- and cross-channel behavioral congruence, such as head gesture to head gesture, and head gesture to facial expression. To assess cross-channel congruence, the categories of each behavior channel were also grouped into positive/neutral/negative. However, no noticeable difference was found, unlike facial congruence. For cross-channel congruence, some difference might be discovered if a more appropriate grouping rule of facial expression and gesture can be found.

4.7 Discussion

This study followed an engineering approach, where an application-oriented problem to be solved by machines, i.e. the automatic inference of perceived empathy between interlocutors in multi-party conversation, determines task design and definition; rather than a common psychological approach, where the definition of the target determines task design. Thus the experimental design is discussed here from several aspects, though the hypotheses of the present study were basically supported by the experiments on the multi-party conversation data.

4.7.1 Cognitive Load on Group Discussion

First, because the present study used group discussion, cognitive load of the interlocutor due to logical thinking will affect the timing of facial actions. So, this effect is discussed here. The point is, like the discussion on real emotion and its perception in Sect. 2, distinguishing how much cognitive load a target interlocutor is really experiencing and how observers perceive it.

As to interlocutors, though not in a conversational context, some researchers reported that the response time of body motion is affected by the degree of cognitive load [37]. This suggests that a similar effect will occur in facial mimicry. In our case, it is implausible to consider that the cognitive load is time-invariant in conversations; it is natural to expect that our data includes a variety of facial interactions made under different cognitive loads. However, the experiment showed the statistically significant differences.

The point of this argument is to focus on the observers rather than the interlocutors. There is a possibility that the observers were consciously or unconsciously estimating the cognitive load of target people while judging their empathy. However, it is expected that the task design preventing observers from accessing the audio makes it difficult to judge the degree of the cognitive load of interlocutors. Some facial expressions of interlocutors, e.g. thinking or puzzled, might indicate high cognitive load. However, such expressions are infrequent in our data ($< 4\%$). The statistical significance suggests that the effect of the cognitive load is not problematic at least for the main purpose of this study, i.e. investigating the relationship between the visual behaviors of interlocutors and the perception of observers. Exploring the relationship between empathy perception and cognitive load is an interesting research topic, but it is out of the focus of this paper.

4.7.2 Differentiation of Empathy from Agreement

Second, one might worry that it is hard to differentiate empathy from agreement in the present task design. They are different in their definitions; empathy in this study is the matching of emotions, while agreement means the matching of opinion. Empathy and agreement in the definitions are not exclusive but compatible. Furthermore, because some nonverbal behaviors, like smiles, indicate both empathy and agreement in conversation [38], their perception would be highly correlated; they are difficult to completely separate. However, the instruction to observers, explained in Sect. 4.4, clearly identified the target as being emotion, although the present study did not give the definition of agreement to the observers.

Judging without recourse to audio is expected to be more appropriate to further alleviate their confusion than that with audio. The reason is as follows: Some studies suggest that prosodic features and backchannel responses are important in judging the agreement of others [38], [39]. In contrast, empathy perception is not expected to require audio information. For example, in targeting rapport, a longer-term interpersonal relationship than empathy but closely related in concept to empathy, [40] reported that the highest accuracy is obtained under the video-only condition than other conditions such as audio-only and video with audio. Accordingly, the recourse to audio would enhance the confusion between empathy and agreement, in the sense that it makes it easier to judge agreement but harder to judge empathy.

4.7.3 Group Size: Dyadic versus Multiparty Conversation

Next, as to the number of interlocutors, the present study focused not on dyadic interactions but on four-party (group) conversations. They can be expected to yield different interlocutor behavior in the following sense. Behaviors of each interlocutor are observable from other multiple interlocutors in multi-party conversations, while the behaviors can be observed only from the single other in dyadic conversations. This suggests that facial expressions are more other-conscious ones, or communicative expressions [41], and their congruence and timing would be different from those in dyadic interaction. Because the observers knew the number of interlocutors, they could compensate this effect while judging empathy. Moreover, four-party conversations include an equilibrium state where the group splits into two conflicting groups; this would be a crucial state for consensus building or the social pressure effect [42].

4.7.4 Display Method

Finally, the present study used the display method of ar-

ranging the bust shots of four persons as a practical way of keeping their facial expressions visible. The display method obscures some visual behaviors such as hand gestures. Accordingly, if researchers want to investigate what visual features observers focus on, a birds-eye-view showing all visual behaviors would be more appropriate. However, [40] suggests that hand gestures and body motions such as preening contribute little to the perception of rapport. Furthermore, such behaviors were not frequent in our data. A possible reason is that most interlocutors kept their hands on their lap, as shown in Fig. 1, except for some cases, e.g. the person in the upper left in the figure. It would be an interesting issue to investigate the effect of display method, including the size, resolution, and position of interlocutors in the display, and the type of device such as traditional monitor or head mount display.

5. Probabilistic Model

Based on the results in Sect. 4, this section proposes a probabilistic model for estimating the distribution of perceived empathy of observers, i.e. voting rates or a ratio of observers who perceive a pair's state as empathy or antipathy.

5.1 Overview

Perceived empathy labels contain a mixture of various types of ambiguities in decision making; the ambiguities about inter-observer difference in the change timing of perceived empathy label, in the definition of empathy and in its perception scheme, and the ambiguity about participant behaviors. These ambiguities are well handled as probabilities in Bayesian theory, where a variety of analytical techniques for estimating model parameters can be applied. Furthermore, though outside the scope of this paper, it also has extensibility to develop a unified model that also includes lower-level layers of automatically inferred interlocutor behaviors such as facial expression and gaze, and higher-level layers such as interpersonal relationship, e.g. hierarchy, without losing the ambiguities at each layer.

Following [6], we treat perception diversity as a probability density distribution that shows how many observers voted for each perception type, i.e. voting rates. More unfocused voting means that the interaction yields greater ambiguity in terms of perception. We consider diversity and ambiguity are essential attributes; this is because humans cannot determine the other's actual emotions, and instead have to guess them from behaviors. To achieve better support of conversations and encourage feelings of satisfaction, these ambiguities must be well handled. By way of contrast, most previous studies consider that low inter-coder agreement rates merely indicate unreliable data.

We propose a naïve Bayes model for estimating the conditional probability density distribution of perceived empathy at time t, e_t, given by the time series of behaviors of a target pair of people, $\boldsymbol{B}$, $P(e_t|\boldsymbol{B})$. The conditional probability is assumed to be independent for each pair of participants. The naïve Bayes model assumes the independence of the probabilistic relationship between the objective variable (observer's perceived empathy here) and each of the explanatory variables (participant behaviors here). Although the naïve Bayes model is simple, its good performance in a variety of areas has been reported [43]. The notable advantages of the naïve Bayes model for the present study are the following two: likelihood functions can be easily added to or deleted from the model, and because joint probabilities among the explanatory variables are not considered, it's easier to avoid overfitting, which often arises if few training samples exist.

In our naïve Bayes model, the conditional probability distribution $P(e_t|\boldsymbol{B})$ is decomposed as:

$$P(e_t|\boldsymbol{B}) :\propto P(e_t)\prod_b P(dt_t^b|c_t^b, e_t)\prod_b P(b_t|e_t), \tag{1}$$

where $P(dt_t^b|c_t^b, e_t)$ denotes the *timing model*, a key component of the present model; it describes how likely an interaction is to be labeled e at time t given the time lag between action and reaction in behavior channel b around t. This model is prepared for each state of their congruence c, i.e. whether the categories of their behaviors are the same or not. The pattern of instantaneous behavioral co-occurrence is modeled with the *static model*, $P(b_t|e_t)$. It describes how likely an interaction is to be labeled with e at time t given the categories of behaviors instantaneously co-occurring in channel b. No reaction of one person to the action of his/her partner is represented by this static model. The following sections detail these two terms. $P(e_t)$ is the marginal probability of e; it describes how likely the target interactions are to being labeled with perceived empathy e without considering any explanatory variable.

5.2 Timing Model

The timing model of perceived empathy of behavior channel b is defined as:

$$P(dt_t^b|c_t^b, e_t) := P(\tilde{dt}_t^b|c_t^b, e_t)^{\pi_t}. \tag{2}$$

That is, it combines the likelihood of perceived empathy e in behavioral congruence/incongruence c with discretized time lag $\tilde{dt}$, $P(\tilde{dt}_t^b|c_t^b, e_t)$, and its weight with regard to the change timing of observer's perception at/or around the behavioral congruence, π_t. Only facial expression is considered in the timing model in the present study.

5.2.1 Time-Lag Function

Time-lag function $P(\tilde{dt}_t^b|c_t^b, e_t)$ describes how likely observers are to perceive empathy state e at time t given congruence state c in behavioral channel b with the time lag of dt. To simplify the mathematics, we use, instead of continuous dt, discrete $\tilde{dt}$ that is the bin number of a histogram. To avoid overfitting due to the limited sample size, the bin size is set to be 300 ms in this paper. Figure 2 is an example of this function with the bin size of 433 and 567

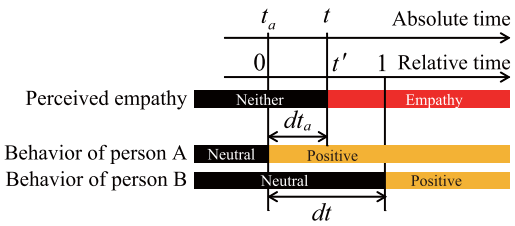

Fig. 3 Time lag between perceived empathy change and the beginning of action and reaction. Different colors mean different categories.

(= 1,000 − 433) ms.

5.2.2 Weight Function π

The weight function π describes at which timing the change in observer's perception, i.e. perceived empathy, will be triggered by the emergence of the action and/or reaction. Note here that the observers were allowed to change the label of perceived empathy at any timing by freely playing the videos. The weight function determines a timing at which the time lag function works in a stochastic form. In other words, π means the percentage of the observers who have changed their perception by this moment in the interaction. For example, $\pi = 0$ means that no observer changes his/her perception label at this moment. In this case, the timing model makes no contribution in Eq. (1). $\pi = 1$ means that every observer who changes his/her perception in the interaction changes the empathy label no later than this moment. So, the timing model makes maximum contribution in Eq. (1).

We use the following ramp function to model the weight function:

$$\pi_t = \begin{cases} 0 & (t' < -\alpha/(1-\alpha) \mid dt > L \mid t - t_a > W) \\ \alpha + (1-\alpha)\cdot t' & (-\alpha/(1-\alpha) \le t' \le 1) \\ 1 & (\text{otherwise}). \end{cases} \quad (3)$$

Variable t' is the relative time defined as $t' = dt_a/dt$, where dt_a is the time lag between the perception change and the beginning of the action of one person, as shown in Fig. 3. That is, $t' = 0$ and $t' = 1$ mean the beginning time of action and reaction, respectively. Variable L is a threshold for the time lag between action and reaction. In this paper, L is set to be 2 s with reference to Jonsdottir et al.'s finding [44] that delay between lexically key phrases and the facial expression of the listener lies approximately in the range of 500–2,500 ms. Condition $t - t_a > W$ means that the current time t is far from the emergence time of the latest action $t_a (< t)$. This models the tendency that once a perception is created, the perception continues for some duration but it is eventually lost in the absence of any new interaction behavior. Threshold W is empirically set to be 4 s. Red and blue bars in Fig. 4 show resulting effective ranges of the timing model, where $\pi > 0$.

The upper part of Fig. 5 shows actual frequencies of the change in perceived empathy around the action and reaction in our dataset. The labels were changed most frequently at the beginning of action ($t' = 0$), and not so much at the beginning of reaction ($t' = 1$). The lower part of Fig. 5 shows accumulated frequency (probability) in the range between the beginning of action and reaction. It shows the probability that if perceived empathy is changed in this range, the change is produced by relative time t'. These points well fit the purple line, plotted by the ramp function $p = \alpha + (1-\alpha)\cdot t'$, where α is 0.2.

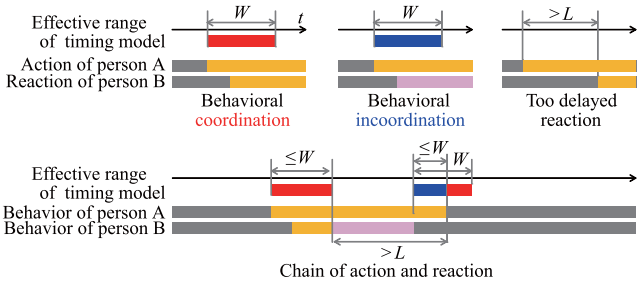

Fig. 4 Effective range of timing model: Upper left two figures show the cases of behavioral congruence and incongruence. Red and blue bars denote the effective ranges of the timing model for behavioral congruence and incongruence, respectively. Upper right figure show the case where the timing model does not work because the time lag between action and reaction is too large, i.e. $dt > L$. Lower figure shows an example where the pair of people are interchangeably displaying actions and reactions. Colors of action and reaction describe categories.

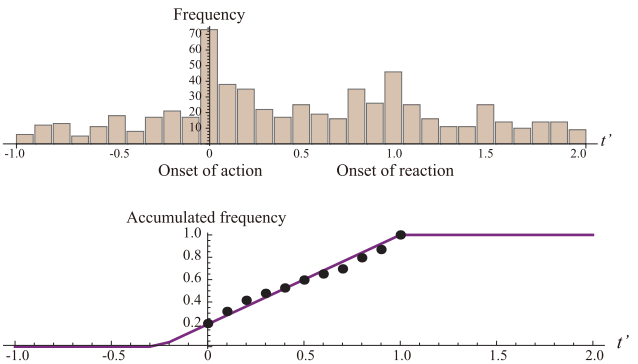

Fig. 5 (Upper) Timing between perceived empathy and action/reaction behaviors. Horizontal axis denotes relative time, t', in each interaction. Relative times $t' = 0$ and $t' = 1$ mean the beginning time of action and reaction, respectively. Vertical axis is the frequency that perceived empathy was changed. (Lower) Accumulated probability of the change in perceived empathy in the range of $t' = [0\ 1]$ (black dots), and the fitted weight function π (purple line).

5.3 Static Model

The static model $P(b_t|e_t)$ describes when a certain combination of behaviors in channel b occurs between a pair of people at time t, how likely the observers are to perceive empathy state e. As behavior channels b, this paper considers facial expression, gaze, gesture, and utterance, while previous works [6], [10] only target facial expression and gaze. This study assumes that they are independent each other given by

Table 4 Estimation accuracy (OA) in the effective ranges of the timing model.

Model	Frame avg.	Type avg.	Type 1 (Emp-dom)	Type 2 (Nei-dom)	Type 3 (Ant-dom)	Type 4 (Emp-inf)	Type 5 (Nei-inf)	Type 6 (Ant-inf)	Type 7 (Flat)
The proposed NB (F+G+H+U+Ft)	.759	.662	.767	.737	.484	.576	.601	.792	.680
The proposed NB (F+G+H+U)	.764	.656	.768	.747	.401	.586	.607	.798	.684
The proposed NB (F+G+H)	.766	.652	.758	.730	.376	.574	.603	.819	.702
Baseline (F+G) [6]	.743	.611	.727	.713	.214	.526	.605	.800	.690

NB: naïve Bayes model. F: facial expression, G: gaze, H: head gesture, U: utterance, and Ft: facial expression timing.
Emp: Empathy, Nei: Neither, and Ant: Antipathy. dom: dominant, and inf: inferior.

Table 5 Estimation accuracy (OA) for all frames.

Model	Frame avg.	Type avg.	Type 1 (Emp-dom)	Type 2 (Nei-dom)	Type 3 (Ant-dom)	Type 4 (Emp-inf)	Type 5 (Nei-inf)	Type 6 (Ant-inf)	Type 7 (Flat)
The proposed NB (F+G+H+U+Ft)	.773	.639	.752	.765	.245	.590	.601	.816	.706
The proposed NB (F+G+H+U)	.774	.639	.752	.766	.239	.591	.603	.817	.707
The proposed NB (F+G+H)	.772	.641	.740	.752	.236	.593	.603	.834	.726
Baseline (F+G) [6]	.746	.617	.692	.735	.211	.541	.614	.814	.714

perceived empathy state e[†]. Head gesture is often produced to show attitude towards other's opinion, and utterance is a measure of conversational role, i.e. speaker or listener. Note that most utterance states can be judged only from images.

The relationship between perceived empathy and these behaviors are modeled with co-occurrence matrices; $P(b_{i,t}, b_{j,t}|e_t)$, where $b_{i,t} = k$ denotes that person i is showing behavior category k in channel b at time t. The number of possible states of the co-occurrence of each behavior channel except for gaze is $N_b \times N_b$, where N_b is the number of behavioral categories in channel b; N_b is six, six[††], and two for facial expression, head gesture, and utterance, respectively. The number of facial expression categories is described in Sect. 5.4. The number of co-occurrence states of gaze is three; mutual gaze, one-way gaze, and averted gaze [6].

5.4 Estimation Experiment Setup

By following [6], we quantitatively evaluated the proposed model based on the similarity of the conditional distributions $P(e_t|\boldsymbol{B})$ to the distributions made by external observers, i.e. voting rates, for each time t. The participant behaviors $\boldsymbol{B}$, i.e. the observation in this study, are the labels that are annotated by three (for facial expression) or one (for other behavior channels) observer(s), as described in Sect. 4. In this experiment, the number of facial expression categories is set to six, because many of the original 18 categories are infrequent in the data, as described in Sect. 4.3. Five of the six categories are the five most frequently used categories, i.e. neutral, smile, thinking, wry smile, and laughter. The sixth category ("others" category) covers all remaining categories. This grouping is expected to be a practical way to enhance system robustness against the ambiguity of hand labeling of infrequent facial expressions.

Inference performances explained below are the average of the three performances, each of which is calculated by using the facial expression labels annotated by one of the three coders. This paper employs the leave-one-conversation-group-out cross validation approach. This evaluates how well perceived empathy distributions created by a specific observer group for an unseen conversation can be replicated by the model; each probability distribution in the right hand of Eq. (1) is trained by using all data except for the target conversation group. These probability distributions are trained based on how often each target state is observed in the training samples.

As the similarity measure between two probability distributions $\boldsymbol{p}$ and $\boldsymbol{q}$ (C-dimensional vectors), this paper utilizes overlap area (OA), because it is a widely used form of similarity [45]. In our case, C is the number of categories of perceived empathy, i.e. $C = 3$. The OA is calculated as $\mathrm{OA}(\boldsymbol{p}, \boldsymbol{q}) = \sum_{i=1}^{C} \min(p_i, q_i)$, where p_i and q_i denote the i-th component of $\boldsymbol{p}$ and $\boldsymbol{q}$, respectively. OA becomes one (zero) at maximum, i.e. perfect inference, (minimum, i.e. worst inference). The present study calculates frame average OAs and distribution type average OAs. By following [6], seven distribution types are defined; X-dominant, X-inferior, and flat distribution types, where X means one of Empathy, Neither and Antipathy.

5.5 Estimation Results

Table 4 shows average OAs in the effective ranges of the timing model, i.e. the regions that satisfy $0 < \pi_t \leq 1$ in Eq. (3). The total number of samples in the ranges is 36,771. The average performance of these distribution types compensates the unbalance of the number of samples among the distri-

[†]As for facial expression and gaze, the previous studies [6], [10] consider the cross-channel co-occurrence between facial expressions and mutual gaze between a pair, while this paper assumes their conditional independence. We have confirmed that this assumption yields comparable results. To avoid overfitting, this paper assumes the independence that reduces model parameters.

[††]Only head gesture, which was originally labeled with 11 categories, was devolved into 6 categories that maximize the inference performance only with head gesture, $P(e|\boldsymbol{B}) := P(e)P(g|e)$, by using the sequential backward selection technique.

bution types. Both head gesture and utterance (H and U in Table 4) well enhance the inference performance compared with the baseline model [6]. Besides, as expected from the results in Sect. 4, the timing model (Ft in Table 4) succeeds in further increasing the OA for antipathy-dominant scenes from OA = .401 without the timing model to OA = .484, without noticeable loss of the OAs for other distribution types. Moreover, we have also confirmed that other similarity measures such as Bhattacharyya coefficient and root mean square error yield comparable results to those with OA.

Table 5 shows average OAs for all frames in a comparison of a family of our naïve Bayes model against a baseline model [6]. The number of samples in total is 297,705, The introduction of head gesture and utterance increases both the OAs of frame average and distribution type average, like Table 4.

5.6 Discussion

The experiment demonstrates the effectiveness of the timing model. However, it is difficult for the timing model in its current form to fully cover all interaction scenes, because people can act or react in other behavioral channels, e.g. head nod to head nod, or smile to utterance. Such temporal structures of behaviors would require a multimodal analysis like [46]. As mentioned in Sect. 4, appropriately handling such cross-channel congruence would extend the effective range and enhance the performance of the timing model. In addition, this study empirically set the end of the effective range of timing model, i.e. W, to be 4 s. To examine how long the same perception is maintained after the emergence of action and reaction is also an interesting issue.

This paper provides examples of the timing at which observers changed their perceived empathy labels during action and reaction of a pair of conversation participants. The present study allows the observers to replay the video as many times as desired, and even to reverse the video to determine the change timing of their perception. However, if they watched the video just once in the forward direction at normal speed, the timing is expected to be at, or just after, the emergence of reaction; the weight function, shown in Fig. 5, might depend on viewing condition. It would be also interesting to compare the perception labels obtained under such conditions with the ones gathered here.

Furthermore, this paper judges facial congruence based on whether the facial expression categories of the pair were the same or not. However, the validity of this semantic categorization was not examined. Because semantic categorization, especially in facial expressions, would differ with the annotators, non-semantic description, e.g. physical-motion-based description like FACS's AU [32], would be more appropriate.

6. Conclusion

The present study analyzed empathy and antipathy aroused between people while interacting in face-to-face conversations. By focusing on the process by which they are perceived by external observers, this paper investigated the perception tendency, and from it developed a computational model for the automatic inference of perceived empathy/antipathy. This study first demonstrated that the observer's perception of an interacting pair is affected both by the time lag between their action and reaction in facial expression and by whether their expressions are congruent or not. Based on the findings, this paper proposed a probabilistic model that relates the perceived emotion of observers to the action and reaction of conversation participants. An experiment conducted on the data of ten conversations hold by 16 women and perceived empathy of nine external observers demonstrated that such a timing cue is helpful in improving the inference performance, especially for antipathy.

References

[1] D. Gatica-Perez, "Analyzing group interactions in conversations: a review," Proc. IEEE Int'l Conf. Multisensor Fusion and Integration for Intelligent Systems, pp.41–46, 2006.
[2] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," Language Resources And Evaluation, vol.42, no.4, pp.335–359, 2008.
[3] Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," IEEE Trans. Pattern Anal. Mach. Intell., vol.31, no.1, pp.39–58, 2009.
[4] M. Nicolaou, H. Gunes, and M. Pantic, "Output-associative RVM regression for dimensional and continuous emotion prediction," Proc. IEEE Int'l Conf. FG'11, 2011.
[5] K. Otsuka, "Conversation scene analysis," IEEE Signal Process. Mag., vol.28, pp.127–131, 2011.
[6] S. Kumano, K. Otsuka, D. Mikami, M. Matsuda, and J. Yamato, "Understanding communicative emotions from collective external observations," Proc. CHI '12 Extended Abstracts on Human Factors in Computing Systems, pp.2201–2206, 2012.
[7] J.N. Cappella, "Mutual influence in expressive behavior: Adult-adult and infant-adult dyadic interaction," Psychological Bulletin, vol.89, no.1, pp.101–132, 1981.
[8] J.K. Burgoon, L.A. Stern, and L. Dillman, Interpersonal Adaptation: Dyadic Interaction Patterns, Cambridge University Press, 1995.
[9] 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.
[10] 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," Proc. IEEE Int'l Conf. FG'11, pp.43–50, 2011.
[11] S. Kumano, K. Otsuka, M. Matsuda, and J. Yamato, "Analyzing perceived empathy/antipathy based on reaction time in behavioral coordination," IEEE Int'l Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp.1–8, 2013.
[12] U. Dimberg, M. Thunberg, and S. Grunedal, "Facial reactions to emotional stimuli: Automatically controlled emotional responses," Cognition and Emotion, vol.16, no.4, pp.449–471, 2002.
[13] T.W. Lee, R.J. Dolan, and H.D. Critchley, "Controlling emotional expression: Behavioral and neural correlates of nonimitative emotional responses," Cerebral Cortex, vol.18, no.1, pp.104–113, 2008.
[14] M. Thunberg, Rapid Facial Reactions to Emotionally Relevant Stimuli, Ph.D. thesis, Uppsala University, Department of Psychology, 2007.

[15] E.J. Moody, D.N. McIntosh, L.J. Mann, and K.R. Weisser, "More than mere mimicry? the influence of emotion on rapid facial reactions to faces," Emotion, vol.7, no.2, pp.447–457, 2007.
[16] J.T. Lanzetta and B.G. Englis, "Expectations of cooperation and competition and their effects on observers' vicarious emotional responses," J. Pers. Soc. Psychol., vol.56, no.4, pp.534–554, 1989.
[17] E.P. Bucy and S.D. Bradley, "Presidential expressions and viewer emotion: Counterempathic responses to televised leader displays," Social Science Information, vol.43, no.1, pp.59–94, 2004.
[18] P. Bourgeois and U. Hess, "The impact of social context on mimicry," Biological Psychology, vol.77, no.3, pp.343–352, 2008.
[19] 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.
[20] R.W. Levenson and A.M. Ruef, "Empathy: A physiological substrate," J. Pers. Soc. Psychol., vol.63, no.2, pp.234–246, 1992.
[21] R. Cowie and R.R. Cornelius, "Describing the emotional states that are expressed in speech," Speech Commun., vol.40, no.1-2, pp.5–32, 2003.
[22] T. Bänziger and K. Scherer, "Using actor portrayals to systematically study multimodal emotion expression: The GEMEP corpus," Proc. Int'l Conf. on Affective Computing and Intelligent Interaction, pp.476–487, 2007.
[23] S. Feese, B. Arnrich, G. Tröster, B. Meyer, and K. Jonas, "Quantifying behavioral mimicry by automatic detection of nonverbal cues from body motion," Proc. Int'l Conf. on Privacy, Security, Risk and Trust (PASSAT), pp.520–525, 2012.
[24] S. Bilakhia, S. Petridis, and M. Pantic, "Audiovisual detection of behavioural mimicry," Proc. Affective Computing and Intelligient Interaction (ACII), pp.123–128, 2013.
[25] L.P. Morency, I. de Kok, and J. Gratch, "A probabilistic multimodal approach for predicting listener backchannels," Autonomous Agents and Multi-Agent Systems, vol.20, no.1, pp.70–84, 2009.
[26] C.D. Batson, "These things called empathy: Eight related but distinct phenomena," in The Social Neuroscience of Empathy, ch. 1, pp.3–15, MIT Press, 2009.
[27] S.D. Preston and F.B. de Waal, "Empathy: Its ultimate and proximate bases," Behavioral and Brain Sciences, vol.25, no.1, pp.1–20, 2002.
[28] E. Stotland, "Exploratory investigations of empathy," Advances in experimental social psychology, vol.4, pp.271–314, 1969.
[29] 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.
[30] N. Chovil, "Discourse-oriented facial displays in conversation," Res. on Lang. and Social Int., vol.25, pp.163–194, 1991.
[31] P. Ekman and W.V. Friesen, "The repertoire of nonverbal behavior: Categories, origins, usage, and coding," Semiotica, vol.1, pp.49–98, 1969.
[32] P. Ekman and W.V. Friesen, The Facial Action Coding System: A Technique for the Measurement of Facial Movement, Consulting Psychologists Press, 1978.
[33] S. Baron-Cohen, Mind Reading: The Interactive Guide to Emotions, Jessica Kingsley Publishers, 2004.
[34] K.L. Gwet, Handbook of Inter-Rater Reliability (3rd Edition), Advanced Analytics, LLC, Gaithersburg, MD, 2012.
[35] W. Sato and S. Yoshikawa, "Spontaneous facial mimicry in response to dynamic facial expressions," Cognition, vol.104, no.1, pp.1–18, 2007.
[36] S. Korb, D. Grandjean, and K.R. Scherer, "Timing and voluntary suppression of facial mimicry to smiling faces in ago/nogo task - an EMG study," Biological Psychology, vol.85, pp.347–349, 2010.
[37] A. Grecuccia, R.P. Cooperb, and R.I. Rumiatia, "A computational model of action resonance and its modulation by emotional stimulation," Cogn. Sys. Res., vol.8, no.3, pp.143–160, 2007.
[38] K. Bousmalis, M. Mehu, and M. Pantic, "Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools," Proc. Conf. on Affective Computing and Intelligent Interaction (ACII), pp.1–9, 2009.
[39] E. Keller and W. Tschacher, "Prosodic and gestural expression of interactional agreement," COST 2102 Workshop, pp.85–98, 2007.
[40] 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.
[41] N. Chovil, The psychology of facial expression, ch. Facing others: A social communicative perspective on facial displays, pp.321–333, Cambridge University Press, Cambridge, UK, 1997.
[42] S. Asch, "Opinions and social pressure," Scientific American, vol.193, no.5, pp.31–35, 1955.
[43] P. Domingos and M.J. Pazzani, "Beyond independence: Conditions for the optimality of the simple Bayesian classifier," Proc. Int'l Conf. on Machine Learning (ICML), pp.105–112, 1996.
[44] G.R. Jonsdottir, J. Gratch, E. Fast, and K.R. Thórisson, "Fluid semantic back-channel feedback in dialogue: Challenges & progress," Proc. Int'l Conf. Intelligent Virtual Agents (IVA), 2007.
[45] S. Cha and S.N. Srihari, "On measuring the distance between histograms," Pattern Recognit., vol.35, pp.1355–1370, 2002.
[46] M.M. Louwerse, R. Dale, E.G. Bard, and P. Jeuniaux, "Behavior matching in multimodal communication is synchronized," Cognitive Science, vol.36, no.8, pp.1404–1426, 2012.

Shiro Kumano received the PhD degree in Information Science and Technology from the University of Tokyo in 2009. He is currently a researcher at NTT Communication Science Laboratories. His research interests include computer vision, human behavior analysis, and automatic meeting analysis, especially in facial expression recognition and empathy inference. He received the ACCV 2007 Honorable Mention Award, the MIRU 2011 Interactive Session Award, and the HCG Symposium 2012 Excellent Interactive Presentation Award. He is a member of the IEEE, and IPSJ.

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.

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 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.