Estimating User's Engagement from Eye-gaze Behaviors in Human-Agent Conversations

Yukiko, I. Nakano
Dept. of Computer and Information Science
Seikei University
y.nakano@st.seikei.ac.jp

Ryo Ishii
NTT Cyber Space Laboratories

ishii.ryo@lab.ntt.co.jp

ABSTRACT
In face-to-face conversations, speakers are continuously checking whether the listener is engaged in the conversation and change the conversational strategy if the listener is not fully engaged in the conversation. With the goal of building a conversational agent that can adaptively control conversations with the user, this study analyzes the user's gaze behaviors and proposes a method for estimating whether the user is engaged in the conversation based on gaze transition 3-gram patterns. First, we conduct a Wizard-of-Oz experiment to collect the user's gaze behaviors. Based on the analysis of the gaze data, we propose an engagement estimation method that detects the user's disengagement gaze patterns. The algorithm is implemented as a real-time engagement-judgment mechanism and is incorporated into a multimodal dialogue manager in a conversational agent. The agent estimates the user's conversational engagement and generates probing questions when the user is distracted from the conversation. Finally, we conduct an evaluation experiment using the proposed engagement-sensitive agent and demonstrate that the engagement estimation function improves the user's impression of the agent and the interaction with the agent. In addition, probing performed with proper timing was also found to have a positive effect on user's verbal/nonverbal behaviors in communication with the conversational agent.

Author Keywords
conversational engagement, eye-gaze, conversational agent, dialogue management.

ACM Classification Keywords
H.5.2 Information Systems: User Interfaces

General Terms
Algorithms, Design, Human Factors

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
IUI'10, February 7–10, 2010, Hong Kong, China.
Copyright 2010 ACM 978-1-60558-515-4/10/02...$10.00.

INTRODUCTION
Recent studies on virtual agents and communication robots have revealed that conversational engagement is fundamental and indispensable in communication between human users and humanoid interfaces [1, 2]. By engagement, we refer to "the process by which two (or more) participants establish, maintain and end their perceived connection", as defined in [3]. If the user is not fully engaged in the conversation, information presented by the system (agent) will not be properly conveyed to the user. Thus, in order to establish natural interactions between users and agents, displaying bodily expressions, such as facial expressions and gestures, to signal that the agent is listening to the user and perceiving nonverbal engagement signals, such as eye gaze and head nods, from the user as a listener are indispensable. If the system can monitor the user's attitude toward the conversation and detect whether the user is engaged in the conversation, then the system can adapt its behavior and communication strategy according to the user's attitude. This is critical in information providing systems, such as explanatory agents, kiosk agents, and instructor agents. If the user does not listen to the agent, the system cannot construct a reliable user model. While tailoring explanations based on the user's understanding is one of the main goals of information providing systems, few studies have considered user engagement as a basis for modeling communication with the user.

To build conversational agents that are sensitive to user engagement, there are two primary aspects to be considered. First, the system must perceive the user's nonverbal behaviors and estimate user engagement based on the sensed information. Thanks to progress in computer vision and human sensing technologies, accurate measurement of human behavior is possible in real time. For example, in an ideal circumstance, eye trackers can recognize the user's gaze points to an accuracy of 0.5 degrees at more than 100 Hz. However, few studies have investigated the interpretation of communication signals or the extraction of communication signals from an enormous number of data.

The second aspect is that the system should exploit the recognized communicative signals in dialogue management and determine the agent's behaviors properly according to the user's engagement status. For instance, if the user is not engaged in the conversation, the system needs to attract the user's attention by changing the topic of conversation. For

this purpose, we propose a dialogue management mechanism that works with an eye-tracking system and presents a prototype system of an automatic conversational agent that can estimate user engagement and determine the agent's response according to the results of the estimation.

In summary, with the goal of improving naturalness in human-agent communications, this paper proposes a method of estimating the degree of engagement by measuring the user's attentional behavior in real time and implementing a conversational agent that decides the agent's response according to the user's level of engagement. Related research is reviewed in the following section. Then, a Wizard-of-Oz experiment on data collection using an eye-tracker is described. After describing the empirical results of analyzing the gaze data, an engagement estimation method is proposed. The second half of this paper describes the implementation of a multimodal dialogue model, as well as a multimodal dialogue manager that can work with an eye-tracking system. The results of evaluating the prototype system are also reported. Finally, we discuss future research.

RELATED WORK

Eye gaze as a communicative signal in face-to-face conversation

During dialogues, two participants repeatedly alternate roles between speaker and listener. Psychological studies reported that eye gazing, specifically accompanied by head nods, serves as positive feedback to the speaker [4] and demonstrates that the listener is paying attention to the conversation. Eye gazing also contributes to smooth turn-taking [5]. Furthermore, Novick, et al. [6] reported that during conversational difficulties, mutual gaze was held longer at turn boundaries. In studies of face-to-face communication, Kendon [7] described various eye gaze functions from an ethnomethodological point of view. These results suggest that speakers distinguish different types of listener's gaze. In fact, Argyle, et al. [8] claimed that gaze is used to send positive feedback accompanied by nods, smiles, and other facial expressions, as well as to collect information from the partner.

Looking at the partner is not the only signal of engagement. When conversational participants share the same physical environment and their task requires complex reference to, and joint manipulation of physical objects, participants do not frequently pay attention to the partner, but rather look at the shared object most of the time [9, 10]. In such situations, paying attention to the shared object signals the listener's engagement in the conversation.

These findings in psychology and communication science provide the basis of the present study. However, to build conversational humanoids, it is also necessary to establish computational models and methods to implement a mechanism that can automatically interpret the user's nonverbal signals from the enormous amount of data obtained by sensing devices.

Sensing user behaviors in conversational systems

The use of head/eye trackers as a component of multimodal conversational interfaces has been successfully demonstrated in previous studies of conversational systems. Prasov and Chai [11] proposed a probabilistic model for reference resolution by combining speech and eye-gaze information. In some studies, eye-gaze information was used to estimate user interest. In Qvarfordt and Zhai [12], maps and pictures are shown on the user's desktop computer display, and based on eye-tracking data, the system estimates the focus of the user's interest and provides information about that place. Eichner, et al. [13] incorporated this user interest estimation mechanism into a presentation system with virtual agents. This system changes the conversational content presented by a pair of virtual agents according to the object of interest. As a conversational agent that communicates directly with a user, Nakano and Nishida [14] used a head-tracker to estimate user interest on a 70-inch big screen, and a tour guide conversational agent changed the topic according to the user's interest. In these systems, using the eye gaze, the user can specify interesting objects even if the name of the object is unknown.

These studies suggest that off-the-shelf head/eye tracking systems are sufficiently accurate and sufficiently stable to use as a component in complex agent systems. Thus, we believe that by combining these sensing technologies with a dialogue management mechanism, conversational agents can become sensitive to the user's level of engagement.

Interpreting nonverbal communicative signals in conversational agents

In previous studies on conversational agents, determining and generating proper communicative signals by the agents has been one of the main issues, and having the agents display nonverbal communicative signals as a listener was demonstrated to be effective in human-agent communications. Pelachaud and Bilivi [15] proposed a gaze model for generating appropriate agent gaze behaviors. Gratch, et al. [16] reported that backchannel feedback from a listener agent is effective in establishing a sense of rapport between a user and a virtual character.

However, in order to determine appropriate nonverbal signals, agents need to be able to sense and interpret signals from the user, because communication is a bilateral process between two parties. As a study focusing on the sensing aspect, Nakano, et al. [17] proposed a gaze model for nonverbal grounding in conversational agents, and, using a head tracker, implemented an agent that can judge whether the information provided from the agent is grounded. The agent also selects a direction giving strategy according to the results of grounding judgment. Morency, et al. [18] used a head tracker to recognize a user's head-nodding behavior. They reported that by combining linguistic contextual information with the results of computer vision processing, the performance of head nod detection was improved. They also applied this technology to a communication robot [19]. More recently,

Bohus and Horvitz [20] proposed a method of predicting the user's engagement intention in multiparty situations using a head tracker. Their system predicts that the user(s) will be engaged in the conversation with a reception agent if the user approaches the system from an F-formation position [21].

In these studies, the user's gaze direction was roughly estimated from the head direction measured by a head tracker. This study attempts to build an information-providing agent that demonstrates and explains products on a display as a virtual salesperson, which must accurately sense the user's attentional behavior. Thus, we use an eye (pupil) tracker to measure gaze information more accurately and estimate the user's level of engagement during the conversation with the agent.

WIZARD-OF-OZ EXPERIMENT TO COLLECT VERBAL AND NONVERBAL DATA

As the first step towards engagement estimation in human-agent communication, it is necessary to collect eye-gaze data and investigate whether eye-gaze behaviors would be useful in estimating user engagement. For this purpose, we use a Wizard-of-Oz experiment. In a Wizard-of-Oz experiment, the agent's actions are limited to the actions that the autonomous agent (to be implemented) can perform, and most of the subjects believe that they are interacting with a real autonomous agent. Therefore, the collected corpus can be used as a basis for designing human-agent interactions.

Since we are interested in information providing systems, the Wizard-of-Oz agent was designed as a salesperson in a mobile phone store in which new models are displayed and the users (subjects) want to collect useful information from the virtual sales agent.

Experiment

Experimental set-up: The experimental set-up is shown in Figure 1. Two people participated in the experiment. One of the subjects was standing in front of a sales agent displayed on a 120-inch screen and communicated with the sales agent. This subject is referred to as the "user". The distance between the sales agent and the user was 1.5 m. The other subject observed the interaction between the user and the agent through a one-way mirror. This subject is referred to as the "observer". The observer was standing 1.5 m from the user. The total number of subjects was 10, and seven of the subjects participated in the next session as an observer after being a user.

User task: Users were instructed to guess the most popular design for specific users (e.g., high school girls, businessmen). The users were promised 1,000 yen for each correct guess. Therefore, users were motivated to carefully listen to the agent's descriptions of all of the cell phones.

Experimental materials: The agent's behaviors shown on the screen were very monotonic. The agent described each of the cell phones while looking at and pointing to the cell phone being described. The agent faced the user for 3 seconds after every 10 utterances.

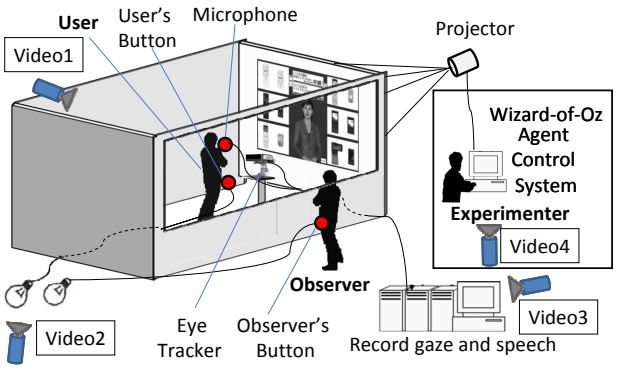


Figure 1. Experimental set-up for data collection

Verbal/nonverbal data

We collected 10 conversations, the average length of which was 16 minutes. The number of utterances by the agent was 951, and the number of utterances by the user was 61. The user's speech was recorded using a pin microphone. We videotaped the user's upper body and the video of the agent's displayed on the screen. The user's gaze data was collected using a Tobii-X50 eye tracker. The frame rate was 50 fps, and the freedom of head movement was 30 x 15 x 20 cm. The eye tracker has an accuracy of 0.5 degrees.

In addition, a push-button device was given to both the user and the observer. The user was instructed to press the button when the agent's explanation was boring and the user would like to change the topic. On the other hand, the observer was instructed to press the button when the user looked bored and distracted from the conversation. Since the button was small and completely hidden in the user's hand, the observer was not able to see whether the user was pressing the button. When these buttons were pressed, lights went on in another room, and these lights were recorded as video data[1].

ANALYSIS

Construction of gaze 3-grams:

We defined four labels to categorize the user gaze direction:

T: looking at the object that the agent is explaining. This object is referred to as the target object. Since the agent is looking at the current target object most of the time, it is presumed that joint attention is established between the user and the agent when the user's gaze label is T.
AH: looking at the agent's head
AB: looking at the agent's body
F: looking at non-target objects, such as other cell phones or advertisement posters (*F1≠F2≠F3*).

[1] We recorded the engagement judgments by video, which is the simplest method, although other methods are possible.

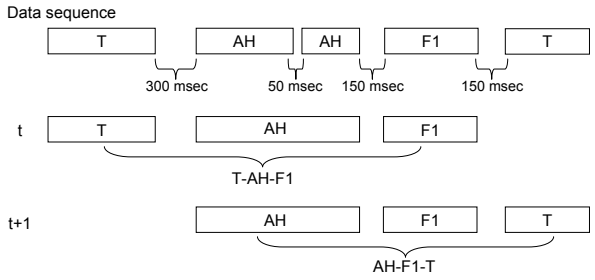


Figure 2. 3-gram construction

During blinking or completely looking down, the eye-tracker cannot measure the pupils' movements. If missing data (blank) occurs for a very short time period (less than 200 msec) and the gaze labels before and after the blank are the same, then these two consecutive gaze data were combined into one block. In contrast, if the gaze label changes after a short blank or if the blank is longer than 200 msec, then these two gaze data are not combined. For example, as shown in Figure 2, suppose that the user's gaze direction shifts as follows: *T*-(300-msec blank)-*AH*-(50-msec blank)-*AH*-(150-msec blank)-*F1*. In this case, the two *AH*s before and after the 50-msec blank are combined into one block. As a result, at time *t*, a 3-gram *T-AH-F1* is constructed from this sequence. If the blank is longer than 1 sec and a 3-gram is not complete, then we ignore the sequence as an incomplete 3-gram and start a new 3-gram at the next gaze data. In our corpus, most of the gaze data construct a 3-gram, and incomplete 3-grams are not very frequent.

Analysis of 3-grams with respect to the degree of disengagement

As part of the nonverbal data, we recorded the button pressing behaviors of the user and the observer as human judgments of disengagement. The reason for collecting reports from both parties is that using self reporting alone is not reliable and using only the observer's report is not reliable because the observer is an overhearer [22], who may not consider the user's nonverbal behaviors as signals directed toward herself/himself. Investigation of the overlap between the user's judgment and the observer's judgment revealed that the observer pressed the button 39.0% of the time when the user also pressed the button, and the user pressed the button 54.4% of the time when the observer also pressed the button. Since the agreement was not high, we decided to use the sum of the judgments, and judged that the user was disengaged if either the user or the observer pressed her/his button.

The average probability of button pressing was then calculated for each type of 3-gram and was used as the degree of disengagement (Figure 3). For example, *F1-AH-AH* co-occurs with button pressing 82% of the time. Thus, the degree of disengagement of this pattern is 82%. For the *AH-T-T* 3-gram, the degree of disengagement was only 45%. This scale indicates that 3-grams with higher degrees violate

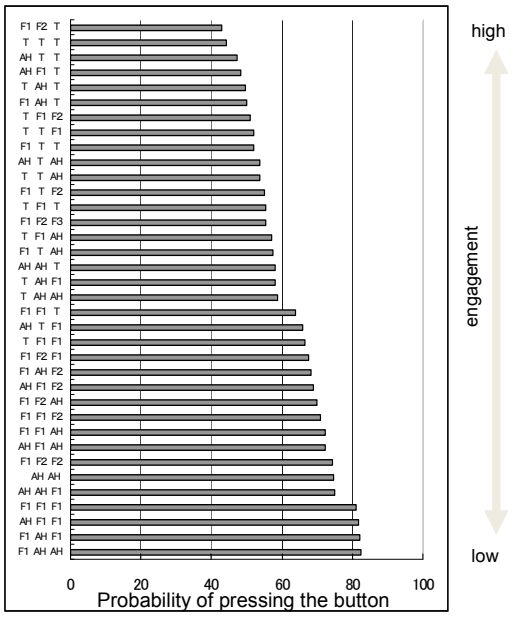


Figure 3. Probability of pressing the engagement judgment button

proper engagement gaze rules and those with lower degrees contribute to the conversation in a proper manner. Note that, as shown in Figure 3, 3-grams containing *T* have lower degrees of disengagement. This suggests that looking at the target object or establishing joint attention with the agent is a positive sign of user engagement. A further analysis of 3-gram patterns is presented in [23].

Analysis of individual difference

As the next step, we investigated the individual difference of the 3-gram distribution. Figure 4 plots the observed 3-grams with respect to degree of disengagement. The X-axis is the timeline (sec), and the Y-axis indicates the degree of disengagement. The rectangles at the top of each graph indicate the time period during which user disengagement was reported (i.e., when either the user or the observer pressed her/his button).

As shown in the graphs, for both User A and User B, 3-grams with higher disengagement values co-occur with human judgments of disengagement. Based on this observation, we set a threshold in order to estimate whether the user is engaged in the conversation.

However, comparison of User A and User B revealed that the distribution of the 3-grams differed depending on the user. In the graph for User A, 3-grams that co-occurred with human judgments of disengagement are shaded (areas (a), (b), and (c)), and the other areas indicate 3-grams that occurred during user engagement (areas (1), (2), (3), and (4)). To distinguish the disengaged areas from the engaged areas, the threshold should be set higher than the upper bound of the engaged areas and lower than the lowest upper bound among the disengaged areas. For example, for User A, the upper

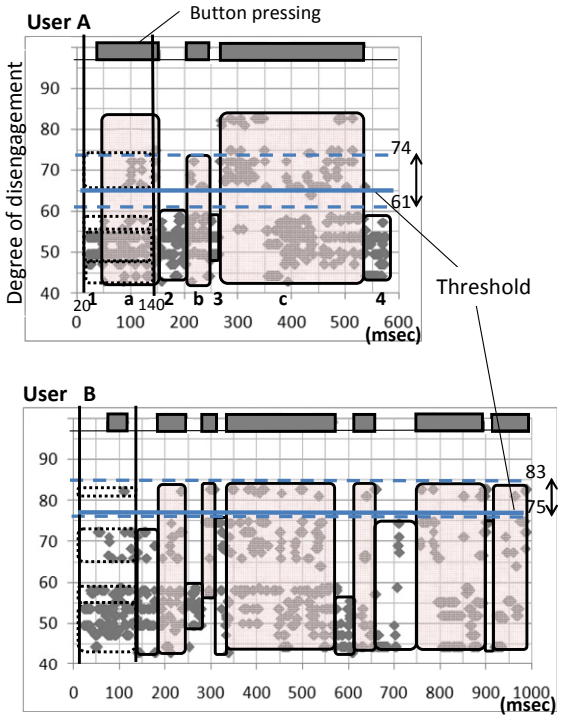


Figure 4. Individual difference of 3-gram distribution

bound of area (b) is the lowest among the shadowed areas (areas (a) through (c)), and the upper bound of area (2) is the highest among areas (1) through (4). In this case, the threshold should be set between the upper bound of area (2) (61%) and the upper bound of area (b) (74%). Likewise, for User B, the threshold should be set between 75% and 83%. The proper threshold ranges are marked with dashed lines and double-headed arrows.

Based on the above analysis, the threshold can be specified as the degree of disengagement that is assigned based on the button pressing probability and should be adapted according to the individual characteristics of the users.

ESTIMATING USER ENGAGEMENT

In order to adapt the threshold for disengagement judgment to individual users, we use a clustering technique. To determine an appropriate threshold according to the user in real time, the clustering algorithm uses the first 120 seconds of gaze data from the beginning of the explanatory conversation. Since the actual explanatory conversation starts 20 second after the greeting, the data sampling ends 140 second from the start of the interaction. In Figure 4, the 3-gram data used in determining a threshold are surrounded by dotted lines.

The data points are clustered according to the degree of disengagement. We use a simple centroid method for this purpose. The clustering procedure is as follows. First, starting with individual data points as a cluster, the Euclidean

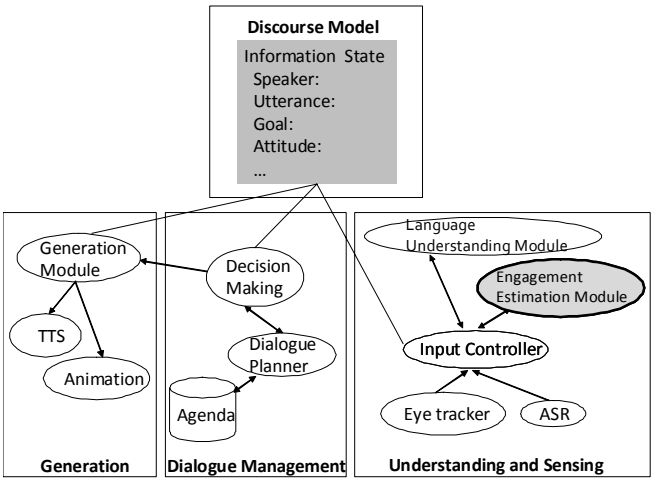


Figure 5. System architecture

distance between the centroids of two clusters is calculated, and the closest clusters are merged. The centroid of the new cluster is then calculated by weighting the centroids of the original clusters according to the number of data points. When the number of clusters becomes four, the process is terminated.

After the clustering procedure, four clusters are obtained. The midpoint between the centroid of the highest disengagement cluster and the centroid of the second highest disengagement cluster is used as the threshold. For example, using this algorithm, the threshold of User A in Figure 4 is determined to be 65, and that of User B is determined to be 76. Note that both of these thresholds fall within the proper ranges, as indicated by the dashed lines in Figure 4.

We evaluated the proposed user adapted engagement estimation method and found that the predictive accuracy is much higher when using the user-adaptive threshold than when applying the same threshold to all users. The details of this evaluation are described in [23].

ARCHITECTURE OF AN ENGAGEMENT-SENSITIVE CONVERSATIONAL AGENT

In this section, we describe the system architecture of the proposed conversational agent. The agent serves as a sales person at a cell phone store and explains about the cell phones in the store one by one. In addition to speech-based communication, this system can estimate the user's conversational engagement through attentional information. Moreover, this system uses the results of estimation in determining the agent's next action. The system architecture is shown in Figure 5. The primary components are described below.

Understanding and Sensing:

Input Controller: The Input Controller receives various types of data from multiple components and updates the state of the dialogue (using the Information State, which will be

```
< InfoState Version="0.0.0">
        <Cell ID="1" name="cVerbal">
...

        <Cell ID="2" name="cNonVerbal">
                <variable varName="cUser1" type="cell"/>
        </Cell>
        <Cell ID="3" name="cUser1">
                <variable varName="Gaze" type="cell"/>
        </Cell>
        <Cell ID="4" name="Gaze">
                <variable varName="Time" type="obj"/>
                <variable varName="CorX" type="obj"/>
                <variable varName="CorY" type="obj"/>
        </Cell>
</InfoState >
```

Figure 6. XML for customizing IS and information flow in the dialogue manager

explained later). The Input Controller receives the recognition results from input devices, such as a speech recognition system and an eye tracker, and obtains the interpretation results from language understanding and engagement estimation. The Input Controller processes these inputs using a queue for their synchronization.

Input Devices: At present, the proposed system has two input devices, namely, a speech recognition system (ASR) and an eye tracker. We use julius-4.0.2 for Windows [9] for Japanese speech recognition. We defined simple recognition rules to recognize user questions, such as questions related to the price and functions of the cell phones. The second input device is the Tobii X-120 eye tracker, which measures the user's gaze behavior. The eye tracker measures the user's gaze points at 50 Hz.

Engagement Estimation Module: We implemented the engagement estimation method proposed in the previous sections as an Engagement Estimation Module. The Engagement Estimation Module receives eye-gaze information from the eye tracker, and, based on the gaze information, this component judges whether the user is engaged in the conversation with the agent. The results of judgment are sent to the Input Controller to update the state of the dialogue.

Discourse Model

The Discourse Model maintains the state of the dialogue. We use the concept of the Information State (IS) [24] to keep track of the state of the dialogue. We modified the original IS to manipulate heterogeneous verbal and nonverbal information, such as symbolic verbal information updated at each utterance and numeric data for gaze points received 50 times per second. The details of the Multimodal Dialogue Management Mechanism are described later.

Dialogue Management

The Dialogue Planner uses a request to explain a cell phone as input and generates a plan for explaining the cell phone. All the communicative goals to be accomplished are added to the Agenda, which is implemented as a stack. The Agenda is also accessed by the Decision Making to determine the agent's next action.

Generation

Recipes of the agent's speech are synthesized by Hitachi Hit-Voice TTS and are saved as .wav files. A sequence of animation commands for each speech is saved as a script file, which is automatically generated by the CAST system [25]. Each animation script is interpreted by the Haptek animation system to generate agent animations that are synchronized with speech.

MULTIMODAL DIALOGUE MANAGEMENT MECHANISM

This section describes the details of the Multimodal Dialogue Management Mechanism (MDMM), which has two primary functions: (1) maintaining the state of the dialogue and (2) determining the agent's next action.

In the engagement-sensitive agent of the present study, the MDMM must update the state of the dialogue according to verbal and nonverbal information, which have different grain sizes. In the current system, gaze information is sent to the Input Controller 50 times per second. On the other hand, verbal information is updated upon each utterance, which is normally several seconds long.

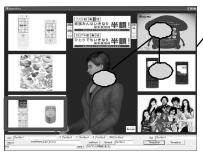
In order to integrate and maintain such heterogeneous information in the MDMM, we use an Information State-based discourse model. The Information State-based dialogue management tool has already been developed as TrindiKit [26] and Midiki [27] in Java implementation. Since these tools were designed to process verbal information in text-based or speech-based dialogue systems, some modifications are necessary in order to process a large amount of nonverbal information, i.e., numerical data sent from motion/eye/head tracking systems.

Multimodal information management: In order to synchronize different types of information that have different grain sizes, a unique time stamp is assigned to all of the information. This allows us to refer to any type of information by specifying the time. For example, by specifying the start and end times of an utterance, the system can access the gaze point information and engagement status during the utterance.

Customizing IS specification: In multimodal dialogue systems, input devices differ depending on the system. Therefore, the IS specification needs to be customizable. The MDMM provides an XML-based IS configuration file that can be edited by system developers. Figure 6 shows an example. The top level contains the "cVerbal" cell and "cNonverbal" cell. In "cNonverbal" cell, one user, "cUser1", is defined, and this cell has a sub cell, "Gaze". Then, the "Gaze" cell has three variables, namely, Time (time stamp), CorX (X coordinate of the gaze point), and CorY (Y coordinate of the gaze point). Subscription and trigger relationships are also defined in XML to specify which component subscribes which cell and which information

(15) A: I have this model, and I found that the peep preventing function is particularly useful.
<User asked a question by interrupting the agent's explanation>
(16) U: What is the price of the D904i?
(17) A: The D904i is 35,300 Yen if you start a new contract.
(18) A: Can I continue explaining?
(19) U: Yes.
(20) A: Let us resume the previous explanation.
…
(25) A: This design is both beautiful and useful. It is 17 mm-thick and fits comfortably in your hands.

<A person interrupted the conversation and the system detected the distraction>
(26) A: Would you like to move on to the next cell phone?
(27) U: Yes.
<Changing the topic after obtaining the user's agreement>
(28) A: Let us resume the explanation. At your left is the kid's model which has many safety features.



< The system detected distraction >
(29) A: Do you understand so far?
< User asked a question after prompting by the agent >
(30) U: What are the game functions like on the kid's model?
(31) A: Both the D904i and the SH904i have intuitive game functions. Can I continue the explanation?
(32) U: Yes.
(33) A: Let us resume the explanation. This phone is easy to handle…

A: Agent's utterance, U: User's utterance
Target objects are marked with squares.
User's gaze points are marked with circles.

Figure 7. Example of user-agent interaction

triggers which component. For example, when the Gaze call is updated, a message is sent to the Engagement Estimation Module, which processes the message to judge whether the user is engaged in the conversation.

Decision Making Module: The Decision Making in the MDMM decides the agent's next action by referring to the IS and the Agenda to find possible subsequent communication goals. As a special case, when the user's distraction is reported to the IS, the Decision Making Module does not choose the next goal in the Agenda, but rather adds a new communicative goal whereby the user is posed a probing question, such as "Do you have any questions?" or "Would you like to move on to the next cell phone?"

EXAMPLE

Figure 7 shows a conversation between a user and the proposed agent. During the agent's explanation of a cell phone, at (16), the user asked a question by interrupting the agent's explanation, at which point the agent responded to the user. This is a typical speech-based interaction between

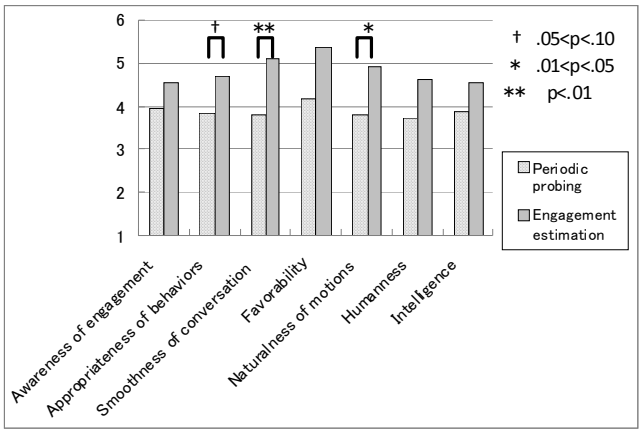


Figure 8. Rating of subjective evaluation

a user and a dialogue system. Just before an utterance at (26), another person approached the user, grabbing his attention. At this time, by processing the eye-tacking data, the engagement estimation module detected that the user was not engaged in the conversation. Then, by referring to the Information State, the MDMM discerned that the user was distracted from the conversation. Based on this information, the Decision Making Module decided to ask the probing question, "Would you like to move on to the next cell phone?" at the next turn. Since the user accepted this proposal, the system changed the topic of conversation to the next cell phone. Later, at (29), the disengagement gaze pattern was detected again. This time, the system asked another probing question, "Do you understand so far?" The system then released its turn to the user, and the user had the chance to ask a question without interrupting the agent's explanation.

EVALUATION

Experimental procedure

To examine whether the agent's capability of estimating user engagement improves the effectiveness of the system and the naturalness of the interaction with a user, we conducted an evaluation experiment. We used three female and six male subjects. None of the subjects had participated in the previous data collection experiment. The subject's task was the same as in the previous experiment, namely, listening to the agent's explanation and guessing the most popular model for female high school students or businessmen. This time, the subject did not report her/his disengagement by pressing a button (because disengagement was judged automatically by the system). A list of questions that the user can ask (regarding, for example, price, game functions, and display size) was displayed in front of the user, and the user wore a headset microphone that was used for the speech input. In the experiment, however, the user's speech was interpreted by an experimenter in order to avoid speech recognition errors that would influence the quality of the interaction. Each subject interacted with the agent under the following two conditions:

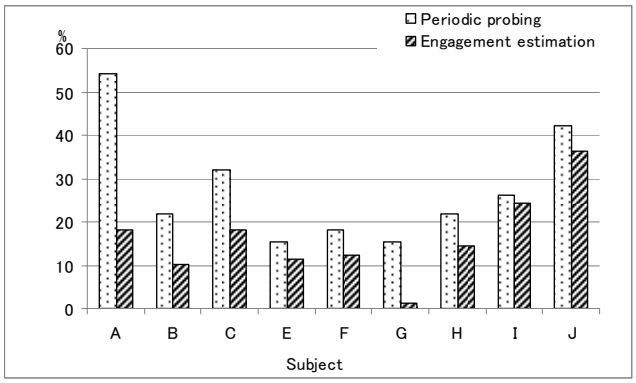


Figure 9. Frequency of disengagement gaze patterns

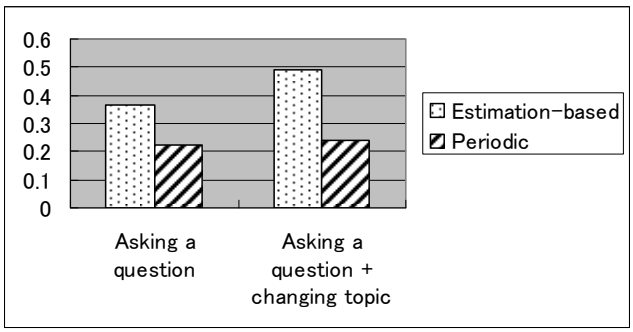


Figure 10. Frequency of user's verbal behaviors triggered by agent's probe

- Probing based on engagement estimation (engagement estimation condition): The agent generates probing questions when the Engagement Estimation Module detects the user's disengagement
- Periodic probing: The agent periodically asks probing questions (after every 10 utterances).

In the engagement estimation condition, a threshold is calculated using the data collected during the first 120 seconds of the explanatory conversation, and the threshold is determined at 140 seconds, as explained in previous sections. For each condition, three cell phones were displayed on the screen as the targets of the agent's explanations.

We used both subjective and objective (behavioral) evaluation measures. As the nonverbal objective measure, the frequency of disengagement gaze patterns was counted. The frequency of asking questions by the subjects was used as the verbal objective measure. As a subjective measure, we used a six-point Likert scale to ask the subjects about their impression towards the agent and the interaction with the agent. The questionnaire contained 33 questions, which were classified into the seven categories shown in Appendix A. Thus, four or five questions were asked for each category, and average values were used in the analysis. Since this experiment uses a within-subject design, each subject completes this questionnaire twice, once after each condition. In order to cancel the order effect, half of the subjects started with the engagement estimation condition and the other half started with the periodic probing condition.

Results

Subjective evaluation

The averages for each question category are shown in Figure 8. All of the scores were higher in the engagement estimation condition than in the periodic probing condition. Specifically, for Appropriateness of behavior and Smoothness of conversation, we found a statistical significance or trend in the two-tailed t-test ($t(8) = 1.96$; $p < 0.10$ for Appropriateness of behavior, and $t(8) = 3.90$; $p < 0.01$ for Smoothness of conversation). These results suggest that selecting the agent's behaviors according to the results of engagement estimation is effective in human-agent interaction.

Another interesting finding is that the subjects felt the agent's animated motions to be more natural in the engagement estimation condition than in the periodic probing condition ($t(8) = 2.32$; $p < 0.05$), although the animations were exactly the same. This suggests that the agent's verbal behavior presented with proper timing improve the user's impression of the agent's nonverbal expressions.

Objective evaluation of nonverbal behaviors

Figure 9 shows the percentage of disengagement gaze patterns in each session for each subject. Note that for all of the subjects (Subjects A through J), the percentage of disengagement gaze patterns decreased in the engagement estimation condition. This difference was found to be statistically significant in a two-tailed t-test ($t(8) = 3.26$; $p < 0.05$). This result suggests that agent's probing questions presented with proper timing prevent subjects from becoming distracted from the conversation.

Objective evaluation of verbal behaviors

As the second behavioral measure, we investigated the subjects' verbal behaviors. We hypothesized that if the agent were to pose probing questions with proper timing, then the subject would be more likely to ask a question or request a change of topic during her/his turn provided by the agent's question. Therefore, such behaviors are expected to be more frequently observed in the engagement estimation condition than in the periodic probing condition. Figure 10 shows the average ratios of (a) subject's asking a question and (b) asking a question and then requesting a change of topic with respect to the total number of the agent's probing questions. In the engagement estimation condition, the subjects asked questions for 37% of the time when an opportunity was presented, although in periodic probing condition, the subjects asked questions 22% of the time when an opportunity was presented. A statistical trend was observed in a two-tailed T-test ($T(8) = 2.066$, $p < 0.1$). Similarly, in engagement estimation condition, the user changed the topic of conversation 49% of the time when an opportunity was

presented, whereas, in the periodic probing condition, the user changed the topic of conversation 24% of the time when an opportunity was presented ($T(8) = 2.387$, $p < 0.05$). This difference is statistically significant. These results suggest that the agent's probes effectively provide opportunities for the subjects to talk to the agent.

Discussion

In the evaluation experiment, we found that, in the engagement estimation condition, not only that the user's impression of the agent was improved, but also that the subjects were less distracted from the conversation than in the periodic probing condition, and that the users asked more questions. These findings suggest that the proposed engagement estimation mechanism can work well in a complex conversational agent system and is useful for improving the quality of the interaction between the user and the agent.

CONCLUSION AND FUTURE WORK

By analyzing gaze patterns observed in a Wizard-of-Oz experiment, we observed that patterns of gaze transition 3-grams are strongly correlated with human subjective or observational judgment of a user's engagement in the conversation. Based on these findings, we applied a clustering technique to gaze 3-gram data and proposed a method of automatically detecting whether the user is engaged in the conversation. Then, we incorporated this mechanism into a conversational agent serving as a salesperson and conducted an evaluation experiment. By using subjective and objective measures for the evaluation, we obtained positive results suggesting that the proposed engagement-sensitive agent improved the human-agent interaction.

Although the proposed method focuses on the transitions of gaze direction, another important aspect is the duration of gaze fixation. Therefore, the proposed method might be improved by weighting each 3-gram according to its temporal duration. As such, whether the model extension contributes to improving the engagement estimation should be investigated.

Finally, we intend to address issues related to how to select the most appropriate probing question according to the user's level of disengagement. In addition to the agent asking probing questions, there may be other possibilities for reacquiring user engagement, such as asking the user's preference or telling the user to disregard other objects. More basic research is necessary in order to select an effective probe. This research may include collecting various types of probes and investigating the correlation between the conversational context and the probe.

ACKNOWLEDGMENTS

This study was supported by the Japan Society for the Promotion of Science (JSPS) through a Grant-in-Aid for Scientific Research in Priority Areas "i-explosion" (21013042).

REFERENCES

1. Sidner, C.L., et al., *Explorations in engagement for humans and robots.* Artificial Intelligence, (2005). 166(1-2): pp. 140-164.

2. Peters, C. *Direction of Attention Perception for Conversation Initiation in Virtual Environments*. in *Intelligent Virtual Agents*. (2005). p. 215-228.

3. Sidner, C.L., et al. *Where to Look: A Study of Human-Robot Engagement*. in *ACM International Conference on Intelligent User Interfaces (IUI)*. (2004). p. 78-84.

4. Argyle, M. and Cook, M., *Gaze and Mutual Gaze*. (1976), Cambridge: Cambridge University Press.

5. Duncan, S., *On the structure of speaker-auditor interaction during speaking turns.* Language in Society, (1974). 3: pp. 161-180.

6. Novick, D.G., Hansen, B., and Ward, K. *Coordinating turn-taking with gaze*. in *ICSLP-96*. (1996). Philadelphia, PA. p. 1888-1891.

7. Kendon, A., *Some Functions of Gaze Direction in Social Interaction.* Acta Psychologica, (1967). 26: pp. 22-63.

8. Argyle, M., et al., *The different functions of gaze.* Semiotica, (1973). 7: pp. 19-32.

9. Argyle, M. and Graham, J., *The Central Europe Experiment - looking at persons and looking at things.* Journal of Environmental Psychology and Nonverbal Behaviour, (1977). 1: pp. 6-16.

10. Anderson, A.H., et al., *The effects of face-to-face communication on the intelligibility of speech.* Perception and Psychophysics, (1997). 59: pp. 580-592.

11. Prasov, Z. and Chai, J.Y. *What's in a Gaze? The Role of Eye-Gaze in Reference Resolution in Multimodal Conversational Interfaces*. in *the 13th international conference on Intelligent user interfaces* (2008). p. 20-29.

12. Qvarfordt, P. and Zhai, S. *Conversing with the User Based on Eye-Gaze Patterns*. in *the Conference on Human-Factors in Computing Systems, CHI 2005*. (2005).

13. Eichner, T., et al. *Attentive Presentation Agents*. in *The 7th International Conference on Intelligent Virtual Agents (IVA)*. (2007). p. 283-295.

14. Nakano, I.Y. and Nishida, T., *Attentional Behaviors as Nonverbal Communicative Signals in Situated Interactions with Conversational Agents*, in *Engineering Approaches to Conversational Informatics*, Nishida, T., Editor. (2007), John Wiley & Sons Inc.

15. Pelachaud, C. and Bilvi, M. *Modelling Gaze Behavior for Conversational Agents*. in *IVA03 International Working Conference on Intelligent Virtual Agents*. (2003). Germany.

16. Gratch, J., et al., *Virtual Rapport*, in *6th International Conference on Intelligent Virtual Agents*. (2006), Springer: Marina del Rey, CA.

17. Nakano, Y.I., et al. *Towards a Model of Face-to-Face Grounding*. in *the 41st Annual Meeting of the Association for Computational Linguistics (ACL03)*. (2003). Sapporo, Japan. p. 553-561.

18. Morency, L.-P., Kok, I.d., and Gratch, J. *Predicting Listener Backchannels: A Probabilistic Multimodal Approach*. in *The 8th International Conference Intelligent Virtual Agents (IVA'08)*. (2008): Springer. p. 176-190.

19. Morency, L.-P., et al., *Head gestures for perceptual interfaces: The role of context in improving recognition.* Artificial Intelligence (2007). 171(8-9): pp. 568-585.

20. Bohus, D. and Horvitz, E. *Learning to Predict Engagement with a Spoken Dialog System in Open-World Settings*. in *SIGdial'09*. (2009). London, UK.

21. Kendon, A., *Spatial organization in social encounters: the F-formation system, Conducting Interaction: Patterns of behavior in focused encounters*. Studies in International Sociolinguistics, ed. Gumperz, J.J. (1990): Cambridge University Press.

22. Schober, M.F. and Clark, H.H., *Understanding by addressees and overhearers.* Cognitive Psychology, (1989). 21: pp. 211-232.

23. Ishii, R. and Nakano, Y. *Estimating User's Conversational Engagement based on Gaze Behaviors*. in *The 8th International Conference Intelligent Virtual Agents (IVA'08)*. (2008): Springer. p. 200-207.

24. Matheson, C., Poesio, M., and Traum, D. *Modelling Grounding and Discourse Obligations Using Update Rules*. in *1st Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL2000)*. (2000). p. 1-8.

25. Nakano, Y.I., et al. *Converting Text into Agent Animations: Assigning Gestures to Text*. in *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004), Companion Volume*. (2004). Boston. p. 153-156.

26. Larsson, S., et al., *TrindiKit 1.0 (Manual)*. (1999). p. http://www.ling.gu.se/projekt/trindi//.

27. *MIDIKI*. [cited; Available from: http://midiki.sourceforge.net/.

Appendix A. List of questions used in subjective evaluation

Labels	Definitions
(a) Awareness of engagement	1. Did you feel that the sales agent was aware of your attitude during her explanation? 8. Did you feel that the sales agent was aware of your
	bored with the conversation? 17. Did you feel that the sales agent catch the atmosphere? 23. Did you feel that the sales agent was aware of your gaze? 31. Did you feel that the sales agent was aware of your facial expressions?
(b) Appropriateness of behavior	2. Did the sales agent adapt her explanation according to your attitude? 9. Did you feel that the sales agent continued her description when you were not bored with the conversation and behaved appropriately when you were bored? 11. Was the timing of the agent's questions (e.g., "Do you understand so far?") proper? 18. Were the contents of the agent's questions (e.g., "Do you understand so far?") proper? 24. Was the number/frequency of the agent's questions (e.g., "Do you understand so far?") proper?
(c) Smoothness of conversation	3. Did you feel that the sales agent was easy to talk to? 12. Did you feel that the conversation with the sales agent was smooth? 25. Did you feel that the conversation was natural? 33. Was it easy to have a conversation with the sales agent? 29. Did you feel that the agent posed too many questions (e.g., "Do you understand so far?") or posed questions too frequently? 32. Did you feel that the agent posed too few questions (e.g., "Do you understand so far?") or did not pose questions frequently enough?
(d) Favorability	4. Did you have a good impression of the sales agent? 13. Did you want to talk to the agent again? 19. Did you feel that the sales agent was friendly? 30. Did you have good impressions of the sales agent's service?
(e) Naturalness of motion	5. Did you feel that the sales agent's actions were natural? 14. Did you feel that the sales agent's motions were smooth? 20. Did you feel that the gestures of the sales agent were natural? 26. Did you feel that the facial expressions of the sales agent were natural?
(f) Humanness	6. Did you feel human likeliness to the sales agent? 10. Did you feel that the sales agent is getting closer to human beings? 15. Did you feel humanity to the sales agent? 21. Did you feel that the sales agent was alive? 27. Did you feel that the sales agent had social skills?
(g) Intelligence	7. Did you feel that the sales agent was intelligent? 16. Did you feel that the sales agent was smart? 22. Did you feel that the sales agent had the ability to learn? 28. Did you feel that the sales agent had the ability to think?