Adaptive Eye Gaze Patterns in Interactions with Human and Artificial Agents

CHEN YU and PAUL SCHERMERHORN
Indiana University, Bloomington, IN, USA
AND
MATTHIAS SCHEUTZ
Tufts University, MA, USA

Efficient collaborations between interacting agents, be they humans, virtual or embodied agents, require mutual recognition of the goal, appropriate sequencing and coordination of each agent's behavior with others, and making predictions from and about the likely behavior of others. Moment-by-moment eye gaze plays an important role in such interaction and collaboration. In light of this, we used a novel experimental paradigm to systematically investigate gaze patterns in both human-human and human-agent interactions. Participants in the study were asked to interact with either another human or an embodied agent in a joint attention task. Fine-grained multimodal behavioral data were recorded including eye movement data, speech, first-person view video, which were then analyzed to discover various behavioral patterns. Those patterns show that human participants are highly sensitive to momentary multimodal behaviors generated by the social partner (either another human or an artificial agent) and they rapidly adapt their gaze behaviors accordingly. Our results from this data-driven approach provide new findings for understanding micro-behaviors in human-human communication which will be critical for the design of artificial agents that can generate human-like gaze behaviors and engage in multimodal interactions with humans.

Categories and Subject Descriptors: H.1.2 [**Information Systems**] – Models and Principles – User/Machine Systems – Human Factors; H.5.1 [**Information Interfaces and Presentations**] - Multimedia Information Systems - Evaluation/methodology; I 2.9 [**Artificial Intelligence**] - Robotics; J.4 [**Computer Applications**] - Social and Behavioral Sciences - Psychology

General Terms: Design, Experimentation, Measurement, Human Factors,

Additional Key Words and Phrases: Multimodal Interface, Gaze-Based Interaction, Human-Robot Interaction

---

AUTHORS' ADDRESSES: C. YU, 1101 EAST 10TH STREET, INDIANA UNIVERSITY, BLOOMINGTON, IN, 47401, USA. E-MAIL: CHENYU@INDIANA.EDU. M., SCHEUTZ, HALLIGAN HALL 107B, TUFTS UNIVERSITY, MA, USA. E-MAIL: MSCHEUTZ@CS.TUFTS.EDU.

1. INTRODUCTION

Interacting agents, such as communicating humans, autonomous robots performing a team task, or avatars on a computer screen interacting with human users, must coordinate their actions to achieve collaborative goals. In human communicative interactions, this coordination typically involves multiple communication channels including looking, speaking, listening, touching, feeling, and pointing, where moment-by-moment bodily actions are indicative of internal cognitive states of the interlocutors. For example, human eye gaze fixations can reveal how much of an utterance has been processed or how much of a situation has been comprehended (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). Thus, it is not surprising that recognizing and reacting appropriately to the observable behaviors of interlocutors turns out to be critical for interaction partners to initiate and carry on effective and productive interactions (Shockley, Santana, & Fowler, 2003).

While the importance of couplings between eye saccades, utterances, and bodily movements such as head turns and gestures has been firmly established for interacting humans, it is less clear whether humans will automatically transfer and apply their expectations from human interactions to interactions with artificial agents. For one, artificial agents can vary greatly in their embodiment, their appearance, and their interaction repertoire and might, as a result, not have the capabilities required for human-like interactions. Yet, to the extent that humans perceive artificial agents as "intentional" agents (e.g., because they have human-like appearance), we would expect that humans might be automatically tempted, at a subconscious level, to evaluate and respond to the agent's behaviors based on their knowledge (and experience) of other intentional agents (e.g., other humans). If this is the case, then artificial agents will have to be sensitive to observable human behaviors and react to them in very much the same way as humans do in order to enable and sustain effective interactions with humans. Specifically, artificial agents will need to (1) spontaneously perceive and correctly interpret dynamic multi-modal cues from the human partner's observable behaviors, including eye gaze, speech and bodily movements, and (2) generate meaningful, carefully timed actions in response that can be in turn perceived and interpreted by humans, often at the subconscious level (e.g., to automatically initiate eye saccades).

Toward this goal, the present paper focuses on one fundamental topic in human-human, and thus human-agent interaction: how two interacting agents dynamically allocate their visual attention when they are engaged in a collaborative task. The overall aim is to determine experimentally whether and where human attentional behaviors differ in interactions with artificial agents compared to human agents. Any systematic differences between human and artificial interaction partners will be critical for agent designers who are interested in guaranteeing smooth and sustained interactions between human users and artificial agents.

2. THE ROLE OF EYE GAZE IN HUMAN INTERACTIONS

We chose to focus on studying gaze patterns for human-agent interaction for several reasons. First, eye movements have been shown to actively gather and select visual information from the external environment for internal human cognitive systems. To achieve this goal, eye movements are also closely tied with other multimodal behaviors.

At the theoretical level of human cognition, it has been suggested that the detailed physical properties of the human body convey extremely important information (Ballard, Hayhoe, Pook, & Rao, 1997). Ballard and colleagues proposed a model of "embodied cognition" that operates at time scales of approximately one-third of a second and uses subtle orienting movements of the body during a variety of cognitive tasks as input to a computational model. At this "embodiment" level, the constraints of the body determine the nature of cognitive operations, and the body's pointing movements are used as deictic (pointing) references to bind objects in the physical environment to variables in cognitive programs of the brain. This theory of embodied cognition can be applied in the context of human-human communication. To do so, one needs to consider the role of embodiment from both the perspective of a speaker and that of a listener. First of all, recent studies show that speech and eye movement are closely linked (Tanenhaus, et al., 1995; Meyer, Sleiderink, & Levelt, 1998a; Rayner, 1998; Griffin & Bock, 2000; Griffin, 2004). When speakers were asked to describe a set of objects from a picture, they usually looked at each new object before mentioning it, and their gaze remained on the object until they were about to say the last word about it (Meyer, Sleiderink, & Levelt, 1998b). Moreover, speakers have a strong tendency to look toward objects referred to by speech and the words begin roughly a second after speakers gaze at their referents (Griffin & Bock, 2000). It has also been shown that eye, head and hand movements are well coordinated in everyday tasks, such as making tea and making peanut-butter-jelly sandwiches, while most often eyes lead the head and hands to provide a reference frame for motor planning (Pelz, Hayhoe, & Loeber, 2001). Therefore, gaze can be viewed as serving as a precursor of the following manual actions (Land, Mennie, & Rusted, 1999; Hayhoe & Ballard, 2005). All of those psychophysics and psycholinguistic results suggest that if we can reliably detect and correctly interpret moment-by-moment eye movements in certain contexts and tasks, we will be able to predict the next speech acts and manual actions even before they start.

Second, humans of all ages communicate with each other using eye gaze and produce a variety of gaze behaviors in everyday social interactions, from face-to-face eye contact, to gaze following, to gaze aversion. These eye gaze behaviors are used to signal communicative intents and to support other multimodal behaviors. The critical role human eye gaze plays in human-human interaction can be seen from the fact that even young infants can already detect and follow their parent's gaze (Butterworth, 2001) and actively gather social information to guide their inferences about word meanings, systematically checking the speaker's gaze to clarify reference (Baldwin, 1993). Moreover, gaze information has been shown very useful in collaborative tasks between human adults. In such tasks, when the linguistic signal is ambiguous, interlocutors need to draw upon additional sources of information to resolve ambiguity and achieve shared understanding (Shintel & Keysar, 2009). It has been suggested that it is the vocal and gestural acts together that comprise our everyday communication – people not only speak, but nod, smile, point, and gaze at each other. Clark and Krych (2004) demonstrated that speakers monitor addressees for understanding by using multiple cues, such as eye gaze, head nods, and head shakes, and alter their spontaneous utterances in progress when necessary. Thus, speakers monitor not just their own actions, but those of their addressees, taking both into account as they speak. Meanwhile, addressees try to keep speakers informed of their current state of understanding through signaling various non-verbal cues, such as gaze and gesture. In a referential communication task, Hanna

and Brennan (2007) explored the time course of flexibility with which speakers' eye gaze is used to disambiguate referring expression in spontaneous dialog and found that eye gaze is a powerful disambiguating cue in facilitating the interpretation of on-going speech. Brennan, Chen, Dickinson, Neider, and Zelinsky (2008) showed that shared gaze affords a highly efficient method of coordinating parallel activities when remotely located pairs of people collaborated on a visual search task.

3. RELATED WORK

Gaze has been used to build better human-agent interfaces with success (Vertegaal, 2003). For example, Mutlu, Shiwa, Kanda, Ishiguro, and Hagita (2009) investigated the role of eye gaze in a story telling robot and found that subjects were better able to recall the story when the robot looked at them more while it was telling the story. Staudte and Crocker (2009) report results from a study where humans watched a video of a robot producing statements about a visual scene in front of it. Eye-tracking data showed different patterns of human eye gaze depending on the robot's gaze and speech, and confirmed that humans are able to comprehend the robot's statements faster when the robot's gaze behavior is similar to that a human would exhibit if she uttered the same sentence. Yamazaki and colleagues (2008) performed experiments with a guide robot designed to use data from human experiments to turn its head towards the audience at important points during its presentation. The results showed that human listeners reacted better non-verbally to human-like head turns of the robot compared to non-human-like head turns. Moreover, gaze information has been used to build multimodal language learning systems showing a significant improvement in learning efficiency by using gaze data compared with not using it (Yu & Ballard, 2004; Qu & Chai, 2010).

In addition to human-robot studies, researchers in human-agent interactions (e.g. embodied conversational agents) have conducted more extensive studies on gaze behaviors accompanying speech. A spectrum of research topics has been studied, from building gaze models that can reveal the virtual agent's emotional state and engagement level, and as well as control turn-taking in multiparty conversations (e.g. Gu & Badler, 2006), to developing gaze detection algorithms to infer a user's internal state (e.g. differentiating whether a user was thinking about a response or was waiting for the agent to take its turn, e.g. Morency, Christoudias, and Darrell, 2006). The most relevant area to the present work is to empirically study and evaluate the role of gaze in human-agent interactions. Nakano, Reinstein, Stocky, and Cassell (2003) investigated the relation between verbal and non-verbal behaviors, showing that speakers looked more at their partners in order to ground references to new entities. Meanwhile, maintaining gaze on the speaker is interpreted as evidence of not-understanding which evokes an additional explanation from the speaker. Rehm and Andre (2003) studied gaze patterns in a multi-party scenario wherein users interacted with a human and a synthetic virtual agent at the same time. This setup allowed the authors to directly compare gaze behaviors in human-human interaction with gaze behaviors in human-agent interaction. They found that the addressee type (human vs. synthetic) did not have any impact on the duration of the speaker's gaze behaviors towards the addressee but meanwhile humans paid more attention to an agent that talked to them than to a human conversational partner. Kipp and Gebhaard (2008) designed an experiment with an avatar generating three gaze strategies (dominant, submissive, or following) based on real-time gaze behaviors from users in an application interview scenario. Their empirical results showed that both dominant and

submissive gaze strategies implemented through a semi-immersive human-avatar interaction system successfully convey the intended impression. More recently, Bee, Andre and Tober (2010) developed an interactive eye gaze model for embodied conversational agents in order to improve the experience of users participating in interactive storytelling. An interactive gaze model for responding to the user's current gaze in real time (i.e., looking into the virtual character's eyes or not) achieved higher user ratings than a non-interactive model. In a follow-up study (Bee, Wagner, Andre, Vogt, Charles, Pizzi & Cavazza, 2010), users' gaze data in the interaction were analyzed and the results showed users looked significantly more often to the virtual interlocutor than they usually do in human-human face-to-face interactions. Another recent study on analyzing users' gaze patterns was reported in Zhang, Fricker, Smith and Yu (2010). In their study, gaze patterns from participants who interacted with a set of virtual agents showing different engagement levels were recorded and analyzed together with user speech. The results revealed that participants were highly sensitive to the agent's behaviors and its internal engagement state, and their adaptive behaviors were not based on the overall engagement level of the virtual agent, but on momentary real-time behaviors of the virtual agent.

To summarize, gaze information is a critical component in understanding human cognitive and communication systems. Most previous studies in artificial intelligence systems found overall improvement in user evaluation or performance by incorporating gaze information in human-agent and human-robot interactions. In addition, some recent studies in embodied conversational agent collected and analyzed eye movement data from users, which led to insightful results to better understand and evaluate human-agent interactions, and to further improve the existing systems/models. The present work is built upon previous success in both human cognitive studies and human-agent interactions studies, and extends previous work in two important directions. First, there are few studies that attempt to analyze temporal dynamics of multimodal data (e.g. gaze, speech and manual actions) at a micro-behavioral level, which is a focus on the present work. Second, instead of studying embodied virtual agents in conversational tasks such as storytelling or interviewing, we designed a teaching/learning task between humans and a physical robot. Unlike face-to-face conversation studies targeted at building virtual conversational partners that can talk to people using both verbal and non-verbal behaviors, our task (as explained more in the next section) features manual actions on physical objects in a shared environment, in addition to speech and gazing behaviors from two interacting agents (e.g. humans or a robot). Because physical manipulation is a common feature of human dialogue, understanding user behaviors in this task will provide a more complete picture of human-agent interactions.

4. AN EXPERIMENTAL PARADIGM TO STUDY MULTIMODAL INTERACTION

Given that critical parts of human joint attention processes naturally occur at a subconscious level and include subtle carefully timed actions (e.g., eye gaze shifts to establish eye contact) and reactions (e.g., eye gaze shifts to objects of interest inferred from perceived eye gaze), we need an experimental paradigm that will allow an agent to interact with humans at this fine-grained level of detail. Failing to respect the subtle time course of human attentional processes will in the best case lead to prolonged, unnatural human-agent interaction; in the worst case, however, it could lead to human lack of interest and trust, frustration and anxiety, and possibly resentment of the agent.

4.1 WORD LEARNING AS AN EXPERIMENTAL TASK

In light of this, we built on our previous experimental work on joint attention in human-robot interaction (Yu, Scheutz, & Schermerhorn, 2010), that allows us to quantify, at high temporal granularity, dynamic attention allocation in real-time human-human and human-agent interactions. In our study, a human participant and an embodied agent sat around a table with multiple visual objects. Participants were asked to engage and attract the agent's attention to the target object of his own interest, and then teach the agent object names (pseudo-English words were used, e.g., "bosa"). In this face-to-face everyday attentional task, participants might switch their attention between those visual objects, and also switch to monitor the social partner's face. They might apply different attention-getting strategies to attract the agent's attention using speech, gaze, gesture and hand actions. Thus, moment-by-moment eye movements generated by participants were complex and served several roles for internal cognitive processes, from facilitating visually guided manual actions on target objects and motor planning, to generating mutual gaze with the agent, to facilitating speech production, to monitoring and using the agent's gaze to infer and then follow its visual attention. In particular, we are interested in the dynamics of how and why participants attend to what object and when, how they sustain and switch their visual attention between objects and the social partner's face, and how their gaze patterns are coordinated with their speech acts and manual actions on objects. Taken together, the experimental paradigm has the following features:

- **Multimodal**: Participants and the agent interact through speech and visual cues (including perceivable information from vision, speech, and eye gaze).
- **Interactive and adaptive**: Human participants perceive the agent's behaviors and can (and will) adjust their behavior in response to the agent's reaction.
- **Real-time**: Human participants' actions are generated in real time as their social partner switches visual attention moment by moment.
- **Naturalistic**: there are no constraints on what participants should or should not do or say in the task.

Additionally, there were two conditions in the present study – participants interacted with an artificial agent in the *agent* condition and they interacted with a human confederate in the *human* condition. Both the artificial agent and the human confederate executed the same action sequence by focusing their attention on a sequence of pre-defined locations during the interaction. This face-to-face joint attention task appears to be relatively simple but it allows us to capture several critical components in joint attention and collaborative tasks. First, we are interested in understanding multimodal micro-level behaviors in human-human interactions. For example, investigating how human users attempt to attract the attention of their social partner can be informative for the design of human-agent interfaces. Those behavioral patterns used to attract the other's attention can be directly implemented in a physical agent or a virtual agent who can therefore attract the human's attention in the same human-like way. Moreover, by studying when and in what way human participants monitor their social partner's attention, we will be able to discover the best time for a future agent to react to the human participant's action – at the exact time when participants attend to the agent's face. Thus, the present study aims at discovering those behavioral metrics and patterns from humans that can either be easily incorporated in agents' sensorimotor systems or provide general principles in the future

design of agent cognitive systems. Second, the task allows us to connect and compare human-human studies with human-agent studies. In this venue, we will report the results from human-human interaction and human-agent interaction side by side. By so doing, we will present two different kinds of results with two different goals: one is to find shared behavioral patterns that can be viewed as reliable patterns found in both human-human and human-agent interactions; the other is to find potential differences between interactions with other humans versus artificial agents, which will shed light on the guiding principles to design more human-like artificial agents.

Critically, compared with other approaches to assessing participants' experiences with artificial agents through subjective measures such as questionnaires alone, the present study specifically uses objective measures, collecting fine-grained multimodal behavioral data during the interactions, including participants' speech, visual information from a first-person perspective, momentary gaze data, and body movement data. We then employ various proven data mining and analysis techniques to discover shared behavioral patterns which can then be compared across two experimental conditions and thus allow us to determine any systematic differences between human-human and human-agent interactions, in particular, systematic differences at the micro-behavioral level that other methods are not able to detect. For example, subjects might spend more time attending to the agent with longer eye fixations in the agent condition. They also might spend more time in attracting the agent's attention before naming objects, and therefore, generate fewer naming utterances. Alternatively, they might more frequently monitor the agent's attention with shorter eye fixations and generate more naming utterances in order to attract the agent's attention through the auditory channel. Furthermore, open questions can be asked about the details of eye fixations in conjunction with naming events: will subjects look more at the agent or more at the named object? Will their eye movement patterns change over the course of naming speech production? Will they sustain and switch their attention differently when interacting with the agent or a human confederate? And if so, in what ways? These kinds of questions can only be answered by collecting and analyzing fine-grained multimodal behavioral data.

4.2 ARTIFICIAL AGENT AND HUMAN ACTOR AS SOCIAL PARTNERS

The employed artificial agent is a humanoid torso with a 2 DoF (degrees of freedom) movable head and two 3 DoF arms. The agent's head includes two 2 DoF “eyes” with integrated Firewire cameras, and two 1 DoF movable eye-brows. In the present experiment, only the head was actuated. The agent's chest also includes integrated speakers and microphones, but there was no speech output and the microphones were used only for recording (no speech recognition was performed). A single Linux PC was used for running all agent control software implemented in previous work (Scheutz, Schermerhorn, Kramer, & Anderson, 2007) and used successfully for previous studies, e.g., Brick & Scheutz (2007) and Kramer & Scheutz (2007). In the agent condition, the agent executed pre-programmed sequences of head movements followed by short pauses where the head remained stationary for a pre-programmed, randomly-generated amount of time. Thus, the agent’s actions were not responsive and completely ignored the participant's behaviors.

In the human condition, a professional male actor was hired as a human confederate that was instructed on what to do at any given point during an experimental run by the agent control program. Thus, he was asked and trained to act exactly like an agent in a tightly

controlled way. Specifically, we used the very same control architecture that guided the artificial agent's behavior for controlling the confederate's behavior by generating and sending spoken commands to the confederate through an ear-phone in place of motor commands. The commands were short one-word codes referring to specific spatial locations at where the confederate was supposed to look. The confederate performed such attention shift as soon as he heard the command. The confederate was trained with all action codes and spoken commands beforehand so that he could quickly execute those actions. In this way, the timing of when attention shifts were initiated during an experimental run was exactly the same in the human and the agent conditions. More precisely, the actor was specifically trained to perform these movements consistently and to the same location. Moreover, the actor was instructed to maintain a neutral expression and not exhibit any facial or other motion response to a human participant regardless of the participant's actions or verbalizations.

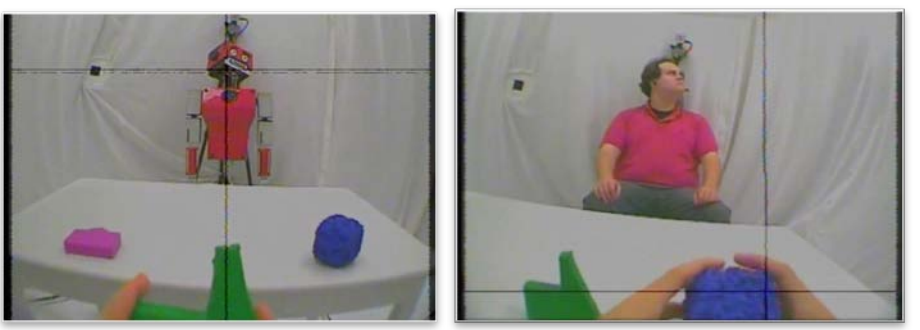

Fig. 1. Two snapshots from the participant's first-person in the two experimental conditions. The black crosshair in each snapshot indicates at where a participant gazed from his first-person view. Left: human-agent interaction -- participants were engaged with an embodied artificial agent in the word learning task. Right: human-human condition -- participants were engaged with a human confederate who was trained to follow spoken commands from the agent control system.

5. EXPERIMENT

The experimental setup is depicted in Figure 1, with a human subject sitting across a table from an agent or a human confederate. The human wears an ASL head-mounted eye-tracker with a head-mounted camera to capture the first-person view from the participant's perspective, and an overhead camera provides a bird's eye view of the scene. A box with two sets of colored objects is located on the subject's side, each set containing three objects with unique colors (e.g. blue, pink, and green). Objects are unique in each set in terms of shapes and names. The object names are artificial words (e.g. "bosa", "gasser") and they are displayed on labels that are attached to the drawers containing the objects and are visible to the participant during the experimental run.

5.1 PARTICIPANTS AND PROCEDURE

There were 48 undergraduate students at Indiana University participating in the study and 6 of them were excluded due to technical problems with their eye tracking. The rest were evenly assigned to two experimental conditions with 21 participants in each condition.

To facilitate visual processing, participants were asked to put on a white lab coat and remove any hand accessories. The eye tracker was calibrated as follows: the subject was seated in front of a calibration board with nine calibration points. The eye tracker was placed on the subject's head, and the experimenter adjusted the alignment to find the position in which the tracker could best track the subject's eyeballs. Next, the experimenter calibrated the eye tracker using the ASL eye-tracker software. When the calibration was complete, the subject was guided into the experimentation room and seated opposite the agent or the human confederate. A head-mounted microphone was attached for recording the subject's speech. The subject was shown the two sets of objects (in their drawers). Subjects were allowed to teach the agent however they wanted and were encouraged to be creative. The subject started by moving objects in the first set from the drawer to the table. The agent control architecture was started, and the experimenter instructed the subject to start. The subject was allowed one minute to teach the social partner (either the agent or the confederate) the names of the three objects in a set. After the minute was up, the experimenter notified the subject to switch to the second set. Subjects alternated between the two sets two times (4 trials in total) in approximately 4 minutes (M=4.32 min, SD=0.35). Within each trial, four head turns were generated by either the human confederate or the robot. In total, there were 20 head turn events in each interaction. In the human condition, each head turn took 1.02 seconds (SD = 0.09 seconds) on average. After a head turn, the confederate oriented his head toward a pre-defined spatial location for 9.42 seconds (SD = 0.26). Low variations in this behavior measure indicate that the human confederate was trained to execute head turn actions in a very consistent and precise way across participants and across multiple instances of head turn within each participant. This consistency is critical for us to analyze fine-grained moment-by-moment behavioral data from users. In the robot condition, each head turn took 1.92 seconds (SD=0.45) and the robot's head was fixed for 9.12 seconds (SD=0.31) before the next head turn. Taken together, the robot's and the confederate's head turn actions were consistent in both conditions and dynamic attentional states (indicated by the confederate's and the robot's head direction) were almost identical in the whole interaction. The only difference was that it took the robot a longer time (900 msec more) to switch its attention from one location to another compared with the actor. Since both the human and the robot received the same commands at the same time intervals from the same control system, this difference was caused by the execution of motor commands in the robot. In total, there were a small proportion of 18 seconds (0.9 sec x 20 instances) among 260 seconds (total interaction time) that the two agents in the two conditions were not at the same state. Many results we will report are based on accumulated statistics and therefore should be reliable even with this minor difference. For the rest derived from zooming into those head moving moments and analyzing temporal patterns (those sensitive to timing), we took this timing difference into account.

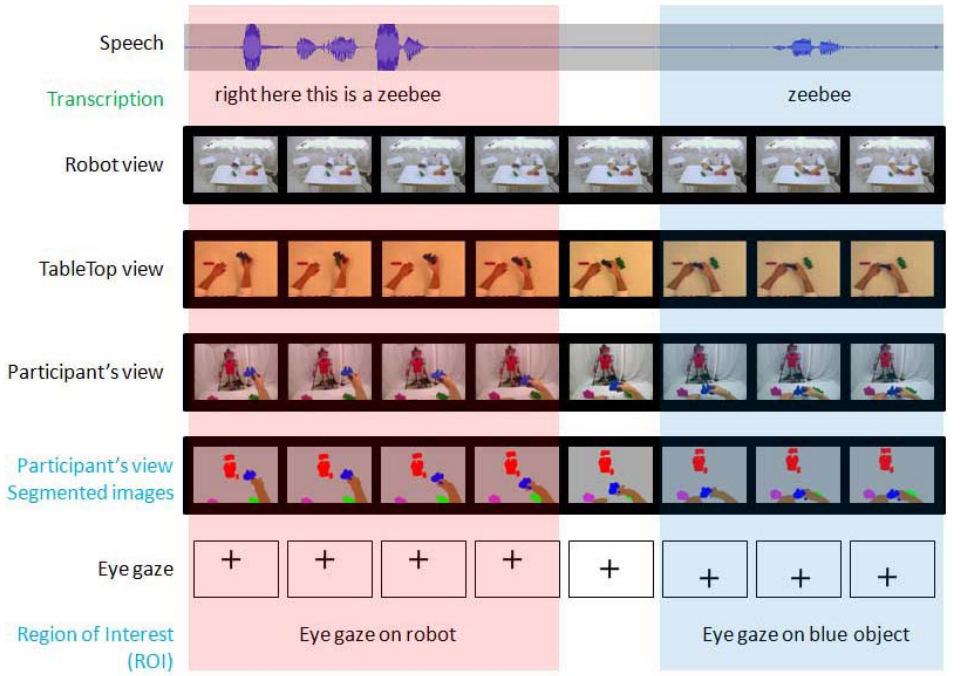

Fig. 2. In our experiment, the collected and processed multi-streaming multimodal data included speech, multiple video streams from different cameras, and eye gaze. The combination of visual processing and gaze information defined temporal Region-Of-Interest (ROI) events.

5.2 MULITIMODAL DATA COLLECTION AND PROCESSING

As shown in Figure 2, we collected multimodal streaming data, including first-person view video from the participant's perspective, first-person view video from the agent, video from a bird's eye camera on the table top, gaze data from the ASL eye tracker, and speech data, all of which will be used in our analyses. In previous studies, we have developed various visual, speech and gaze-data processing tools to automatically process and manually annotate (if needed) raw data, and to facilitate the discovery of interesting patterns from multimodal data (Yu, Smith, Hidaka, Scheutz, & Smith, 2010; Yu, Zhong, Smith, Park, & Huang, 2009). The following is a brief description on how those techniques were used in the multimodal data collected in the present study.

Visual data: Since the interaction environment was covered with white curtains and visual objects were made with unique colors, the detection of objects in view can be done easily based on color blobs. Similarly, we used red color to detect the agent in the agent condition and skin color to detect the confederate's face in the human condition. We also detected the body trunk of either the agent or the human. As shown in Figure 3, five Regions Of Interest (ROIs) were defined and calculated frame by frame: the agent, the three objects, and other body parts. In practice, we decomposed the whole video from the participant's first-person view into an image sequence and ran the ROI detection program

to compute the 5 ROIs. In addition, we manually annotate hand actions from participants, coding frame by frame when and which object they were holding in the interaction.

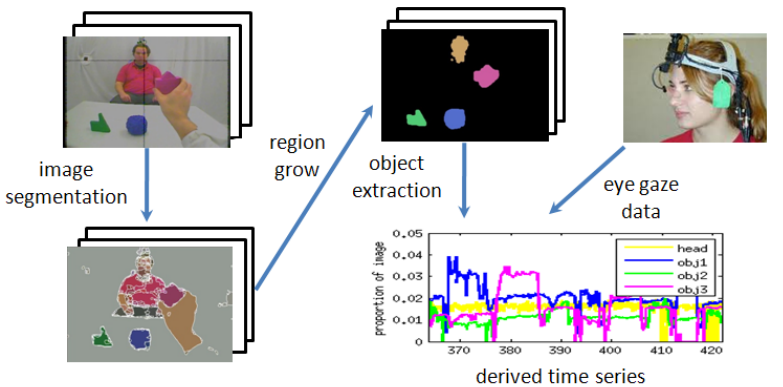

Figure 3: Areas of interests are extracted and detected based on color blobs in an image. Gaze data at a moment was superimposed on the first-person view image to detect ROIs that a participant attended to. Taken together, multiple derived time series were generated based on image processing and gaze data.

Gaze data: We computed eye fixations using a velocity-based method to convert continuous eye movement data into a set of fixation events. For each fixation, we superimposed (x,y) coordinates of eye gaze onto the image sequence of the first-person view to calculate the corresponding regions-of-interest (ROIs) moment by moment. As a result, a temporal sequence of attended regions was generated and used in the following data analyses. Since the primary goal of the present study is to understand user's attention and attention switches between different objects and the social partner's face, in the following calculations, we count gaze duration as the total amount of time on an ROI (which may consist of gazing at different sub-regions in the same overall ROI, objects or face) before switching to another ROI. That is, as far as participants' gaze data fall within the agent's face (maybe different parts of the face) before moving to another ROI, these gaze data are counted as one big fixation on the face ROI.

Speech: We implemented an endpoint detection algorithm based on speech silence to segment a speech stream into several spoken utterances, each of which may contain one or multiple spoken words. We then transcribed speech into text. The final result is a temporal stream of utterances, each of which is coded with onset and offset timestamps and a sequence of spoken words in the utterance. In addition, we manually labeled these spoken utterances containing object names.

As a result of data processing, we converted raw multiple multimodal data into multiple derived time series from which we extracted and discovered various behavioral patterns. All of the following statistics and analyses were based on the whole data set across 42 subjects (21 in each condition), containing about 302,400 image frames, 483,840 gaze data points, 16,800 eye fixations, 4,500 spoken utterances, 1,200 manual actions and 1,700 naming utterances.

6. RESULTS

We analyzed this rich, multimodal, fine-grained behavioral dataset at four different levels and aspects and the results are organized as follows:

- Accumulated statistical patterns from two types of behaviors separately: speech acts and gaze.
- Time-course analysis of the dynamics of gaze patterns when participants were naming or manually holding visual objects. To do so, we integrated speech, manual action and eye movement data, zoomed into the moments when they named or held objects and extracted dynamic time-course patterns around those events.
- Responsive gaze behaviors to either the robot or the human confederate's head turns.
- Timing patterns of multimodal behaviors to measure the synchrony of multimodal data within an agent and across two social partners.

In the following data analyses, we compare two experimental conditions and focus on extracting and interpreting both shared and different patterns in the human and robot conditions.

6.1 ACCUMULATIVE STATISTICS OF MULTIMODAL BEHAVIORS

Analyses of speech acts

We first calculated basic statistics related to participant's speech acts (see Table I). On average, participants in the human condition produced slightly more distinct word types (M_{human} = 103.54; M_{agent} = 92.50; t(40)=5.12, p<0.001, two-tailed), but a similar number of spoken utterances (M_{human} = 107.35; M_{agent} = 111.72, t(40) =1.12, p=0.267, two-tailed). As a result, participants in the human condition generated more tokens[1] (M_{human} = 444.08) than those in the agent condition (M_{agent} =400.90; t(40)=5.84, p<0.001, two-tailed), as well as longer spoken utterances (4.07 words per utterance) than the agent group (3.55 words per utterance; t(40)=3.17, p<0.005, two-tailed). In particular, participants in the human condition produced fewer one-word utterances than those in the agent condition did (M_{human} = 32.95; M_{agent} = 42.10; t(40) = 3.56, p<0.001, two-tailed). More interesting in the context of a word learning task are naming utterances: participants in the human condition produced a similar number of naming utterances compared with those in the agent group (36.01 versus 42.10), yet they produced more one-word naming utterances than those in the agent group (M_{human} = 8.50; M_{agent} =6.50; t(40)=4.53, p<0.001, two-tailed), despite having produced fewer one-word utterances overall.

Taken together, speech acts from participants in the two conditions are overall similar. However, participants in the human condition tended to use more words and more words per utterance, and more one-word naming events. One plausible explanation is that participants in the human condition attempted to use their speech to attract the confederate's attention while participants in the agent condition attempted to attract the agent's attention first before uttering object names.

[1] The number of tokens counts the number of occurrences of all distinct word types.

Table I: Overall statistics of vocabulary (**indicating statistical significance, P<0.005, t-test, two-tailed * indicating statistical significance, P<0.01, t-test, two-tailed)

	human	agent
number of words (*)	103.54	92.50
number of tokens (**)	444.08	400.90
number of utterances	107.35	111.72
words per utterance (*)	4.07	3.55
number of naming utterances	36.01	40.01
Number of one-word utterances (*)	32.95	42.10
number of one-word naming utterances (*)	8.50	6.50

Analyses of Eye Movement Data

As shown in Table II, the overall eye movement patterns from the two experimental conditions are similar in terms of the total number of eye fixations (M_{human} = 96.02; M_{agent} = 92.40; t(40) = 1.42, p<0.001, two-tailed). Moreover, human participants in both groups generated similar gaze behaviors with respect to the social partner. For example, the total number of eye fixations on the agent is comparable between the two groups (M_{human} = 26.25; M_{agent} = 29.20; t(40)=0.31, p=0.75, two-tailed). However, a closer examination of gaze data revealed differences between the two experimental conditions. Even though the total number of fixations on the social partner was similar, participants in the human group pay less attention to the human's face (M=16%) than those in the agent group (M=24%; t(40)=4.23, p<0.001, two-tailed). This partially relates to the fact that participants in the human group in general generated shorter average fixations overall (M=0.43 sec) than those in the agent condition (M = 0.52 sec).

Figure 4 shows a comparison of fixation durations in the two conditions. We found no significant difference in the human condition between average fixation durations on the agents or objects (p=0.23; t-test, two-tailed), and no significant difference in object fixation durations between the two conditions (p=0.35; t-test, two-tailed). However, the average face fixation duration in the agent condition was significantly longer than the average fixation duration on objects in the same condition, and also significantly longer than the average face fixation duration in the human condition. There are two plausible interpretations. First, it may take more time to ascertain the agent's visual attention than the other human being's attention. An alternative explanation is that participants may not have previous experiences of interacting with an agent and their curiosity with agent appearance made them look longer at the agent.

Next, we integrated gaze data with speech acts and manual actions by examining gaze patterns during those events. Figure 5 (left) shows that, during holding actions, participants in the human condition fixated on the target object more often than those participants in the agent condition, but no difference in face fixations was observed. Thus, given a time window of 7-8 seconds for each holding action, on average, participants in both conditions checked the agent's face more than twice. Figure 5 (right) compares gaze behaviors during naming events. Note that the average length of an object naming event is around 1400 msec (much shorter than the duration of an object holding event), so there are fewer fixations during naming events than during holding events. Similar to the holding event results, participants in the human condition generated more

target object fixations than participants in the agent condition. Unlike the holding events, however, agent condition participants generated more face fixations than did human condition participants.

Table II: Eye movement measures averaged across trials (approximately 1 minute per trial, * indicating statistical significance, P<0.05, t-test, two-tailed)

	human	agent
number of attention switches (fixations)	96.02	92.40
average fixation duration (seconds) (*)	0.43	0.52
number of fixations on the agent's face	26.25	29.20
number of fixations on objects	63.50	60.25
Proportion of time on the agent's face (*)	16%	24%
Proportion of time on objects	63%	56%

In sum, the analyses of eye gaze data show that 1) humans were sensitive to whom they interacted with and accordingly generated different gaze patterns; 2) they spent more time overall on the social partner's face in interactions with the artificial agent than with another human; and 3) although the overall eye gaze patterns were (more or less) similar between two groups, fine-grained data analyses exposed various differences, ranging from increased face fixations when holding or naming an object, to the duration and number of fixations.

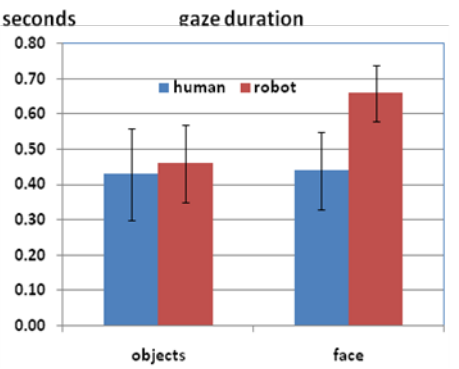

Fig. 4. A comparison of average fixation durations on either the agents' face or visual objects in two experimental conditions.

6.2 ANALYSES OF TEMPORAL DYNAMICS IN MULTIMODAL DATA

The above analyses of speech, manual action and gaze data focused on aggregated statistical patterns from each data stream. In an effort to better understand the dynamic processes that lead up to moment-by-moment events as participants attempted to direct the social partner's attention, we next present a time-course analysis of gaze data. In particular, we are primarily interested in gaze dynamics relative to three events: when participants were holding an object, when they were naming an object, and when the agent was switching its attention.

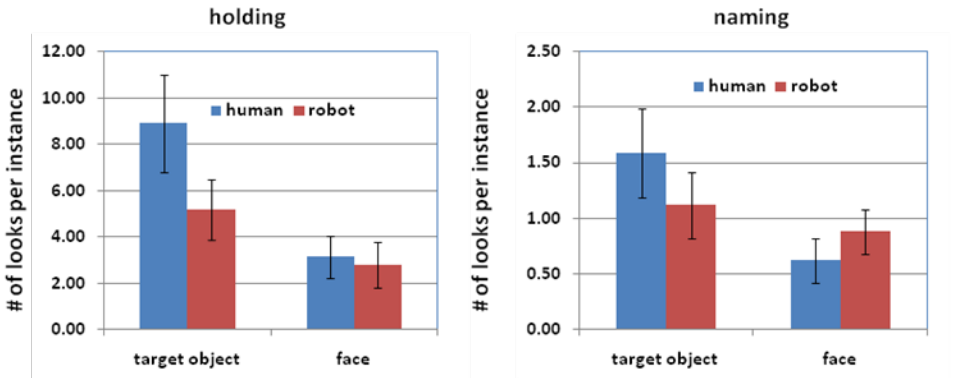

Fig. 5. The number of looks on the target object or the agent's face when naming or holding a visual object.

Speech actions and gaze patterns

We focused first on naming utterances and the moments immediately surrounding them, using participants' visual attention to integrate speech data with gaze data. This approach is based on methods used in psycholinguistic studies to capture temporal profiles across a related class of events (Allopenna, Magnuson, & Tanenhaus, 1998). Such profiles allow us to discern potentially important temporal moments within a trajectory and compare temporal trends across trajectories. In the current study, the relevant class of events is a naming utterance; we aligned all the instances of naming events based on the onset and offset of each instance. Figure 6 shows the average proportion of time across all naming events (which can be viewed as a probability profile) that participants looked at the agent's face, the target object mentioned in the naming utterance, the other two objects, or other body parts. Thus, each trajectory in a plot shows the probability that participants looked at the associated type during the 5 seconds immediately prior to naming events (the left two plots), during naming events (the middle two plots; a fixed temporal window of 2 sec was used since name utterances had different lengths), or during the 5 seconds immediately after naming events (the right two plots). We observe several interesting patterns: 1) Overall, participants in both conditions spent a majority of the time during naming events looking at the named objects and the social partner's face, and therefore only a small amount of time looking at other objects or body parts (the two middle panels). 2) Participants in the human condition began fixating on the named object before the onset of a naming utterance. This is a well-documented behavior pattern in speech production, as fixating on the target object facilitates the planning of speech acts. This result is also consistent with the findings of a previous study (Griffin & Bock, 2000) showing that in face-to-face interaction, speakers are likely to look at the target object when they produced an object name. 3) However, participants in the agent condition spent a large proportion of the time gazing at the agent both before, during and after naming events. Our results thus show that maintaining and monitoring joint attention in face-to-face human-agent interaction significantly changes participants' visual attention: they looked at the agent more frequently than expected in face-to-face human-human referential communication tasks (Griffin & Bock, 200), and they did so even during naming events. 4) Participants in both conditions shifted attention to the agent's face immediately following the onset of a naming utterance, as demonstrated by an increase of

the face profile (the green line in the two middle panels). 5) Participants' visual attention on the target object decreased after a naming event.

In sum, during the moments surrounding a naming event, human condition participants paid more attention to the named object than to any of the other categories. This particular gaze pattern is in line with results from psycholinguistic studies on the coupling between speech and gaze. However, this pattern was not replicated in the agent condition before and during naming, suggesting that participants perceived the agent's random behaviors and attempted to adjust their own behaviors. By doing so, they failed to demonstrate typical behaviors that are well-documented in psycholinguistic studies on speech and simultaneous eye gaze.

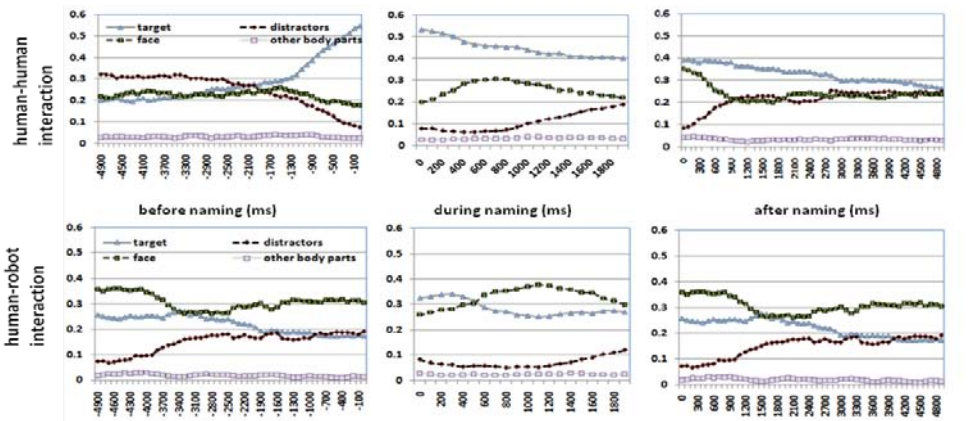

Fig. 6. Gaze probabilistic profiles before/during/after naming an object in human-human and human-agent conditions representing the proportion of time that participants spent looking at the agent, the named target object, the other two objects, and the agent's other body parts, before (left), during (middle), and after (right) a naming utterance. The top three plots represent naming events by participants in the human condition, while the bottom three are derived from those of participants in the agent condition.

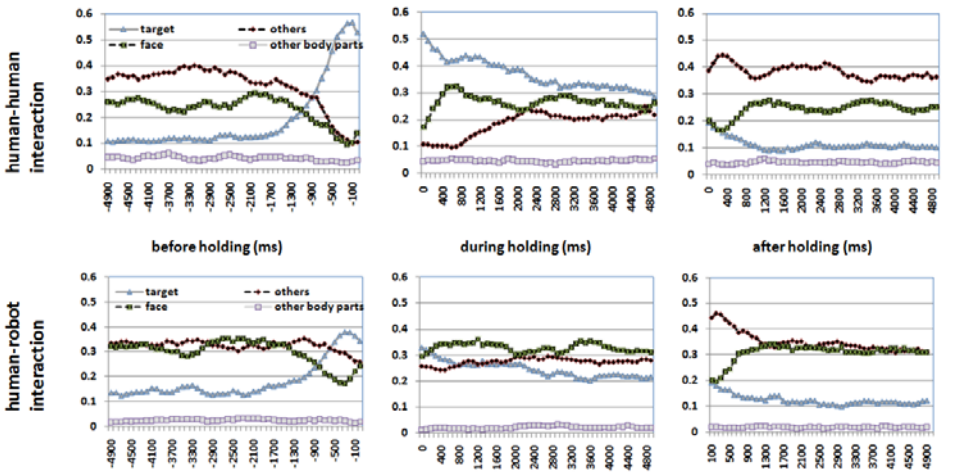

Fig. 7. Gaze probabilistic profiles before/during/after holding an object in human-human and human-agent conditions representing the proportion of time that participants spent looking at the agent, the held target object, the other two objects, and the agent's other body parts, before (left), during (middle), and after (right) a holding action. The top three plots represent holding events by participants in the human condition, while the bottom three are derived from those of participants in the agent condition.

Manual actions and gaze patterns

Eye-head-hand coordination is well-documented in naturalistic everyday tasks, such as making tea, and making peanut-butter-jelly sandwiches (Pelz, et al., 2001). We next investigate the extent to which either group exhibits standard coordination behaviors, as well as how and whether coordinated patterns may differ between the two conditions. Figure 7 shows the time course profiles before, during and after a holding event, with both similarities and differences evident in gaze patterns between the two conditions: 1) Participants in both conditions were more likely to look at the target (i.e., to-be-held) object just before grasping it. The stabilization of gaze on the target object may facilitate the ongoing reaching action by providing a reference frame in motor planning (Hayhoe, Shrivastava, Mruczek, & Pelz, 2003). However, those in the human condition paid much more attention to it than those in the agent condition. 2) Immediately after fixating on the target and even before grasping it, there was an increase of attention to the social partner's face (the green lines in the two left panels), suggesting that participants checked the agent's face to monitor its attention and potential reactions to the reaching and holding actions. 3) During holding, participants in the human condition were more likely to look at the target object than at the other objects, while agent condition participants were no more likely to look at the target object than at the others. 4) During holding and manipulating, participants in both conditions regularly monitored the agent's face, with those in the agent condition looking at the face slightly more often (35%) than those in the human condition (27%).

More generally, as shown in Figure 7, participants in the human condition demonstrated more natural gaze behaviors – they fixated on the target object before holding it,

maintained their attention on that object during holding and switched their attention to other objects after holding. In addition, immediately prior to the holding action, their attention began to shift to the face (as indicated by an increase in face fixation probability), presumably to monitor the agent's attention. In contrast, participants in the human-agent condition showed quite different gaze patterns. They paid less attention to the target object prior to grasping it, showed an approximately equal amount of attention on the target object, the agent's face, and other objects during holding, and were more likely to fixate on the agent's face after holding. In short, participants in the agent condition were less engaged with the target object and more concerned about the agent's attention.

6.3 RESPONSIVE ACTIONS TO THE AGENT'S HEAD TURNS

The results reported thus far are based on aligning the instances of two participant-originated event types – naming and holding – without regard to the agent's status. Next, we directly examine participants' responses to the agent's actions. One critical design aspect of the present study is that both the human agent and the robot agent executed the same pre-defined action sequence. Thus, regardless of the participants' efforts, the agents always generated the same action at the exactly same moment in the two conditions, which allows us to directly compare, moment-by-moment, users' responsive actions in the two conditions. For example, sometimes the agent's attention shifts happened to coincide with participants' naming or holding actions. Would participants be sensitive to the agent's head turns and attention switches at those moments?

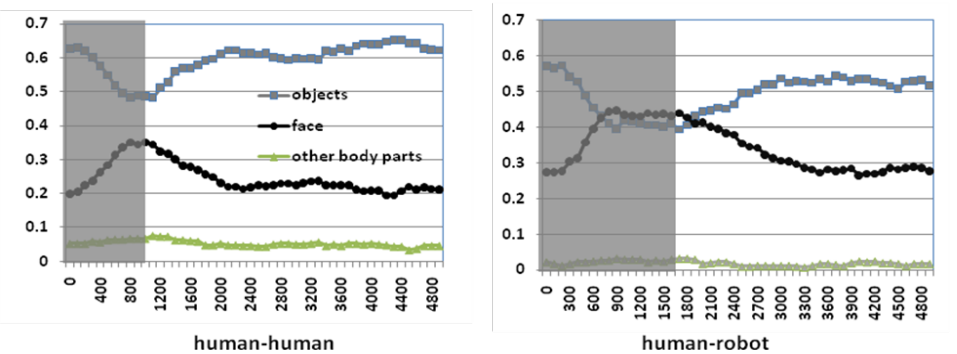

Fig. 8. Gaze probabilistic profiles during/after the agent's head turns. In both figures, 0 in the x-axis indicates the onset of a head turn, and the shaded areas indicate the span of the agent's head turn. Note that in the human-agent condition, the agent took a longer time to generate the same head turn compared with what the human agent did. Nonetheless, the results showed that in both conditions, participants quickly switched their attention to the agent's face soon after the onset of the head turn, and then back to the target object right after the offset of the head turn.

Figure 8 shows the temporal profiles of gaze patterns relative to agent-generated attention shifts (head movements from one pre-defined spatial orientation to another). As noted above, participants in both conditions focused more on the objects than on the agent's face (63% of the time in the human condition, 56% in the agent condition). However, they quickly switched their attention to the agent's face right after the onset of the agent's

head turn, suggesting that they used their fovea to attend to the objects (probably for motor planning and visually guided actions) while simultaneously using peripheral vision to monitor the agent's attention. As a result, they detected the agent's head turn in real time and immediately switch their attention to the agent's face. In addition, because the agent's slower head movement led to longer attention shift durations, participants in that condition demonstrated more sustained attention on the agent's face before returning back to the objects. In both conditions, when the agent's head stopped moving (and therefore participants knew to where the agent finally attended), participants quickly switched their attention away from the agent's face and back to visual objects.

In summary, participants in both conditions continuously monitored the agent's attention. And they were sensitive to the agent's head actions. Their attentional strategy was to use their focal vision on visual objects and their peripheral vision to monitor the agent's status while the agent's head was stationary. This allowed them to immediately switch attention to the agent's face whenever its head started moving. This attention tracking strategy was unexpected -- our original hypothesis was that participants would frequently switch their attention between objects and the agent's face in order to monitor the agent's attention. Compared with the peripheral vision strategy actually applied by participants, such an explicit monitoring approach would be less effective, as the agent's attention shifts would go unnoticed until the next "checking" event. This result is a good example of how this data-driven approach, based on fine-grained multimodal data, will not only quantify and connect with previously-observed phenomena, but can also lead to discoveries that were not expected in advance – the true value to advance our knowledge on human-human and human-agent communications.

6.4 TIMING OF FACE LOOKING

The above results based on time-course analyses provide a dynamic picture of participants' gaze behaviors over time in the course of interaction. One strength of this approach is that it can identify general temporal trends by aggregating behaviors across multiple instances. However, this approach cannot capture the exact timing and synchrony between multiple behaviors and events. Our next step is to zoom in on individual instances and examine the synchronized gaze patterns elicited by salient events.

We examine the exact timing of two multimodal behaviors/events, drawn from observations made in the time-course analyses above: 1) that participants were likely to look at the agent's face in response to agent attention shifts, and 2) that participants were likely to look at the agent's face immediately after initiating a naming utterance. In each case, we calculated, for each instance, the timing between the onset of the event (either a head turn or a naming utterance) and the first fixation on the agent's face following that event. We report here the statistical patterns observed in these synchrony measures.

Figure 9 summarizes the observed gaze behaviors in response to the agent's head turns. At the onset of head turns, participants had already fixated on the agent's face 20% of the time in the human condition and 27% of the time in the agent condition. These, of course, cannot be treated as responsive actions. However, in cases where participants were *not* already looking at the agent's face, agent attention shifts led to a gaze fixation on the face within a window of 2 seconds after the onset of a head turn 42% of the time in the human

condition, compared to 60% in the agent condition. Taking as a baseline the overall likelihood of face fixation (16% in the human condition and 24% in the agent condition in Table II), there was a clearly substantial increase in fixations on the agent's face driven by the response to the agent's head turns. Further, we calculated the timing delay between the onset of the head turn and the onset of the first gaze fixation on the agent's face (among those instances with face fixation within 2 seconds), and found most fixations happened at around 800-900 milliseconds, as shown in Figure 7. Considering that it takes adults approximately 200-300 milliseconds to plan and execute a gaze shift and that there was also a perceptual processing time to detect the head turn action after it initiated, it is clear that participants in both conditions rapidly detected the head turn event and as a response, quickly switched their attention to the agent's face in real time. In addition, although participants in the agent condition were more likely to shift attention to the face, the timing between head turn and the first face gaze was similar in both conditions, suggesting that the timing discovered here is reliable.

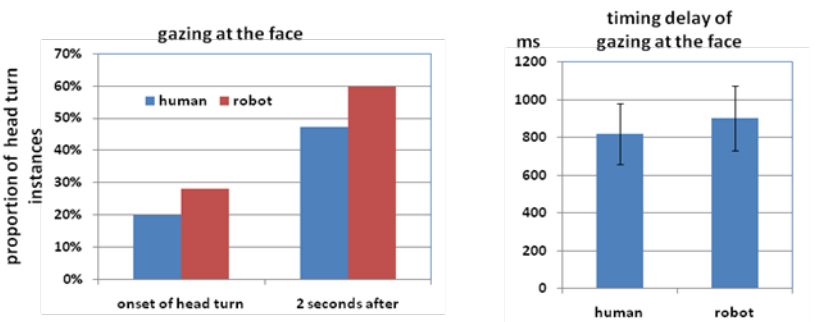

Fig. 9. The participants' responsive actions to the agent's head turns. Left: the proportion of time that participants were already gazing at the agent's face at the onset of head turn in both conditions. Also shown is the proportion of instances that participants focused on other ROIs (Regions of Interest) at the onset of the head turn but shifted gaze to the face within a window of 2 seconds. The results show that those in the agent condition were more likely to respond to the agent's head turn than those in the human condition. Right: the exact timing delay between the onset of head turn and the onset of the first fixation on the face. The timing delay was similar in both conditions: 800-900 milliseconds.

Next, we measured the synchrony between the onset of speech naming acts and the onset of the first follow-up look to the agent's face. We already know that participants in both conditions were likely to check the agent's face after they named the object. However, what is unknown is how this naming act was synchronized with their agent face fixations. Figure 10 shows the results. First, participants in the agent condition were more likely to gaze at the agent's face at the onset of a naming event (26%) than those in the human condition (20%). In both cases, the results were similar to the baselines reported above (16% in the human condition and 24% in the agent condition). However, within a window of 2 seconds after a naming event onset, the likelihood of face fixation increased by more than 10% in both conditions. More interestingly, the calculation of the timing between the onset of a naming event and the onset of the first look yields a reliable timing metric of 600-700 milliseconds. This suggests that right after the naming event

started, a significant proportion of participants immediately switched their attention to the agent's face to check whether the agent was paying attention to the named object.

We note that our gaze data are from user's naturalistic behaviors without any instructions/constraints to bias participants on where they are supposed to look and at what moments. Therefore, they were free to shift their gaze moment by moment. In this context, it is clearly not the case that all the participants generated exactly the same behaviors every time an event happened. Thus, they might or might not look at the agent face after naming and even if they do so, they might generate the first face fixation at different timing moments. Considering those potential variations from naturalistic gaze behaviors in both human-human and human-agent interactions, the exact timings extracted here are informatively reliable.

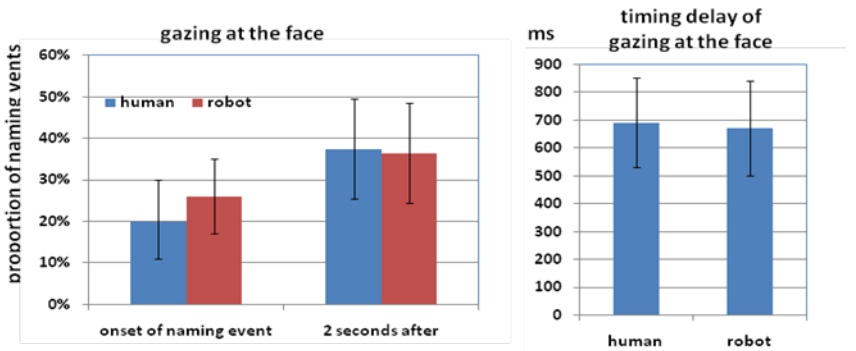

Fig. 10. The participants' gaze patterns after object naming. Left: the proportion of naming instances in which participants in both conditions were already attending to the agent's face at the moment of the onset of a naming utterance and within a window of 2 seconds. Right: the exact timing between the onset of a naming event and the first gaze on the agent's face. The timing delay is 600-700 milliseconds after the onset of the first spoken word in a naming utterance.

7. GENERAL DISCUSSION AND FUTURE WORK

The present study set out to investigate micro-level behaviors of humans in interactions with human and artificial agents. The starting point was human-human interaction, for which there is a wealth of empirical evidence of what may be considered "normal" responses and behaviors in various laboratory tasks. The experimental paradigm presented here examined multimodal couplings of key sensorimotor behaviors during face-to-face interactions in a word learning task, in particular, how gaze changes are related to and affected by one's own speech acts and manual actions, and the actions of the social partner. In addition, the comparative experimental design explored similarities and differences in human micro-level behaviors during interactions with another human versus an artificial agent. Uncovering differences at this micro-level is important, as it allows us to establish patterns of differences that can inform future agent and human-agent interaction designs. The data-driven approach described above yielded a varied spectrum of results. However, one consistent theme emerged: participants in the artificial agent group paid significantly more attention to the agent's face than those in the human group, who instead tended to fixate more on the target (named or grasped) object. This could be due to any number of factors (e.g., the novelty of the agent, more difficulty

determining its gaze target, the difference in head motion durations; we discuss this in more detail in the following subsection), but whatever the reason, it is an important phenomenon of artificial agent interactions and it is critical to firmly establish it before any attempts can be made to explain it or to mitigate its effects.

Additionally, the data-driven approach also allowed for the identification of interesting consistencies in the two experimental conditions that are informative to agent and agent interaction designers. In particular, we discovered the exact timing of participants' eye gaze at their social partner's face during the interactions. Participants in both groups exhibited very similar face looking patterns after the onset of agent head movements (800-900 ms) and after the onset of naming acts (600-700 ms). Regularities such as these should be exploitable by agent and agent interaction designers, as they can be directly incorporated in real-time human-agent interfaces to enhance efforts both to predict human social peers' actions by using gaze as a precursor, and to identify the exact timings for the agent to execute head turn and gaze shift actions when social partners are gazing at the agent's face. In addition, the results from the present study can also be used to discriminate human-human and human-agent interactions when both interactions happen in the same environment with multiple human users and artificial agents working together in a collaborative task (e.g. Katzenmaier, Stiefelhagen, & Schultz, 2004).

Each of the results summarized here, as well as the many others described in Section 5, could be useful when designing artificial agents that need to interact with humans. Whether they are relevant in any particular context will, of course, be up to the designer to decide. The important point here is that these similarities and differences need to be exposed. And the data-driven, micro-level approach pursued in this paper represents the first steps toward this direction.

High-Level and micro-level user data

The primary goal of most human-agent interaction studies is to develop new computational models to generate more human-like behaviors. These behaviors are then evaluated by asking users to answer questions on post-experimental surveys after they had interacted with the intelligent agent. For the purpose of evaluating human-agent interfaces and models, this methodology is sufficient as the ultimate goal of building intelligent systems/interfaces is to enable human-like interaction so that users can enjoy interacting with artificial agents (as much as they enjoy interacting with other human beings, at any rate). Thus, survey data obtained contain direct feedback from users that can be used to compare and evaluate different interfaces and models.

Table III: Summary of primary results (without temporal dynamics results)

behavioral type	descriptions of measurement	results (H: stands for the human condition; R: stands for the robot condition)
speech	# of word types	H: more distinct words
	# of spoken utterances	no difference
	# of tokens	H: more tokens/words
	# of one-word utterances	H: more one-word utterances
	# of naming utterances	no difference
	# of one-word naming utterances	H: more one-word naming utterances
eye gaze	# of eye fixations	no difference
	# of eye fixations on the social partner	no difference
	prop. of time on face	R: more looking
	fixation duration on face	R: longer
	fixation duration on objects	no difference
	fixating on the target object when holding it	H: more looking
	fixating on the target object while naming it	H: more looking
response to the social partner's head turns	prob. of gazing at the face at the moments of head turns	R&H: 20-25%
	prob. of gazing at the face right after head turns	R&H: 50-60%
	timing latency of gazing at the face after head turn onsets	R&H: 800-900ms
multimodal synchrony between speech acts and face looking	prob. of gazing at the face at naming onsets	R&H: 20-25%
	prob. of gazing at the face right after naming onsets	R&H: ~35%
	timing latency of gazing at the face after the naming onset	R&H: 600-700ms

Besides user evaluation, another important reason to acquire user's feedback from interacting with current models or interfaces is to collect empirical evidence that can be used to guide us on designing better ones in the future. While survey-based measures are

ideal for user evaluation, they are rather high-level and subjective for providing insights for better designs. For example, users may report that they find artificial agents “eerie" (MacDorman & Ishiguro, 2006) which makes them not want to interact with agents again; or they may report that they find an agent “friendly” and that they would like to interact with more in the future. While such self-reports serve the general purpose of user evaluation, they provide few details that could be used to precisely “pin down” where the agent failed to meet expectations, and in what ways designers can improve their systems and models based on current results. This is because, even though users may have high-level feelings and impressions of how well they interacted with an agent, they cannot recall details about their subconscious reaction to the agent, nor might they be able to remember important subconscious effects and micro-behavioral differences in real-time interactions. For example, typically participants are not consciously aware of or able to recall these micro-level behaviors (e.g., where their gaze fell during the past two minutes, how long they looked at a certain location, how many fixation events they generated, etc.). This is because users' reactions to an agent typically happen at both a conscious and subconscious level and only parts of the conscious level (which is the basis for verbal reports on surveys) are accessible to introspection.

Table III summarizes a set of results from the current study. These results are informative for two reasons. First, they represent a degree of micro-level detail which cannot be obtained by asking participants survey questions. We argue here that high-level feelings of the agent and interaction are influenced by and reflected in these micro-level behaviors. Second, these results can be directly incorporated in future agent designs. If we want to understand how micro-behavioral events that are often being processed only at the subconscious level can cause such overall evaluations, we need to focus on events such as head movements, gaze shifts, speech prosody changes, and even postural adjustments (e.g., leaning fractionally toward or away from the social peer). Each of these subtle external cues reflects the internal states of ongoing cognitive processes, whether interacting with an artificial agent or another human. Yet, those micro-level behaviors may be key to understanding what makes some interactions feel more natural than others. Therefore, incorporating these micro-level behavioral patterns into future designs has a promise to lead to better human-agent interactions.

Connecting micro-behaviors with high-level psychological constructs

While the current study focused on analyzing micro-level data, our results can also be connected with high-level psychological concepts. For example, the difference in how much participants looked at the human confederate or the agent may relate to intimacy regulation (e.g., Argyle & Dean, 1965, Argyle & Ingham, 1972). Research on this topic suggests that people avoid gaze to control the level of intimacy between themselves and their partners. This explanation is consistent with low levels of gaze between strangers and high levels of gaze with inanimate characters (e.g., people do not avoid gaze when they watch characters on TV). In the current study, participants might be regulating intimacy with the human confederate more but less with the robot. In addition, the difference may also be caused by uncertainty in the agent’s capabilities to understand/learn names, and uncertainty in the robot’s social skills (maintaining joint attention). A third high-level factor is the unfamiliarity of participants with interacting with a robot. As a result, they may look at the robot

more due to their curiosity. Even though we didn't find any difference in their face looking behaviors at the beginning vs. the end of the interaction, this novelty effect may still play a role given the overall interaction is relatively short (<30 minutes including instructions, eye-tracking calibration and breaks between trials). As further studies are required to obtain a clear answer, connecting micro-level patterns with high-level constructs and theories can be an important component in this data-driven approach. First, high-level theories may allow us to make predictions of micro-behavioral patterns which can be used to guide pattern discovery in data analysis. More importantly, micro-level patterns may reveal how high-level behaviors, e.g. intimacy regulation, are implemented through moment-by-moment multimodal behaviors. In this way, bottom-up behavioral patterns and top-down theoretical knowledge can work together to lead to a better understanding of human-agent interaction and better agent designs.

There are two factors at the sensorimotor level that may be relevant to the results reported in this paper. First, there might be differences in how easy it is for participants to read the human's gaze and the robot's gaze, even though they observe from the same distance in the current experimental setup. Second, as reported earlier, differences in the duration of motion exhibited by the human confederate and the robot during attention shifts may not only cause different responsive behaviors at those moments (only a small proportion) but also have a carryover effect through the whole interaction. Both factors are worth further study in future work. One way to answer these questions is to use embodied virtual agent platforms. With the advance of computer graphics technologies, visual details of virtual characters can be rendered with high resolution, making them much more human-like. Along this line, research on android science has successfully built human-like robots with artificial skins (e.g. MacDorman & Ishiguro, 2006). However, the cost and effort of this embodied approach are prohibitive; using virtual agents is effective at relatively low cost and in a short development time. In addition, the actions executed by virtual agents are not constrained by motor and mechanical delays in a physical robot, and therefore they can be precisely controlled and programmed to match with the timing of natural human action. One noticeable difference between a physical robot and an embodied virtual agent is the physical co-locatability – a robot can share the physical environment with human users while virtual agents need to be rendered in a display and cannot directly execute an action to change the physical environment that users live in. Recently, Hasegawa, Cassell and Araki (2010) conducted a comparison study in an interaction task in which either a physical robot, a virtual agent or a GPS device gave directions to human participants. They found no difference in direction-giving performance between the robot and the embodied conversational agent, but did find behavioral differences (e.g., participants interacting with the virtual agent used more complementary gestures than those interacting with the physical robot). With interactive systems available in different forms, comparisons of different systems in the same experimental setting, such as the present work and the study reported in Hasegawa, et. al (2010), can generate new empirical evidence and insights that cannot be obtained from studying a single system alone.

Experimental control vs. free interaction

In the current experimental paradigm, the behaviors of the robot and the human confederate were pre-determined. We chose this controlled design to allow us to compare, in a rigorous way, how human users may respond differently when they interact with two types of agents with the same action sequences. That is, any differences or consistencies found are due to differences between the two agents (the visual appearance and the participants' mental models of the agents). This controlled experimental design is typical of most empirical studies. Meanwhile the interaction itself was free-flowing without any constraints on the user end in terms of what they should do or say. An alternative approach is to study human-agent and human-human interactions in completely unconstrained naturalistic contexts. For example, Gergle & Clark (2011) used a naturalistic paradigm in which two participants were asked to freely interact with each other in a conversation elicitation task. The naturalistic aspect of interaction studies is critical because the ultimate goal of human-agent interactions is to design an agent/robot that can interact with users in an unconstrained environment with naturalistic behaviors. From an ecological perspective, there is always a need to confirm whether results and findings from well-controlled experiments can be easily extended and generalized to more naturalistic situations. Compared with well-controlled experimental paradigms, a challenge in naturalistic studies is to analyze rather complex data (generated from free interactions) and reach reliable conclusions by discovering interesting patterns and factors (and meanwhile excluding confounding factors). In brief, these two experimental paradigms are complementary and ultimately converging evidence is expected to be collected from both paradigms. In addition, we can build a closed loop – results from naturalistic studies can be used to guide the design of experimental controls, while the results from well-controlled studies can provide insights for naturalistic studies on which patterns and factors should be focused on in order to generate reliable results while maintaining the naturalistic nature of interaction. In this way, these two lines of research can inform and bootstrap each other.

Future work

The employed experimental paradigm (from data collection, to data processing, to data analysis) can be easily extended to other collaborative human-human and human-agent interaction tasks, allowing further investigations to both refine and generalize our present findings in other contexts. Along this line, we are also interested in comparing virtual agents rendered by computer graphics with physical robots, while continuing to use the human confederate as a baseline. This will allow us to determine how factors such as 2D vs. 3D, physical embodiment, and rendering quality may create differences in user responses at the micro-behavioral level. Another important direction for future work is to use our experimental findings to guide the development of an artificial gaze control system. For example, we can predict based on our data what object a participant will name based on their eye gaze patterns in the 2 to 3 seconds immediately preceding the naming event. Similarly, we can predict which object a participant is likely to reach for even before the grasping action starts. Hence, an artificial agent with access to real-time human eye gaze data could be programmed to change its focus of attention to that object as soon as the human intention is discovered. This could result in better coordination between humans and artificial agents, or reduced human cognitive loads. Moreover, by studying in detail the temporal coupling of eye gaze and naming, we will be able to design behavioral scripts that will allow artificial agents to assume the role of a teacher

that is intuitive and easy to follow for human learners. Incorporating those findings into a gaze model and implementing that within virtual and physical agents will allow us to explicitly test the real-time mechanism. In turn, we will acquire additional empirical results from those studies which will guide us in improving the agent model. In this way, we can establish a positive feedback loop between knowledge discovery from empirical studies and improved model building. Overall, we believe that the kind of empirically-grounded and sensorimotor-based study of human-agent interactions exemplified in this paper will ultimately allow us to systematically investigate important aspects of human social interactions that can form the foundation for developing truly natural human-agent interactions.

References

Allopenna, P., Magnuson, J., & Tanenhaus, M. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*(4), 419-439.

Argyle, M., & Dean, J. (1965). Eye contact, distance and affiliation. Sociometry, 1965, 28, 289-304

Argyle, M., & Ingham, J. (1972). Gaze, mutual gaze, and proximity. Semiotica, 1972, 6, 32-49.

Baldwin, D. (1993). Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental Psychology, 29*(5), 832-843.

Ballard, D., Hayhoe, M., Pook, P., & Rao, R. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences, 20*(04), 723-742.

Bee, N., Wagner, J., Andre, E., Vogt, T., Charles, F., Pizzi, D., Cavazza, M. (2010). Discovering eye gaze behavior during human-agent conversation in an interactive storytelling application. *In Proceeding of International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*.

Bee, N., Wagner, J., Andre, E., Vogt, T., Charles, F., Pizzi, D., Cavazza, M. (2010). Multimodal interaction with a virtual character in interactive storytelling. *In Proceedings of* AAMAS 2010, 1535-1536.

Breazeal, C., & Scassellati, B. (2000). Infant-like social interactions between a robot and a human caregiver. *Adaptive Behavior, 8*(1), 49.

Brennan, S., Chen, X., Dickinson, C., Neider, M., & Zelinsky, G. (2008). Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition, 106*(3), 1465-1477.

Brick, T., & Scheutz, M. (2007). *Incremental natural language processing for HRI*.

Butterworth, G. (2001). Joint visual attention in infancy. *Blackwell handbook of infant development*, 213–240.

Clark, H., & Krych, M. (2004). Speaking while monitoring addressees for understanding* 1. *Journal of Memory and Language, 50*(1), 62-81.

Gergle, D., & Clark, A.T. (2011). See what I'm saying? Using dyadic mobile eye tracking to study collaborative reference. *Proceedings of CSCW 2011*, 435-444

Griffin, Z. (2004). Why look? Reasons for eye movements related to language production. *The interface of language, vision, and action: Eye movements and the visual world*, 213–247.

Griffin, Z., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science, 11*(4), 274.

Gu, E., Badler, N.I. (2006). Visual Attention and Eye Gaze During Multiparty Conversations with Distractions. *In Proceedings of IVA'2006*, 193~204

Hanna, J., & Brennan, S. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language, 57*(4), 596-615.

Hasegawa, D., Cassell, J., Araki, K. (2010) "The Role of Embodiment and Perspective in Direction-Giving Systems" in Proceedings of AAAI Fall Workshop on Dialog with Robots.

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences, 9*(4), 188-194.

Katzenmaier, K., Stiefelhagen, R., & Schultz, T. (2004). Identifying the addressee in human-human-robot interactions based on head pose and speech. *In Proceedings of ACM 6th International Conference on Multimodal Interfaces*, 144-151.

Kipp, M., & Gebhard, P. (2008). IGaze: Studying reactive gaze behavior in semi-immersive human-avatar interactions, *In Proc. of IVA2008*, 191-199.

Kramer, J., & Scheutz, M. (2007). *ADE: A framework for robust complex robotic architectures*.

Land, M., Mennie, N., & Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *PERCEPTION-LONDON-*, *28*(11), 1311-1328.

MacDorman, K., & Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies*, *7*(3), 297-337.

Meyer, A., Sleiderink, A., & Levelt, W. (1998a). Viewing and naming objects: Eye movements during noun phrase production. *Cognition*, *66*(2), B25-B33.

Meyer, A., Sleiderink, A., & Levelt, W. (1998b). Viewing and naming objects: Eye movements during noun phrase production. *Cognition*, *66*(2), 25-33.

Morency, L., Christoudias, C., & Darrell, T. (2006). *Recognizing gaze aversion gestures in embodied conversational discourse*.

Mutlu, B., Shiwa, T., Kanda, T., Ishiguro, H., & Hagita, N. (2009). *Footing in human-robot conversations: how robots might shape participant roles using gaze cues*.

Nakano, Y., Reinstein, G., Stocky, T., & Cassell, J. (2003). Towards a model of face-to-face grounding. *In Proceedings of the 41st meeting of the Association for Computational Linguistics*, 553–561.

Pelz, J., Hayhoe, M., & Loeber, R. (2001). The coordination of eye, head, and hand movements in a natural task. *Experimental Brain Research*, *139*(3), 266-277.

Qu, S., & Chai, J. (2010). Context-based word acquisition for situated dialogue in a virtual world. *Journal of Artificial Intelligence Research*, *37*(1), 247-278.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, *124*, 372-422.

Rehm, M., & Andre, E. (2005). Where do they look? Gaze behaviors of multiple users interacting with an embodied conversational agent. *In Proc. of IVA'05*.

Scheutz, M., Schermerhorn, P., Kramer, J., & Anderson, D. (2007). First steps toward natural human-like HRI. *Autonomous Robots*, *22*(4), 411-423.

Shintel, H., & Keysar, B. (2009). Less is more: A minimalist account of joint action in communication. *Topics in Cognitive Science*, *1*(2), 260-273.

Shockley, K., Santana, M., & Fowler, C. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology*, *29*(2), 326-332.

Staudte, M., & Crocker, M. (2009). *Visual attention in spoken human-robot interaction*.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632-1634.

Trafton, J., Cassimatis, N., Bugajska, M., Brock, D., Mintz, F., & Schultz, A. (2005). Enabling effective human-robot interaction using perspective-taking in robots. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, *35*(4), 460-470.

Vertegaal, R. (2003). Attentive user interfaces. *Communications of the ACM*, *46*(3), 31-33.

Yamazaki, A., Yamazaki, K., Kuno, Y., Burdelski, M., Kawashima, M., & Kuzuoka, H. (2008). *Precision timing in human-robot interaction: coordination of head movement and utterance*.

Yu, C., & Ballard, D. (2004). A multimodal learning interface for grounding spoken language in sensory perceptions. *ACM Transactions on Applied Perception (TAP), 1*(1), 57-80.

Yu, C., Scheutz, M., & Schermerhorn, P. *Investigating multimodal real-time patterns of joint attention in an hri word learning task.*

Yu, C., Smith, L., Shen, H., Pereira, A., & Smith, T. (2009). Active Information Selection: Visual Attention Through the Hands. *IEEE Transactions on Autonomous Mental Development, 2*, 141–151.

Yu, C., Smith, T., Hidaka, S., Scheutz, M., & Smith, L. A Data-Driven Paradigm to Understand Multimodal Communication in Human-Human and Human-Robot Interaction. *Advances in Intelligent Data Analysis IX*, 232-244.

Yu, C., Zhong, Y., Smith, T., Park, I., & Huang, W. (2009). Visual data mining of multimedia data for social and behavioral studies. *Information Visualization, 8*(1), 56-70.

Zhang, H., Fricker, D., Smith, T., & Yu, C. (2010). Real-time adaptive behaviors in multimodal human-avatar interactions. In Proceedings of International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction.