Analyzing Perceived Empathy/Antipathy based on Reaction Time in Behavioral Coordination

Shiro Kumano, Kazuhiro Otsuka, Masafumi Matsuda, and Junji Yamato

Abstract— **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 estimation 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 estimation performance, especially for perceived antipathy.**

I. 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], [2], [3], [4], [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 estimate what type of basic emotion a target person is really feeling. Recently, from a more communication-directed viewpoint, Kumano et al. [6] proposed to estimate how emotions aroused between a pair of people in multi-party conversation are perceived by external observers. Their targets were empathy as emotional contagion, and antipathy or counter-empathy as emotional conflict. Their work extends the research area on automatic meeting analysis and emotion estimation.

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], [8], [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 (FEs) 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 coordination and response time 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 similar to those of human physical reactions to emotion-eliciting stimuli. Previous psychological studies, e.g. [11], [12], [13], [14], have studied and documented the characteristics of imitated facial reactions when exposed to another human's

S. Kumano, K. Otsuka, M. Matsuda and J. Yamato are with NTT Communication Science Laboratories, 3-1 Morinosato-Wakamiya, Atsugi, Kanagawa, Japan. kumano@ieee.org

 Find the published version of this article under http://dx.doi.org/10.1109/FG.2013.6553812.

TABLE I

KEY HYPOTHESIS OF THE PRESENT STUDY: EXPECTED DOMINANT OBSERVER PERCEPTION

Behavioral coordination between a pair	Reaction time	
	Rapid (300-400 ms)	Delayed (500-1,000 ms)
Congruent	Empathy	Antipathy
Incongruent	Antipathy	Empathy

facial expression. In short, as described in detail in II, both the reaction time and the interpersonal relationship between the pair change the coordination/incoordination of the interaction. Table I 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, because it's currently intractable to fully consider all possible combinations of single- and cross-channel behavioral coordination 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 IV, our experiment demonstrates that the timing model is helpful in improving the estimation performance, especially for antipathy, and so is superior to the use of the static model only.

The remainder of this paper first introduces related works to position this study and to derive the present hypothesis in II. Next, our definition of perceived empathy/antipathy is explained in III. A psychological study that assesses the impact of timing and coordination between action-reaction is detailed in IV. A probabilistic model for estimating perception from participant behaviors is described and evaluated in V. A discussion and the potential for future growth are given in VI. Finally, we summarize this study in VII.

II. 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 receiver* of emotion-eliciting stimuli, including other's facial expressions. For example, when observing other's emotional face, people involuntarily and rapidly mimic the presented FE, even for negative emotional displays [11], e.g. to smile at a happy face or frown at an angry face. Many previous studies, e.g. [11], [12], [13], [14], reported that the response time in facial electromyographic (EMG) reactions is around 300-400 msec, if the subject is prepared to mimic the presented FE. 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 msec [11], [12]. These mimicking patterns, whether congruent or not, depend strongly on context, e.g. relationship between the subject and the persons issuing the target FE [15], [16], [17]. For example, congruent reaction, i.e. a reaction the same as or similar to the target FE, is likely to be produced for a cooperative partner, while an incongruent reaction is expected from a competitive partner [15].

In pioneering works on perceived empathy, Ickes et al. [18] and Levenson et al. [19] 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 [18] or external observers [19]. Although Levenson et al. [19] demonstrated physiological linkage between the target and the external observers, neither of them focused on behavioral coordination between a pair or its time lag, nor proposed any computational models of perceived empathy that could realize automatic estimation.

Kumano et al. [6] recently proposed a research framework for estimating how differently empathy/antipathy as emotional contagion/conflict between a conversational pair will be perceived by external observers. Their key contributions are to focus on the distribution of perception made by multiple observers, i.e. inter-observer difference, which is ignored by most previous studies which use averaging or majority voting. They posed the problem setting of estimating the distribution among target categories (empathy, antipathy or neither), i.e. voting rates. 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 referenced paper considers neither other behavioral channels, e.g. head gesture and utterance, nor dynamic cues.

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. [20] 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, but did not address the perception of such interactions.

III. OUR 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 [18] described in II. On the other hand, perceived empathy in [6] 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.

Among the eight distinct phenomena of empathy, the definition of empathy in [6] is most strongly associated with "emotional contagion" and "imagine-other perspective" [21]. "Emotional contagion" is the state wherein the emotion of

the subject is the same as or similar to that of the object [22]. Like [6], this paper defines empathy as emotional contagion, and antipathy as emotional conflict. Emotional contagion provides a good match to our concept that emotions in conversation are shared between participants via their interaction. Understanding this phenomenon demands the pairwise treatment of emotions. "*Imagine-other* perspective" [23] is a way of viewing others; the observer imagines how the target participants are feeling. Because the observed others in the present study case are a pair of participants, the perspective of [6] can be called "*imagine-pair* perspective". 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 [24]. By following the guideline in [24], 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.

IV. Timing Analysis: A Psychological Study

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

A. Hypotheses

From existing findings on the characteristics of human reaction to emotion-eliciting stimuli described in II, we explore the possibility that time lags and coordination can be used to estimate observers' perceived empathy. As a key nonverbal behavior for observers to judge emotions of conversation participants, facial expressions (FEs), i.e. facial action and reaction, are focused here; FE is the primary channel for the transmission of emotion [25].

Our basic hypothesis is that the external observer consciously or unconsciously utilizes the characteristics of human reaction 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 coordination of the interaction [15], [16], [17], 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 [21], [22]; another aspect is behavioral coordination [21].

The hypothesis is decomposed into the following four sub hypotheses in terms of the coordination and delay between action and reaction: (H1.1) Observers are likely to perceive empathy when the FEs of the target pair are *congruent* with lag of *300-400* msec. (H1.2) Observers are *unlikely* to perceive empathy when the FEs of the target pair are *congruent* with lag of *500-1,000* msec. (H2.1) Observers are likely to perceive antipathy when the FEs of the target pair are *incongruent* with lag of *300-400* msec. (H2.2) Observers are *unlikely* to perceive antipathy when the FEs of the target pair are *incongruent* with lag of *500-1,000* msec. Table I summarizes these hypotheses.

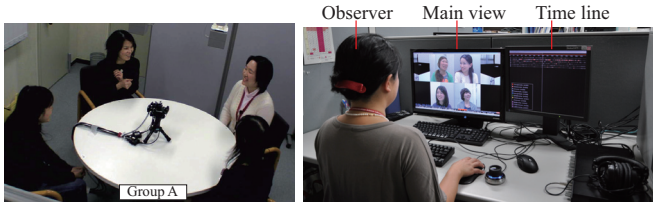

Fig. 1. Snapshots of conversation (left) and labeling (right) scenes.

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 300-400 msec) 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 (H1.1). If an incongruent response is rapidly returned, e.g. negative FE to smile, the observer would perceive spontaneously displayed antipathy (H2.1). If the listener displayed a smile but after some delay (lag of 500-1,000 msec), 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 (H1.2).

B. 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.

C. 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 discussion topics were "Who are more beneficial, men or women?", "Is marriage and romantic love the same or different?" etc. 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 [22]. They had not met before the experiment. They were asked to first hold a short chat with self-introduction, then hold about seven discussions with some intervals between them.

Most discussion topics were assigned to the groups on the basis of the participants' opinions found in pre-questionnaires so as to cause frequent concordance and disagreement. The discussions were held on a single day for each group. Focusing on the most lively exchanges, this paper picks up and analyzes ten discussions: four 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.). All conversations were captured at 30 fps by IEEE1394 color cameras. One observer 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. The label sets for FE, head gesture, and utterance are {neutral, smile, laughter, wry smile, thinking, others}[10], {no gesture, nod (3 levels), shake (3 lv.), tilt (3 lv.), their combination}, and {speaking, silence}, respectively.

D. Procedure: labeling of perceived empathy

Videos were viewed and labeled using our original software [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 given labels. 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). 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.

E. Analysis

This analysis aims to investigate whether the perceived empathy of observers is really affected by both the time lag and coordination 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 coordination.

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 V. The frequency is separately calculated for each time lag, dt, and each coordination state, c, between action and reaction. Time lag dt is grouped into 0-500 msec (rapid) and 500-1,000 msec (delayed) in this section. Coordination state c is a binary state describing whether the action and reaction were the same or not. In judging coordination, the six categories of FE were grouped into three emotional tones; positive (smile or laughter), neutral (neutral or thinking), and negative (wry smile or others). 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 coordination 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.

TABLE II

ORIGINAL FREQUENCY OF PERCEIVED EMPATHY

(a) Congruent facial expressions

Reaction time	Empathy	Neither	Antipathy
Rapid (0-500 msec)	314	95	8
Delayed (500-1,000 msec)	185	67	10

(b) Incongruent facial expressions

Reaction time	Empathy	Neither	Antipathy
Rapid (0-500 msec)	102	43	15
Delayed (500-1,000 msec)	67	56	8

Each frequency distribution pair between different conditions are compared by using the chi-square test ($df = 2$). Four conditions yield six ($={}_4C_2$) pairs. 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 coordination affect observers' perceived empathy. Note that the original, i.e. unnormalized, frequencies are used in the chi-square test.

F. Results and Discussion

Table II shows the unnormalized frequencies of perceived empathy. Fig. 2 shows the distributions of normalized frequency[1] N' for each condition with statistically significant difference in distribution between conditions. In short, as expected, significant differences ($p < 0.05$) were found among most condition pairs. These results basically support the validity of our hypotheses; observer perception, especially antipathy, strongly depends on both the time lag and coordination between action and reaction. More observers

[1]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.

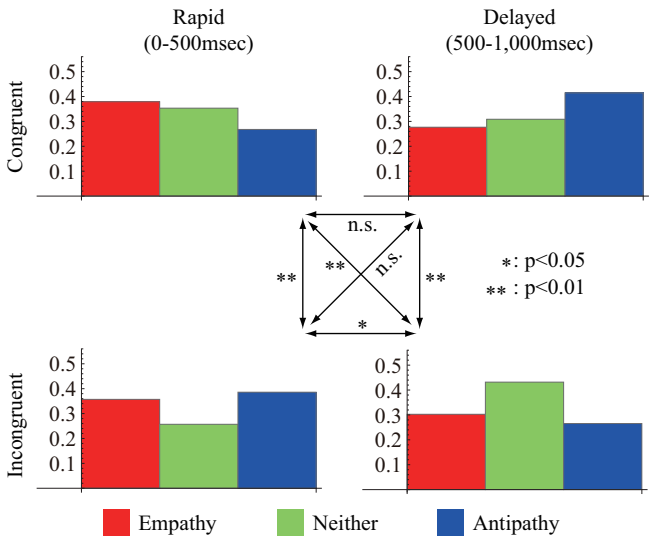

Fig. 2. Normalized frequencies of perceived empathy for each coordination (row) and reaction time (column), i.e. $N'_{dt,c}(e)$: Vertical axes denote normalized frequencies. Most of these distribution pairs with regard to the original, i.e. unnormalized, frequencies, i.e. $N_{dt,c}(e)$, are significantly different ($p < 0.05$).

TABLE III

SUMMARY OF RESULTING DOMINANT TENDENCY IN OBSERVER PERCEPTION

FE coordination between a pair	Reaction time	
	Rapid (0-500 msec)	Delayed (500-1,000 msec)
Congruent	Empathy and Neither	Antipathy
Incongruent	Empathy and Antipathy	Neither

and conversations might increase the statistical significance for most distribution pairs.

Fig. 2 demonstrates that when facial interactions are congruent and rapid (upper left distribution), makes their perception more likely to be Empathy or Neither. Though no remarkable difference between Empathy and Neither can be found, H1.1 can be partly accepted. In contrast, congruent and delayed interactions (upper right distribution) are hard to perceive as Empathy, so it's reasonable to accept H1.2. Incongruent and rapid FEs (lower left distribution) are the most likely to be recognized as Antipathy, so it is reasonable to accept H2.1. Although Neither is the most frequent in incongruent and delayed responses (lower right distribution), Antipathy is infrequent on this condition. Thus, some part of H2.2 is acceptable. Fig. III 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 FE incoordination and delayed FE coordination, unlike the other reactions. Although the targets are different, these tendencies are consistent with those obtained in previous works, e.g. [19].

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

V. PROBABILISTIC MODEL

Based on the results in IV, 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.

A. 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 handled as probabilities in the framework of Bayesian theory. 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 posterior 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 posterior 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 [26]. 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 posterior probability distribution $P(e_t|\boldsymbol{B})$ is decomposed as:

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

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 coordination c, i.e. whether the categories of their behaviors are the same or

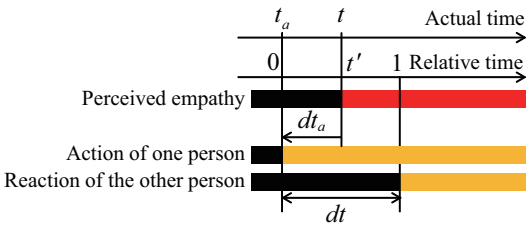

Fig. 3. Time lag between perceived empathy change and the beginning of action and reaction. Different colors mean different categories.

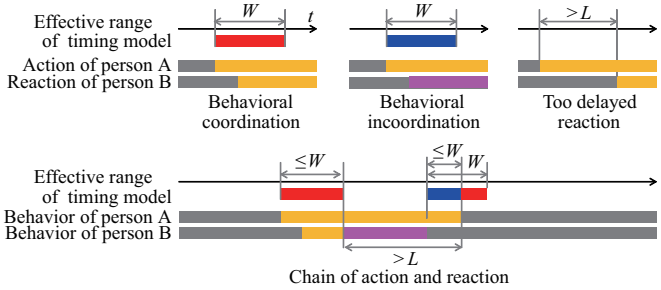

Fig. 4. Effective range of timing model: Upper left two figures show the cases of behavioral coordination and incoordination. Red and blue bars denote the effective ranges of the timing model for behavioral coordination and incoordination, respectively. Upper right figure show the case where the timing 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.

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 prior probability of e; it describes how likely the target interactions are to being labeled with perceived empathy e without considering any explanatory variable.

B. Timing model

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

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

That is, it combines the likelihood of perceived empathy e in behavioral coordination/incoordination 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 perceived empathy at/or around the behavioral coordination, π_t. Only FE is considered in the timing model in the present study.

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 coordination 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

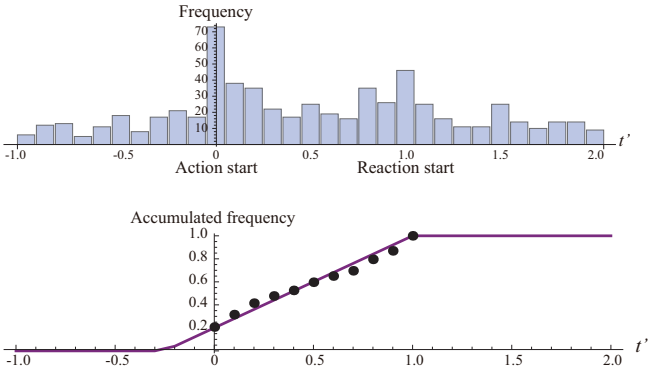

Fig. 5. (Upper) Timing between perceived empathy and action/reaction behaviors. Horizontal axis denotes relative time, t', in each interaction. Relative times $t' = 0$ and $t' = 1$ mean the beginning time of action and reaction, respectively. Vertical axis is the frequency that perceived empathy was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

of a histogram. To avoid overfitting due to the limited sample size, the bin size is set to be 300 msec in this paper. Fig. 2 is an example of this function with the bin size of 500 msec.

2) Change timing function π: The change timing function π describes at which timing the change in 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. It determines the weight and effective range of the timing model in the full model ((1)).

We use the following ramp function to model the change timing:

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

where α is 0.2. $\pi = 0$ means that the timing model makes no contribution in (1). 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 sec with reference to Jonsdottir et al.'s finding [27] that delay between lexically key phrases and the facial expression of the listener lies approximately in the range of 500-2,500 msec. 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 sec. 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 $p = \alpha + (1 - \alpha) \cdot t'$.

C. 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 for FE and gaze, we follow [6] and [10], which demonstrated the effectiveness of modeling the co-occurrence of FEs for each gaze state between a pair. The gaze state is mutual gaze, one-way gaze, or averted gaze [6]. The present study additionally introduces static models for head gesture and utterance. 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 number of possible states of the co-occurrence of head gestures is, for example, $N_g \times N_g$, where N_g is the number of head gesture categories[2]. Utterance is modeled in the same manner.

D. Estimation experiment setup

By following [6], we quantitatively evaluated the proposed model based on the similarity of the posterior distributions $P(e_t|\boldsymbol{B})$ to the distributions made by external observers, i.e. voting rates, for each time t. The participant behaviors $\boldsymbol{B}$, annotated by one observer as described in IV, are taken as the observation in this study. This paper employs the leave-one-conversation-group-out cross validation approach. This evaluates how well perceived empathy distributions created by an unseen observer group can be replicated by the model; each probability distribution in the right hand of (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 [28]. 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 estimation, (minimum, i.e. worst estimation). The present study calculated 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.

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

E. Estimation results

Table IV shows average OAs in the effective ranges of the timing model. As expected from the results in IV, the timing model (Ft in Table IV) succeeds in significantly increasing the OA for antipathy-dominant scenes from OA = .696 without the timing model to OA = .831, 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 V shows average OAs for all frames in a comparison of a family of our naïve Bayes model against a baseline model [6]. The introduction of head gesture and utterance increases both the OAs of frame average and distribution type average. Like Table IV, the timing model further improves the estimation performance especially for antipathy, though some antipathy-dominant scenes lay out of the effective range.

VI. 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 head nod to smile. As mentioned in IV, appropriately handling such cross-channel coordination 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 sec. 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. It would be also interesting to compare the perception labels obtained under such conditions with the ones gathered here.

Furthermore, this paper judges FE coordination based on whether the FE 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 [29], would be more appropriate.

VII. 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

TABLE IV

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+X+G+U)	.757	.693	.765	.750	.696	.577	.625	.783	.654
The proposed NB (F+X+G+U+Ft)	.755	.709	.763	.743	.831	.565	.622	.784	.654

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

TABLE V

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)
Baseline (F+X) [6]	.732	.621	.679	.687	.250	.521	.620	.850	.710
The proposed NB (F+X+G+U)	.766	.631	.742	.773	.253	.541	.616	.798	.693
The proposed NB (F+X+G+U+Ft)	.765	.632	.741	.773	.265	.538	.616	.798	.693

for the automatic estimation 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 estimation performance, especially for antipathy. We believe that the present study will enlarge the scope of research areas on automatic meeting analysis, emotion model and its estimation, and emotion in psychology.

REFERENCES

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

[2] 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. PAMI*, 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," in *Proc. IEEE Int'l Conf. FG'11*, 2011.

[5] K. Otsuka, "Conversation scene analysis," *IEEE Signal Processing Magazine*, 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," in *Proc. CHI '12 extended abstracts on Human factors in computing systems*, 2012, pp. 2201–2206.

[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," in *Proc. IEEE Int'l Conf. FG'11*, 2011, pp. 43–50.

[11] 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.

[12] 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.

[13] M. Thunberg, "Rapid facial reactions to emotionally relevant stimuli," Ph.D. dissertation, Uppsala University, Department of Psychology, 2007.

[14] 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.

[15] 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.

[16] 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.

[17] P. Bourgeois and U. Hess, "The impact of social context on mimicry," *Biological Psychology*, vol. 77, no. 3, pp. 343–352, 2008.

[18] 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.

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

[20] 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.

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

[22] 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.

[23] E. Stotland, "Exploratory investigations of empathy," *Advances in experimental social psychology*, vol. 4, pp. 271–314, 1969.

[24] 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.

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

[26] P. Domingos and M. J. Pazzani, "Beyond independence: Conditions for the optimality of the simple Bayesian classifier," in *Proc. ICML*, 1996, pp. 105–112.

[27] G. R. Jonsdottir, J. Gratch, E. Fast, and K. R. Thórisson, "Fluid semantic back-channel feedback in dialogue: Challenges & progress," in *Proc. Int'l Conf. Intelligent Virtual Agents (IVA)*, 2007.

[28] S. Cha and S. N. Srihari, "On measuring the distance between histograms," *Pattern Recognition*, vol. 35, pp. 1355–1370, 2002.
[29] P. Ekman and W. V. Friesen, *The Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, 1978.