# Automatic Inference of Cross-modal Nonverbal Interactions in Multiparty Conversations

## "Who Responds to Whom, When, and How?" from Gaze, Head Gestures, and Utterances

Kazuhiro Otsuka
NTT Communication Science Laboratories
3-1, Morinosato-Wakamiya
Atsugi, 247-0198 Japan
otsuka@eye.brl.ntt.co.jp

Hiroshi Sawada
NTT Communication Science Laboratories
2-4, Hikaridai, Seika-cho
Kyoto, 619-0237 Japan
sawada@cslab.ntt.co.jp

Junji Yamato
NTT Communication Science Laboratories
3-1, Morinosato-Wakamiya
Atsugi, 247-0198 Japan
yamato@brl.ntt.co.jp

## ABSTRACT

A novel probabilistic framework is proposed for analyzing cross-modal nonverbal interactions in multiparty face-to-face conversations. The goal is to determine "who responds to whom, when, and how" from multimodal cues including gaze, head gestures, and utterances. We formulate this problem as the probabilistic inference of the causal relationship among participants' behaviors involving head gestures and utterances. To solve this problem, this paper proposes a hierarchical probabilistic model; the structures of interactions are probabilistically determined from high-level conversation regimes (such as monologue or dialogue) and gaze directions. Based on the model, the interaction structures, gaze, and conversation regimes, are simultaneously inferred from observed head motion and utterances, using a Markov chain Monte Carlo method. The head gestures, including nodding, shaking and tilt, are recognized with a novel Wavelet-based technique from magnetic sensor signals. The utterances are detected using data captured by lapel microphones. Experiments on four-person conversations confirm the effectiveness of the framework in discovering interactions such as question-and-answer and addressing behavior followed by back-channel responses.

## Categories and Subject Descriptors

H1.2 [**Models and Principles**]: User/Machine System — Human Information Processing

## General Terms

ALGORITHMS, HUMAN FACTORS


ICMI November 12-15, 2007, Nagoya, Aichi, Japan.


## Keywords

Face-to-face multiparty conversation, Eye gaze, Head gestures, Nonverbal behaviors, Bayesian network, Markov chain Monte Carlo, Gibbs sampler, Semi-Markov process

## 1. INTRODUCTION

Face-to-face conversation is one of the most basic forms of communication in our life and is used for conveying/sharing information, understanding others' intention/emotion, and making decisions. To enhance our communication capability beyond conversations on the spot, the automatic analysis of conversation scenes is a basic technical requisite to enable effective teleconferencing, archiving/summarizing meetings, and to realize communication via social agents and robots. The conversation scene analysis targets various aspects of conversations, from individual/group behaviors such as "*who is speaking now?*" and "*who is talking/listening to whom?*", to context/mental status such as "*who made him angry?*" and "*why is she laughing?*".

In the face-to-face setting, the messages include not only verbal but also nonverbal messages. The nonverbal messages are expressed by nonverbal behaviors in multimodal channels such as eye gaze, facial expressions, head motion, hand gesture, body posture and prosody; psychologists have elucidated its importance in human communications [1, 26]. Therefore, it is expected that conversation scenes can be largely understood by observing people's nonverbal behaviors with sensing devices such as cameras and microphones.

As a preliminary step, authors have focused on eye gaze as a nonverbal cue for recognizing addressing/listening behaviors [22], based on the importance of gaze functionality in conversations [16, 12]. We have conducted a frequency analysis of the set of gaze directions of all participants, we call it *gaze pattern*, and hypothesized that the topology of gaze patterns (convergence and mutual gaze) can indicate the pattern of conversations such as monologue and dialogue; we call these the *conversation regimes*. As an example, one relationship is "*hearers tend to look at speaker in monologues*". To model the relationships between gaze patterns and regimes, we proposed a probabilistic conversation model based on a dynamic Bayesian network; the conversational regime controls the dynamics of gaze patterns and utterances; the gaze pattern is a hidden variable and is esti-

mated from head-direction measurements. With this model, the regimes and the gaze patterns are jointly estimated from the utterances and the head directions measured with sensors [22] or face tracking in videos [23]; the estimation is implemented using the MCMC(Markov chain Monte Carlo) method.

This paper tries to extend our framework to a new target, the automatic inference of nonverbal interaction structures in multiparty conversations; the goal is to determine "*who responds to whom, when, and how*". In contrast to authors' previous work [22, 23], this paper tries to recognize more direct nonverbal interactions in conversations. We particularly focus on head gestures (nod, shake, and tilt) and utterances as nonverbal cues, and try to discover the action-reaction pairs of participants' behaviors such as question-and-answer and addressing followed by back-channel responses. The target interactions will yield cross-modal formations such as an utterance acknowledged by nodding where the nod triggers the other's utterances. The interaction structure is the basic primitive in conversations and can reveal how messages are exchanged among people; it can be a clue for inferring how the attitude and minds of participants change. As far as the authors know, this paper is the first one to shed light on the explicit structural analysis of nonverbal interactions in conversations.

We formulate this problem as the probabilistic inference of the causal relationships among participants' behaviors, we call these relationships the *interaction structures*. To solve this problem, this paper proposes a hierarchical probabilistic model; the interaction structures are probabilistically generated from gaze and conversation regimes; the interaction structures then determine how head gestures and utterances relate to each other, i.e. "*which behavior is triggered by which behavior*", as well as "*which behaviors are spontaneous and which are reactive*". Based on the model, the interaction structures, gaze patterns, and conversation regimes, are simultaneously inferred from head directions, head gesture intervals, and utterance intervals, using a MCMC.

One of the key features differentiating our model from existing interaction models is the modeling concept: explicit representation of the causal relationships among behaviors in Bayesian network form; the configuration of which is ruled by upper-layer processes, i.e. regimes and gaze. Another key feature is the use of a semi-Markov process [14] to accurately model the temporal structures of interactions; it permits arbitrary distributions of behavior timing, such as duration and pause length. As one such timing distribution, this study employs a Weibull distribution [21] due to its expressiveness. So far, several interaction models have been proposed for conversation scene analysis, based on the coupled-HMM [4] and its derivatives such as the influence model [2]. However, due to the Markov property of these models, the only exponential temporal distributions are supported, which does not necessarily match actual phenomena. Moreover, interaction modeling has, so far, mainly targeted audio modality [2, 7], and the modeling of multimodal interactions remains an open problem.

This paper focuses on head gestures such as nodding, shaking, tilt, for the following reasons. First, it is well known that head gestures play important roles in face-to-face conversation for both speakers and hearers [18]. The speaker's head gestures appear as visible signs of actions such as addressing, questioning, and stressing. The hearers' head gestures can be interpreted as signs of listening, acknowledgement, agreement/disagreement, and the level of understanding. These gestures are used to regulate various interactions in conversations, such as question & answer, addressing & back-channel response, and turn-taking/yielding. Therefore, head gestures are considered to be a rich information source for understanding conversations.

Several head gesture recognition methods have been proposed for man-machine interfaces using techniques such as HMM [15] and FFT+SVM(Support Vector Machine)[20]. Unlike interactions with artificial agents, human-human conversations exhibit a wide variety of gestures, in terms of periodicity, speed, and dynamic range, which are mixed together with other head motions such as those synchronized to utterances, turning head when changing gaze direction, and so on. To handle such gestures, this paper proposes a novel gesture recognition technique that consists of Wavelet-analysis of head pose sequences; SVM is used as a discriminator.

This paper is organized as follows. Section 2 overviews related works. Section 3 proposes our conversation model, and Section 4 presents an estimation algorithm based on the model. Section 5 describes the experiment conducted to verify the effectiveness of our method. Section 6 presents our conclusion and some discussions.

## 2. RELATED WORKS

In recent years, conversation scene analysis has emerged as an attractive research area [10], and two streams of research have been gaining attention: the automatic recognition of meeting actions, and using annotated data to explore human mechanisms in meetings.

The former study stream aims to realize the automatic recognition of meeting actions such as monologue, dialogue, discussion, note-taking, and presentation, from audio/visual signals. To do this, most prior studies employed low-level features such as global image motion and geometric image primitives detected from video, and tried to build statistical models on machine learning techniques that linked the signals to meeting actions. So far, a number of models have been proposed based on the HMM(Hidden Markov Model) [19], layered-HMM [27], coupled-HMM [2], and dynamic Bayesian networks [9]. However, the explicit measurement and modeling of human behaviors in meetings remain as open problems. On the other hand, another line of studies, motivated by the desire to explore human mechanisms in meetings, takes the psychological point of view. In pioneering work, a group led by Quek focused on the floor control function; they used a multimodal meeting corpus [6], created by a human expert, for analyzing speech patterns such as interruption and delegation of the floor [5]. So far, their research has revealed that multimodal cues such as gaze, gesture, speech, have important roles in floor control. However, full-automatic data annotation remains a future work.

Given the current status of the field, we have been trying to bridge the gap between the two streams of studies mentioned above: *automatic* understanding of meeting scenes from *direct* measurements of nonverbal behaviors, and *explicitly* modeling the relationship between individual behaviors and the status of conversation, with the help of *psychological* findings.

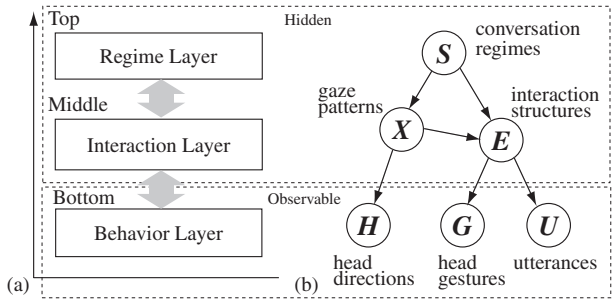


**Figure 1: Conversation model, (a) concept, (b) graphical model.**

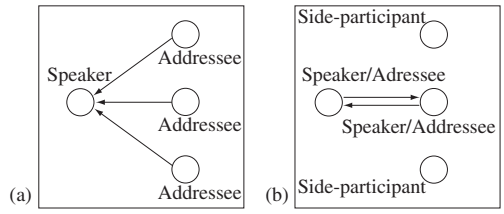


**Figure 2: Gaze patterns and interaction patterns in (a)convergence regime and (b)dyad-link regime.**

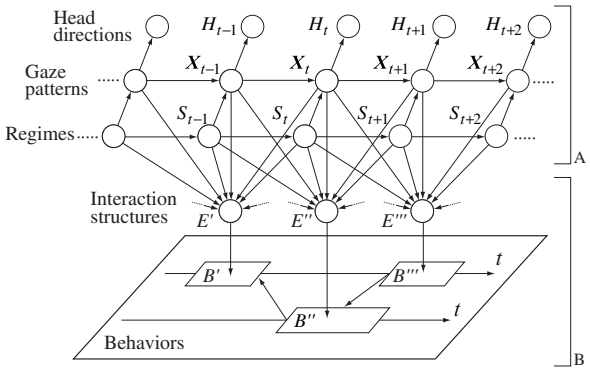


**Figure 3: Temporal representation of model.**

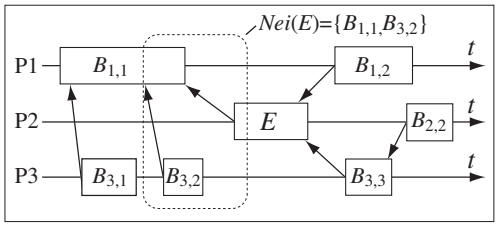


**Figure 4: Interaction network representing causal relationship determined from interaction structures.**

# 3. CONVERSATION MODEL

## 3.1 Model Concept

This study focuses on group conversations held in a closed environment; the number of participants is $N \geq 3$. As shown in Fig. 1(a), we assume that conversations have hierarchical structures. Further, we hypothesize that a high-level process, called a conversation regime, governs how people interact with each other on the interaction layer, and the interaction process governs how each individual behaves on the behavior layer.

In [22], authors have proposed conversation regimes as a global status of conversations, which correspond to addressing/listening patterns such as monologue and dialogue. We focused on the eye gaze of each participant as an interactive behavior. Also, we hypothesized that the gaze pattern of participants can indicate the structure of conversations, and proposed three regimes: *Convergence*, *Dyad-Link*, and *Divergence*. The regime called *Convergence* (also called monologue) corresponds to the situation that a speaker addresses all others, and the addressees listen to the speaker. This regime is indicated by the convergence of the addressees' gaze onto the speaker, as shown in Fig. 2(a). Second, the regime called *Dyad-Link* (also called dialogue) corresponds to the situation that two people are talking to each other, and the others are side-participants. This regime is indicated by mutual gaze between the two, as shown in Fig. 2(b). Third, the regime called *Divergence* (also called others) corresponds to situations other than convergence and dyad-link regimes; every one is silent and/or no organized conversation exists. The gaze pattern does not exhibit any organized pattern.

This paper extends the authors' framework in [22] to infer cross-modal nonverbal interactions. Of particular note, this paper newly introduces head gestures (nod, shake, tilt) and utterances as the interacting nonverbal cues, and tries to find the causal relationship amongst them them. Fig. 1(b) provides a graphical representation of our new conversation model, consisting of hidden variables (conversation regimes $\boldsymbol{S}$, gaze patterns $\boldsymbol{X}$, and interaction structures $\boldsymbol{E}$) and observable variables (head directions $\boldsymbol{H}$, head gestures $\boldsymbol{G}$, and utterances $\boldsymbol{U}$). This model assumes that the interaction structures are probabilistically generated depending on the conversation regimes and gaze patterns and the head gestures and utterances are probabilistically generated by the interaction structures.

To establish the link between conversation regimes and the interaction structures, we hypothesize that the pattern of interactions resemble the gaze patterns as shown in Fig. 2. For example, in regime convergence, the addressees often respond to the speaker with nods as the sign of listening, sometimes accompanied with short utterances like "*hmm*" and "*yeah*". These addressing/back-channel responses are basic primitives of interactions in conversation. It is assumed that the direction of the responses follows the same pattern noted in the case of gaze, as shown in Fig. 2(a). On the other hand, there is another type of interaction, called question-and-answer; it is an interaction for exchanging messages between two persons. We assume that it often appears in the regime dyad-link, and it takes the form depicted in Fig. 2(b).

## 3.2 Model Structure

Fig. 3 provides a graphical representation of the proposed conversation model with temporal information. The upper part (A) is the same as the proposed in [22], and the lower part (B) is the novel extension made in this paper. In the upper part, we denote the sequence of regime states as $\boldsymbol{S} = \{S_1, S_2, \cdots, S_T\}$; we target the discrete temporal interval $[1, T]$. The regime state $S_t$ at time $t$ takes one of $N$ convergence regimes, ${}_NC_2$ dyad-link regimes and the divergence regime. The regime changes are considered to follow a discrete Markov process. The sequence of gaze patterns is represented as $\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_T\}$, where

the gaze pattern $\boldsymbol{X}_t$ at time $t$ is composed of the set of gaze directions of all participants, $\boldsymbol{X}_t = \{X_{i,t}\}_{i=1}^{N}$; it takes $N$ discrete directions: look at other's face or avert from all of them. The sequences of head directions is denoted as $\boldsymbol{H} = \{H_1, H_2, \cdots H_T\}_{t=1}^{T}$, and the head direction of each participants, $H_t = \{h_{i,t}\}_{i=1}^{N}$, is observed as continuous azimuth (horizontal) angle.

The lower part (B) in Fig. 3 represents the relationship between the interaction structures $\boldsymbol{E}$ and behaviors consisting of temporal intervals of head gestures $\boldsymbol{G}$ and utterances $\boldsymbol{U}$. In this paper, a head-gesture detector detects the presence/absence of head gestures at each time step, and a voice activity detector detects that of utterances at each time step (See also 5.4 and 5.3). From the detection results, the temporal intervals of continuous head gestures are extracted; here we define a temporal interval $G \in \boldsymbol{G}$ as that is bounded by the beginning/ending time step at which head motion starts/ends. The same definition can be applied to the utterance interval $U \in \boldsymbol{U}$. Here, we denote a set of gesture and utterance intervals as $\boldsymbol{B} = \boldsymbol{G} \cup \boldsymbol{U}$; hereafter we refer to *behavior* $B \in \boldsymbol{B}$ unless it is necessary to distinguish between gestures and utterances. Note this paper targets nod, shake, tilt, and treats them as the same behavior, because it mainly focuses on the temporal aspect of gestures, not the meaning of gestures.

This paper assumes that each behavior is triggered by another's behavior, or appears spontaneously. The interaction structures $\boldsymbol{E}$ determine the causal relationship among behaviors. The proposed model assumes that the interaction is probabilistically generated based on the states of regimes and gaze patterns. Fig. 4, which corresponds to the lower part of Fig. 3, visualizes the causal relationship among behaviors assigned by the interaction structures; we call it the *interaction network*. In Fig. 4, boxes indicate behaviors and an arrow from a box indicates the reaction target that triggered the behavior. A box without any outgoing arrow indicates spontaneous behavior. The interaction network can be considered as a Bayesian network with inverse arrow directions.

The interaction structures consist of a set of elements, called *action unit*, $E \in \boldsymbol{E}$, which corresponds to each behavior interval, where gesture and utterance intervals are linked if their beginning times are similar. Each action unit, $E$, has attributes including spontaneous-reactive class and a reaction target. The spontaneous-reactive class indicates that the action is spontaneous (denoted $E \rightarrow \emptyset$) or a reaction to another's behavior (denoted $E \rightarrow B, B \in Nei(E) \subset \boldsymbol{B}$), where $B$ denotes the reaction target that triggered action $E$. $Nei(E)$ denotes a set of others' behaviors that occur in the temporal vicinity of $E$, as shown in Fig. 4. This paper assumes there is only one reaction target for each action unit, at most. Here, reactive behavior is defined as the direct and immediate response to others' behavior. Spontaneous behavior is one that is not reactive behavior. Typical examples of spontaneous behaviors are addressing and questioning behavior of speakers. On the other hand, typical reactive behaviors include the hearers' back-channel responses and answers to questions posed.

## 3.3 Model Definition

Based on the conditional dependency depicted in Fig. 1(b), the joint probability distribution of the model is defined as

$$p(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{E}, \boldsymbol{H}, \boldsymbol{U}, \boldsymbol{G}, \varphi) \propto F_H(\boldsymbol{H}|\boldsymbol{X}, \varphi) \cdot F_B(\boldsymbol{U}, \boldsymbol{G}|\boldsymbol{E}, \varphi) \cdot P(\boldsymbol{E}|\boldsymbol{X}, \boldsymbol{S}, \varphi) \cdot P(\boldsymbol{X}|\boldsymbol{S}, \varphi) \cdot P(\boldsymbol{S}|\varphi) \cdot p(\varphi), \quad (1)$$

where $\varphi$ denotes the set of all model parameters. Eq.(1) is composed of the product of the likelihood functions for observed data and the prior distribution of all hidden variables. This paper employs the same definitions used in [22] for the priors of regimes $P(\boldsymbol{S}|\varphi)$ and gaze patterns $P(\boldsymbol{X}|\boldsymbol{S}, \varphi)$, and for the likelihood for head directions $F_H(\cdot)$, which assumes that head direction follows a Gaussian distribution for any given gaze direction. The prior $p(\varphi)$ of model parameters is defined as the product of that of each of the parameters; this assumes the independency of individual parameters.

In Eq. (1), $P(\boldsymbol{E}|\boldsymbol{X}, \boldsymbol{S}, \varphi)$ represents the probability that interactions $\boldsymbol{E}$ occur in given regimes $\boldsymbol{S}$ and gaze patterns $\boldsymbol{X}$. This paper decomposes this into the product of the probabilities of each action unit $E \in \boldsymbol{E}$, as written in

$$P(\boldsymbol{E}|\boldsymbol{X}, \boldsymbol{S}, \varphi) = \prod_{E \in \boldsymbol{E}} P(E|\boldsymbol{X}, \boldsymbol{S}) \cdot \prod_{B \in Nei(E)} \psi(E, B) \quad (2)$$

where the first term, $P(E|\boldsymbol{X}, \boldsymbol{S})$, represents the probability of the action unit state and the second term is a penalty to suppress the case of two behaviors responding to each other ($\psi(E, B) = 0$), otherwise $\psi(E, B) = 1$.

In Eq. (2), $P(E|\boldsymbol{X}, \boldsymbol{S}, \varphi)$ represents the probability that action unit $E$ is in response to another's behavior or is spontaneous. To define this, this paper introduces spontaneous probabilities and directional probabilities, and defines $P(E|\boldsymbol{X}, \boldsymbol{S}, \varphi)$ as their product. The former is the probability that an action is spontaneous behavior. This paper assumes that it depends on the regime and role of the person making action unit $E$, as summarized in Table 1; each probability is a hidden variable to be estimated. In Table 1, values inside the parentheses are examples obtained from manually-annotated data (C1) (See 5.1); they indicate that for a monologue, speaker's behaviors is far more spontaneous than that of the addressees. On the other hand, the directional probability represents the probability that action unit $E$ is in response to the target person; it is assumed to depend on gaze and regime. Table 2 summarizes the directional probabilities. As mentioned earlier, we assume that addressees often respond to the speaker in regime convergence, and respond to each other in dyad-link regime. Also, the reaction target tends to be a gazee, because people tend to look at the target when they respond.

The second component, which is newly defined for the joint density in Eq. (1), is the likelihood of interactive behavior $F_B(\boldsymbol{U}, \boldsymbol{G}|\boldsymbol{E}, \varphi)$ for given interaction structures $\boldsymbol{E}$ in Eq. (1). This paper defines this component as the product of the likelihood function of each behavior interval, $f_B(B|E)$,

**Table 1: Spontaneous probabilities**

| Regime | Monologue | | Dialogue | | Others |
|---|---|---|---|---|---|
| Role | speaker | addressee | dyad | others | — |
| Utterance | $\eta_{\mathrm{SMSU}}$ (0.95) | $\eta_{\mathrm{SMAU}}$ (0.06) | $\eta_{\mathrm{SDDU}}$ (0.00) | $\eta_{\mathrm{SDSU}}$ (0.00) | $\eta_{\mathrm{SOU}}$ (0.78) |
| Gesture | $\eta_{\mathrm{SMSG}}$ (0.93) | $\eta_{\mathrm{SMAG}}$ (0.05) | $\eta_{\mathrm{SDDG}}$ (0.50) | $\eta_{\mathrm{SDSG}}$ (0.00) | $\eta_{\mathrm{SOG}}$ (0.71) |

**Table 2: Directional probabilities**

| | | |
|---|---|---|
| Monologue | Addressee to speaker | $\eta_{\mathrm{DA}}$ (0.88) |
| Dialogue | One in dyad to another | $\eta_{\mathrm{DD}}$ (1.00) |
| — | Look at response target | $\eta_{\mathrm{DG}}$ (0.88) |

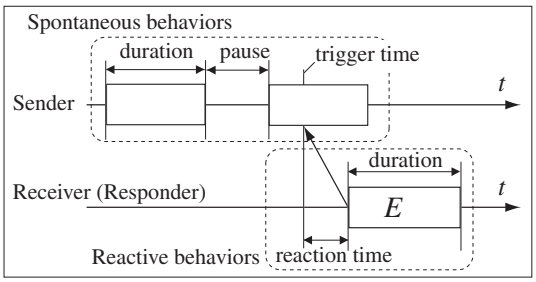


**Figure 5: Duration, pause length, and reaction time.**

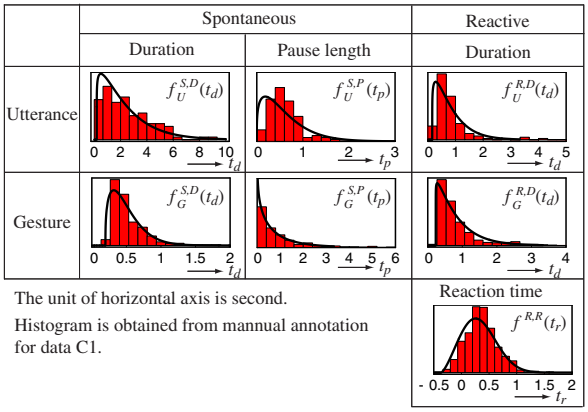


**Figure 6: Weibull distributions for duration, pause length, and response time.**

in each action unit, by assuming the conditional independency of each behavior for given interaction structures $\boldsymbol{E}$. This likelihood calculation is based on the temporal distributions of duration, pause, and reaction time of gestures and utterances (See Fig. 5). Fig. 6 summarizes Weibull models employed to represent the distributions. The model in Fig. 6 indicates the tendency observed in the timing of behaviors; e.g. spontaneous utterances are longer than reactive ones, but spontaneous gestures tend to be shorter than reactive ones. Using these models, the likelihood $f_B(B|E)$ can be defined separately for each case as in

$$f_B(B|E) = \begin{cases} f_B^{S,D}(t_d) \cdot f_B^{S,P}(t_p) & \text{if } E \text{ is spontaneous,} \\ f_B^{R,D}(t_d) \cdot f^{R,R}(t_r) & \text{if } E \text{ is reactive,} \end{cases} \tag{3}$$

where $t_d$, $t_p$, and $t_r$ denote the duration, pause length, and reaction time of behavior $B$, respectively. Note that the Weibull parameters are hidden variables to be estimated. If the target behavior is a gesture, the trigger time is set to the beginning of the interval. Otherwise, trigger time is considered to be a hidden random variable, which follows the probability distribution of response time.

# 4. ESTIMATION ALGORITHM

Based on the model defined above, the problem is to estimate the interaction $\boldsymbol{E}$, the regime $\boldsymbol{S}$, gaze pattern $\boldsymbol{X}$, and model parameters $\varphi$ from measurements $\boldsymbol{Z} = \{\boldsymbol{H}, \boldsymbol{G}, \boldsymbol{U}\}$. We employ a Bayesian approach [3] to estimate the joint posterior distribution $p(\boldsymbol{E}, \boldsymbol{S}, \boldsymbol{X}, \varphi|\boldsymbol{Z})$ of all unknown variables from the given measurements. To estimate the joint posterior, this study uses the Markov chain Monte Carlo method called the Gibbs sampler [11], which has an advantage when dealing with complex models. The Gibbs sampler repeatedly generates random samples from the full-conditional posterior distributions of each unknown variable, which constitute a Markov chain whose invariant distribution equals the desired joint posterior. The full-conditional distribution is the distribution of a variable when other variables are given. From the random samples after the Markov chain has converged, the maximum a posterior estimate is calculated for discrete variables, and the minimum mean-squared error estimates are calculated for continuous variables. Note this estimation algorithm is a form of unsupervised learning, which does not need training data to obtain the model parameters $\varphi$. Instead, we need to experientially determine hyper-parameters of the prior distributions of the parameters $\varphi$.

The full-conditional distribution of the interaction structure of an action unit can be derived from the joint distribution in Eq. (1), and is written as

$$\begin{aligned} &P(E \to B|\boldsymbol{S}, \boldsymbol{X}, \boldsymbol{E} \setminus E, \varphi, \boldsymbol{Z}) \\ &\quad \propto \textstyle\prod_{B' \in \boldsymbol{B}(E)} f_B(B'|E) \cdot P(E \to B|\boldsymbol{X}, \boldsymbol{S}) \cdot \psi(E, B), \end{aligned} \tag{4}$$

where $\boldsymbol{B}(E)$ denotes a set of behaviors included in action unit $E$. According to Eq. (4), the reaction target (also spontaneous-reactive class) $B \in Nei(E) \cup \emptyset$ of each action unit $E$ is sampled. The trigger time for each of candidate utterance intervals in neighborhood $Nei(E)$ is sampled from the reaction time distribution. The full conditional of each spontaneous action and directional probability becomes a Beta distribution when assuming Beta priors, and these probabilities are sampled from the corresponding Beta full-conditional posteriors. For Weibull models, we assume Gamma priors for the Weibull's shape and scale parameters, and truncated uniform prior for the location parameter; these priors are used to represent a priori knowledge about the timing distributions. For other variables and parameters, this paper follows the procedures described in [22].

# 5. EXPERIMENTS

## 5.1 Data

This paper targets 4-person group conversations. The participants were four women within the same age bracket; they were seated as shown in Figure 7. They were instructed to hold a discussion and try to reach a conclusion as a group for a given discussion topic within five minutes. The discussion topics were "*Should tax breaks be given to full-time housewives, or not?*" and "*Is marriage and romantic love the same or different?*"; hereafter the recorded conversations are referred C1 and C2, respectively. The head directions were measured at 30 Hz using magnetic-based sensors (POLHEMUS Fastrak$^{\text{TM}}$), which were attached to their heads on hair bands. Audio data were recorded by lapel microphones attached to each participant. Also, video sequences, whole shot (Figure 7(b)) and bust shots (Figure 9(a)), were recorded at 30 frames/sec. These data were synchronized at the unit-time step of 1/30 sec. The lengths of data were 10000 and 9100 frames (5.6 and 5.1 min) for C1 and C2, respectively.

## 5.2 Manual Annotation

The raw data was manually annotated to permit a quantitative evaluation. Annotation was mainly performed by one female in her 20s'. Fig. 8(a) shows a part of the

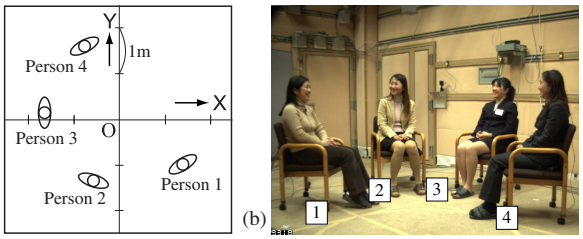


**Figure 7: Overview of scene. (a)plan view of participants' location, (b)whole view of participants.**

manual annotation ($\simeq$ 20 sec.). For each person, P1$\sim$P4, the thick bands shows the utterance intervals, manually detected based on IPU(Inter Pausal Unit) of $\geq$0.3 sec. The line segments beneath the utterances indicate gesture intervals, manually detected by visual inspection of the video. The target gestures were nodding, shaking, and tilting. Other head motions were excluded. Next, for each utterance interval and each gesture interval, spontaneous-reactive class and the reaction target were determined. Also, trigger time was given for each reactive behavior. In Fig. 8(a), small circles represent the beginning of action units and the arrows from them indicate the reaction targets, while circles with no arrow indicate spontaneous actions. The position of arrow's head indicates the trigger time step. Each vertically-elongated ellipse indicates an integrated action unit consisting of an utterance and a gesture. Also, the ground truth of gaze directions and regimes was manually created by watching the video sequences.

## 5.3 Voice Activity Detection

To automatically detect utterance intervals from the audio signals captured by lapel microphones, this paper employed a voice activity detection (VAD) method [24] that can robustly detect each person's utterance separately by clustering each person's signal in the time-frequency domain. The detected utterances output by the VAD method were reformed by filling short-term gaps to satisfy the IPU criteria, and eliminating very short intervals as noise. Fig. 8(b) shows some of the utterance intervals so detected. Compared to the manual detection results in Fig. 8(a), the automatic result includes detection lapses due to whisper-like utterances, and over-detection due to breathing, rustling, and coughing. Table 3(a) shows the accuracy of utterance detection in terms of precision, recall, and hit ratio. Here, the hit ratio is the ratio of correct frames to all frames. Table 3(a) confirms that the automatic voice detection method used was highly accurate and robust even though the amount of cross-talk was significant.

## 5.4 Gesture Recognition

The head gestures were detected with a new technique based on discrete Wavelet transform (DWT). First, DWT features are separately calculated for each head pose component; the components are azimuth (horizontal), elevation (vertical), and roll (in-plane rotation). This study applied the Daubechies wavelet of order 10 (db10) and decomposition scale was set to 2-4; windows size was 16. At each time step, we calculated the DWT coefficients of details D2-D4 and final approximation A4, and then calculated the maximum, minimum, mean, standard deviation of the wavelet coefficients in each sub-band, as the feature vector of gestures. These statistics were used in EEG signal analysis [13]. Next, we trained an SVM to classify the feature vector into two categories; gesture or non-gesture, at each time step. This paper employed a polynomial kernel of order 5 and a soft margin criterion. Training and classification was done for each person in each conversation. C1 was classified using the SVM trained with manually detected data of C2, and vice versa. The output of the SVM was then reformed in a manner similar to the utterance intervals to yield the final gesture intervals.

Fig. 8(b) shows some of the detected gestures. A comparison to Fig. 8(a) shows that there are some errors; over-detection occurs due to continuous gestures and small head movements. Table 3(b) shows the accuracy of gesture detection, and indicates that the detection was moderately successful, despite the huge dynamic range of gestures, from almost invisible ones to very large ones.

## 5.5 Experiment Setting

This paper employed the same values as used in [22], for hyper-parameters, which determine the prior distributions of gaze, regime, and head directions. The Gibbs sampling iteration was 10000, and statistics were calculated from samples obtained from the 5000th$\sim$10000th iterations. The same parameter set were used for both data $C1$ and $C2$.

## 5.6 Qualitative Evaluation

Fig. 8(b) shows a part of the interaction structures inferred from automatically detected behaviors. Fig. 9 shows three snap shots to illustrate the flow of the conversation. In this scene, speaking turn changes over time; P4$\rightarrow$P2$\rightarrow$P1. We have confirmed that gaze patterns and regimes were successfully estimated for this scene. First, P4 started to give her opinion to others who listened to P4. During this (P4's) turn, others responded to P4 with utterances and gestures. They synchronized their responses to a break point in P4's discourse. Fig. 8(b) indicates that these back-channel responses were successfully determined. Next, at the end of P4's utterances, she asked the others for agreement and tried to confirm their attitudes; their answers were correctly inferred. At the same time, P2 overlaid her utterance with the end of P4's sentence, and took over the speaking turn. P1, P3, and P4 turned their gaze to P2 and acknowledged her turn. Also, they responded to her tag-question with positive answers; their responses were successfully determined, even though the turn taking by P2 was abrupt. P1 then took advantage of a momentary chance, and took the turn. P2, P3, and P4 laughed at what P1 said; it appeared to be a humorous phrase. These responses toward P1 were successfully determined.

Fig. 8(b) indicates that most question/answer and addressees' back-channel responses toward speakers were accurately estimated, and followed the changes in speaking turn. A visual inspection of all inferred interactions confirmed that the inferred interaction structures were reasonably accurate; a few flaws were present.

## 5.7 Quantitative Evaluation

Table 4 shows the result of the quantitative evaluation of interactions. Table 4(a) shows the ratio of spontaneous-reactive class (correctly estimated) to the manual annotation. Table 4(b) shows the ratio of the number of action units whose target person was correctly inferred, to the number of all action units whose spontaneous-reactive class was

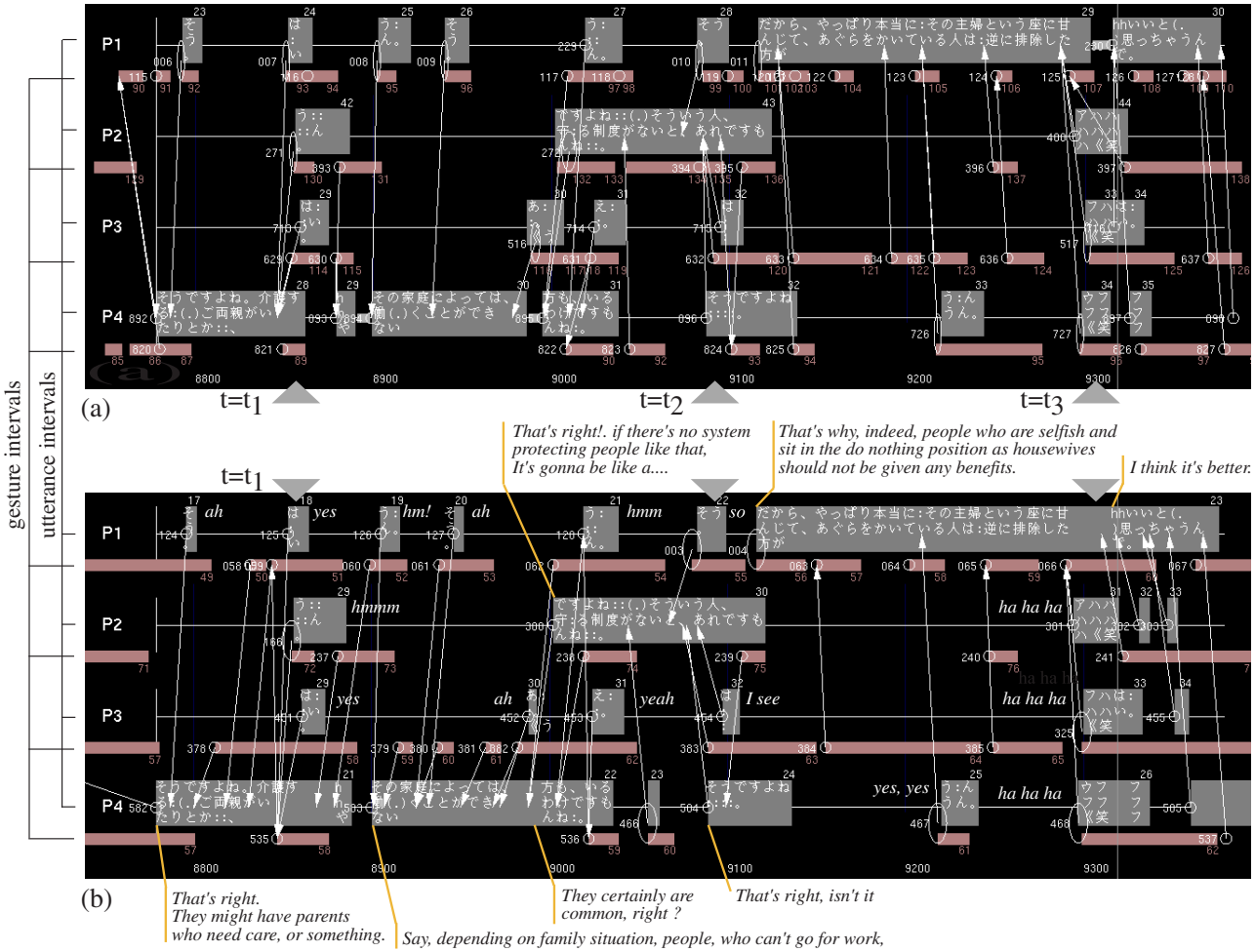


**Figure 8: Interaction network representation of interaction structures, (a)manual annotation, (b)inference result from automatically detected behaviors, (data = C1), 1 frame = 1/30 sec., Display length ≃ 20 sec.**

correctly determined. Table 4(c) shows the ratio of the number of action units in which the difference between the estimated trigger time and the one from the corresponding annotation was less than or equal to 0.3 sec.; the denominator of this ratio is the number of action units whose target person was correctly inferred. Table 4(a) shows that the accuracy of determining spontaneous-reactive class is rather modest; miss-classification happens often, especially when the action unit has a short preceding pausal length. Table 4(b) indicates that target persons were accurately inferred, and Table 4(c) suggests that the accuracy of identifying the trigger time was reasonably high. C1 yielded superior performance to C2, because C2 was a more complex conversation, with a lot of rapid turn changes. In general, the results gained from automatically detected behaviors were basically comparable to those from the manually detected ones; this verifies the effectiveness of VAD and the gesture detection technique described here.

Despite the limited dataset and annotation, the above results suggest the effectiveness of the proposed method in analyzing nonverbal interactions in multiparty conversations. Future work includes evaluations using a comprehensive dataset that includes various group and topics, as well as examining the consistency of manual annotations (used as ground truth) given by different annotators.

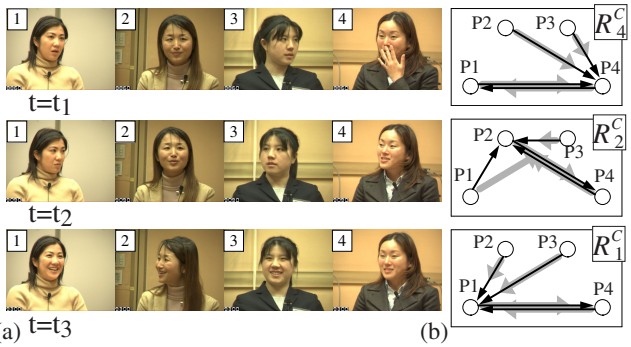


**Figure 9: Snap shots of three time steps, $t_1, t_2, t_3$ in Fig. 8. (a)each participant, (b)regime estimates and gaze patterns (solid arrows: estimates, wide arrows: ground truth). $R_i^C$ denotes P$i$'s monologue regime.**

## 6. CONCLUSION AND DISCUSSION

This paper proposed a novel target of conversation scene analysis, the automatic inference of interaction patterns from participants' nonverbal behaviors in multiparty conversations. To that purpose, a hierarchical probabilistic conversation model was introduced. Even though this paper focused on simple interactions conditioned on conversation regimes, gaze patterns, and temporal structures such as duration,

**Table 3: Detection accuracy of utterance (a) and gesture (b)**

| | (a)Utterance | | | (b)Head Gesture | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | Hit | Precision | Recall | Hit |
| C1 | 95.2 | 85.2 | 95.0 | 60.0 | 86.6 | 73.3 |
| C2 | 91.5 | 81.3 | 93.6 | 75.1 | 60.3 | 77.2 |

**Table 4: Acuracy of estimated interactions: (Manual)manual annotation, (Auto)automatically detected behaviors**

| | (a)Spont. | (b)Person | (c)Trigger |
|---|---|---|---|
| C1 (Manual) | 88.3 | 97.7 | 81.2 |
| C2 (Manual) | 77.1 | 92.2 | 67.7 |
| C1 (Auto) | 75.7 | 95.9 | 74.5 |
| C2 (Auto) | 71.1 | 91.4 | 70.1 |

pause, and reaction times, the proposed framework is considered to be noteworthy in that it provides a basic methodology for analyzing nonverbal cross-modal interactions in face-to-face conversations, and offers several prospective directions.

First, the interaction structures discovered with the proposed framework can be used as a clue for understanding the mental/context level aspects of conversations. The first step would be classifying the responses into positive and negative types, which can be indicated by head gesture classes. The proposed gesture detector is powerful enough to distinguish various head motions such as nod, shake, and tilt. The problem is to establish useful links between motion features and the inner state of the person, like the degree of agreement. Moreover, it can provide a basic element for analyzing how one person's opinion can spread throughout a human network and how a group's concordance is formed over time. It is also interesting work to relate our framework to psychological/linguistic studies such as adjacency pair analysis [25] and synchrony analysis [8].

The interaction structures can be a useful element of meeting annotation for automatic archiving/summarizing systems. For example, it could provide more semantic-based retrieval capability such as "*who had positive/negative on his opinion?*" and "*who's opinion was the most influential?*". Also, the identified interaction structures can be used to improve automatic video editing so that viewers can more clearly understand who responds to whom. Furthermore, it is worth considering a system that can quantify communication skill and provide users with feedback to improve human communication skill in organizations.

To realize real-time applications, future works include the real-time simultaneous tracking of faces from low-resolution video sequences, image-based head gesture recognition, and voice detection/separation captured with microphone arrays. Our framework can also easily incorporate other modalities such as prosody and facial expressions. We are currently developing a facial expression recognition technique that is robust against head-pose changes [17].

Finally, authors believe that this work will contribute to opening up a new research field that can explore various aspects of nonverbal cross-modal interactions in multiparty conversations, and bridge related disciplines such as psychology, social linguistics, and multimodal applications.

# 7. REFERENCES


[1] M. Argyle. *Bodily Communication – 2nd ed.* Routledge, London and New York, 1988.

[2] S. Basu. *Conversational Scene Analysis*. Ph.D thesis, Massachusetts Institute of Technology, 2002.

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

[4] M. Brand, N. Oliver, and A. Pentland. Coupled hidden Markov models for complex action recognition. In *Proc. CVPR'97*, pages 994–999, 1997.

[5] L. Chen, M. Harper, A. Franklin, T. R. Rose, I. Kimbara, Z. Huang, and F. Quek. A multimodal analysis of floor control in meetings. In *Proc. MLMI'06*, 2006.

[6] L. Chen, R. T. Rose, F. Parrill, X. Han, J. Tu, Z. Huang, M. Harper, D. M. F. Quek, R. Tuttle, and T. Huang. VACE multimodal meeting corpus. In *Proc. MLMI*, 2005.

[7] T. K. Choudhury. *Sensing and Modeling Human Networks*. Ph.D thesis, MIT, 2004.

[8] W. S. Condon and M. B. Ogston. Sound film analysis of normal and pathological behavior patterns. *J. Nervous and Disease*, 143:338–347, 1966.

[9] A. Dielmann and S. Renals. Dynamic Bayesian networks for meeting structuring. In *Proc. IEEE ICASSP'04*, 2004.

[10] D. Gatica-Perez. Analyzing group interactions in conversations: a review. In *Proc. IEEE Int. Conf. Multisensor Fusion and Integration for Intelligent Systems '06*, pages 41–46, 2006.

[11] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC, 1996.

[12] C. Goodwin. *Conversational Organization: Interaction between Speakers and Hearers*. Academic Press, 1981.

[13] İ. Güler and E. D. Übeyli. Multiclass support vector machines for EEG-signals classification. *IEEE Trans. Information Technology in Biomedicine*, 11:117–126, 2007.

[14] J. Janssen and R. Manca. *Applied Semi-Markov Processes*. Springer, 2006.

[15] A. Kapoor and R. W. Picard. A real-time head nod and shake detector. In *Proc. Workshop on PUI*, 2001.

[16] A. Kendon. Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26:22–63, 1967.

[17] S. Kumano, K. Otsuka, J. Yamato, E. Maeda, and Y. Sato. Pose-invariant facial expression recognition using variable-intensity templates. In *Proc. ACCV'07*, 2007.

[18] S. K. Maynard. Interactional functions of a nonverbal sign: Head movement in japanese dyadic casual conversation. *J. Pragmatics*, 11:589–606, 1987.

[19] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang. Automatic analysis of multimodal group actions in meetings,. *IEEE Trans. PAMI*, 27(3), 2005.

[20] L.-P. Morency, C. Sidner, C. Lee, and T. Darrell. Contextual recognition of head gestures. In *Proc. ICMI'05*, pages 18–24, 2005.

[21] D. N. P. Murthy, M. Xie, and R. Jiang. *Weibull Models*. John Wiley & Sons, Ltd., 2004.

[22] K. Otsuka, Y. Takemae, J. Yamato, and H. Murase. A probabilistic inference of multiparty-conversation structure based on Markov-switching models of gaze patterns, head directions, and utterances. In *Proc. ICMI'05*, 2005.

[23] K. Otsuka, J. Yamato, and H. Murase. Conversation scene analysis with dynamic Bayesian network based on visual head tracking. In *Proc. ICME'06*, 2006.

[24] H. Sawada, S. Araki, K. Otsuka, M. Fujimoto, and K. Ishizuka. Voice activity detection for multiple speakers with multiple pin microhpones. In *2007 Spring Meeting, Acoustical Society of Japan*, 2007.

[25] E. A. Schegloff and H. Sacks. Opening up closings. *Semiotica*, 8:289–327, 1973.

[26] R. Virginia and M. James. *Nonverbal behavior in interpersonal relations 5th Ed.* Allyn & Bacon, 2003.

[27] D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan, and G. Lathoud. Modeling individual and group actions in meetings: A two-layers HMM framework. In *Proc. 2nd. IEEE Workshop on Event Mining*, 2004.