See You See Me: The Role of Eye Contact in Multimodal Human-Robot Interaction

TIAN (LINGER) XU, Indiana University at Bloomington
HUI ZHANG, University of Louisville
CHEN YU, Indiana University at Bloomington

We focus on a fundamental looking behavior in human-robot interactions—gazing at each other's face. Eye contact and mutual gaze between two social partners are critical in smooth human-human interactions. Therefore, investigating at what moments and in what ways a robot should look at a human user's face as a response to the human's gaze behavior is an important topic. Toward this goal, we developed a gaze-contingent human-robot interaction system, which relied on momentary gaze behaviors from a human user to control an interacting robot in real time. Using this system, we conducted an experiment in which human participants interacted with the robot in a joint-attention task. In the experiment, we systematically manipulated the robot's gaze toward the human partner's face in real time and then analyzed the human's gaze behavior as a response to the robot's gaze behavior. We found that more face looks from the robot led to more look-backs (to the robot's face) from human participants, and consequently, created more mutual gaze and eye contact between the two. Moreover, participants demonstrated more coordinated and synchronized multimodal behaviors between speech and gaze when more eye contact was successfully established and maintained.

CCS Concepts: • **Human-centered computing** → **Empirical studies in interaction design**; *Laboratory experiments*; • **Applied computing** → *Psychology*

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

ACM Reference Format:
Tian (Linger) Xu, Hui Zhang, and Chen Yu. 2016. See you see me: The role of eye contact in multimodal human-robot interaction. ACM Trans. Interact. Intell. Syst. 6, 1, Article 2 (May 2016), 22 pages.
DOI: http://dx.doi.org/10.1145/2882970

1. INTRODUCTION

In human-human social communication, it is well established that social understanding is facilitated by paying attention to other people and subtle social cues they generate in real time [Kendon 1967]. Gaze, or looking, is of central importance in social behaviors due to the two special roles it plays in maintaining everyday interactions [Kendon and Cook 1969; Mundy and Newell 2007; Senju and Johnson 2009]. First, eye gaze serves

The reviewing of this article was managed by special issue associate editors Yukiko Nakano, Roman Bednarik, Hung-Hsuan Huang, and Kristiina Jokinen.
This work is supported by NSF BCS 0924248 and NIH R01 HD074601.
Authors' addresses: T. Xu, Department of Computer Science, Indiana University Bloomington, Bloomington, Indiana, 47405; email: txu@indiana.edu; H. Zhang, CECS Department, University of Louisville, Louisville, Kentucky, 40292; email: h0zhan22@louisville.edu; C. Yu, Psychological and Brain Science, and School of Informatics, Indiana University Bloomington, Bloomington, Indiana, 47405; email: chenyu@indiana.edu.

DOI: http://dx.doi.org/10.1145/2882970

as a non-verbal signal [Argyle 1988]. In particular, the act and manner of gazing at a social partner has meanings as a signal, showing the amount of interest to another person. Second, gaze is also a channel to perceive the expressions signaled by others. For instance, to follow the other person's visual attention, one needs to look at the person's eyes and use eye direction to infer where the person is gazing [Frischen et al. 2007]. Taken together, gaze serves as both a signal and a channel: as a signal used by the gazer to pass communicative information, and as a channel for the social partner (as a recipient) to receive social information conveyed by the gazer and also to infer the gazer's attention.

Due to the importance of eye gaze in human-human communication, gaze behavior has been widely studied. It has been shown that during face-to-face conversation, two people look at one another quite often [Argyle and Graham 1976; Freeth et al. 2013]. On average, the speaker tends to look at the listener's face around 40% of the time, while the listener looks at the speaker's face 75% of the time. In addition, mutual gaze, when partners look at each other's face, accounts for 30% of those face looks. There is a dramatic decrease in face looking when there are other targets to look at, especially when there is an object of legitimate mutual attention. Argyle and Graham [1976] found that the gaze level in dyads fell from 76.3% when there was nothing much else to look at, to 6.4% when they were discussing a holiday plan with a relevant map between them. In addition, gaze, like touch and physical attractiveness, is a powerful reinforcement to strengthen social influence. For example, teachers who look more at their students in schools promote more work and learning from their students [Kleinke 1986].

Gaze has also been used as an important social behavior to build intelligent human-computer interactions, in particular, in human-robot interfaces [Admoni et al. 2014; Liu et al. 2012; Mutlu et al. 2009; Scassellati 1999; Vertegaal 2003; Yu et al. 2010]. Previous studies have demonstrated effective ways to improve overall user evaluation or performance using gaze cues in human-robot interactions [Hosoda et al. 2004; Kamashima et al. 2004]. For example, Mutlu et al. [2009] investigated the role of eye gaze in a story-telling robot and found that participants were better able to recall the story when the robot looked at them more while it was telling the story. Moreover, when interacting with a group of participants, the robot's looking behaviors directly influenced who would take a turn to speak next in the conversation. Yamazaki and colleagues [Yamazaki et al. 2008] performed experiments with a guide robot designed to use data from human experiments to turn its head toward the audience at important points during its presentation, which made participants demonstrate more non-verbal actions with precise timing as a response. Yoshikawa et al. [2006] built a robot that could move its gaze responsively to its interaction partner's gaze, showing responsive gaze to the partner's face could give stronger feelings of being looked at by the robot. Staudte and Crocker [2011] showed that human gaze was modulated by both robot speech and gaze, and that human comprehension of robot speech could be improved when the robot's language-related gaze behavior was similar to that of humans.

Even with the advances in robotics and sensing techniques, only a few studies have designed robotic systems that were able to access, process, and react to human participants' gaze behaviors in real time to study how real-time behaviors from a robot influence human participants' behaviors [Admoni et al. 2013; Yoshikawa et al. 2006]. Such systems need to implement gaze-contingent platforms in free-flowing human-robot interaction, and to collect and analyze micro-level behavioral patterns in human participants [Rich et al. 2010; Xu et al. 2013; Yu et al. 2010]. Toward this goal, the present article focuses on investigating real-time coupling of face looks in human-robot interactions and how specific face looks influence participants' multimodal coordinated behaviors, with three specific goals: (1) the study aims at providing empirical evidence on whether human participants naturally respond in real time to momentary gaze

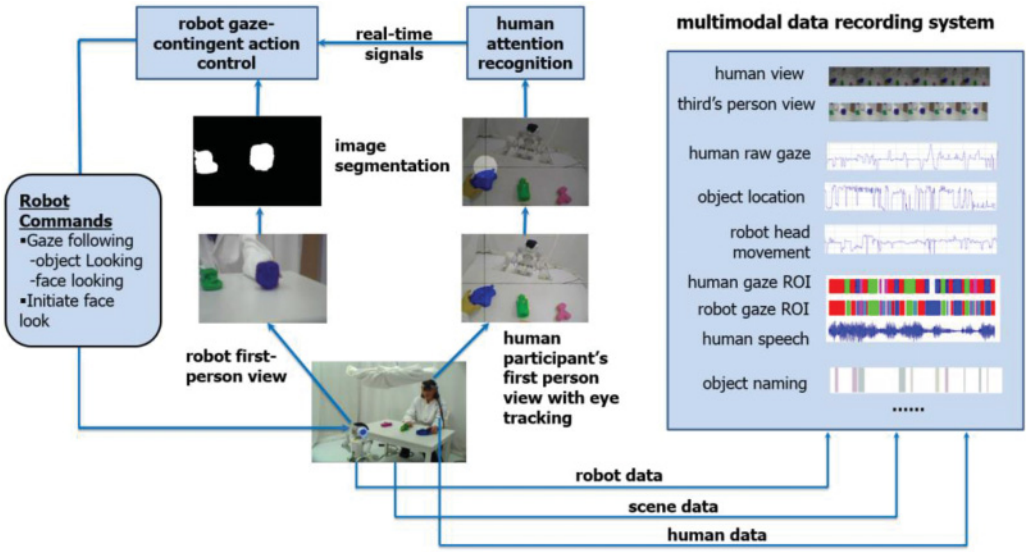

Fig. 1. An overview of real-time human-robot interaction. Left: a real-time human attention recognition system based on processing first-person view video and human gaze data. The system detected the attentional target from a human user moment by moment and passed such information to the robot control system. The robot then detected the same target from the robot's view and turned his head directly toward that target. Right: we recorded and analyzed multimodal data, including video, audio, and eye and head movement data from both the human and robot sides.

behavior generated by a robot partner in a free-flowing interaction; (2) we examine what looking behaviors from the robot can successfully elicit responses from human participants; and (3) we study how different mutual gaze patterns between the human and the robot may facilitate overall human-robot interaction.

The article is organized as follows. We start with describing a real-time gaze-contingent robotic platform developed for this study. We then present an experiment design with three conditions and report results with detailed patterns of human gaze and speech during human-robot interaction. We conclude with discussions on how to directly incorporate the findings from the present study into a robot control system to improve its performance [De Barbaro et al. 2013] and how to take advantage of guiding principles derived from the present study in future interactive intelligent system design.

2. REAL-TIME HUMAN-ROBOT INTERACTION

Figure 1 shows the structure of multimodal real-time human-robot interaction in which a human teacher attempted to teach a robot learner a set of novel object names (experimental details are described in the next section). In this context, the human manually manipulated the objects to move them to different spatial locations on the table throughout the entire experiment. In our gaze-contingent multimodal interaction system, the robot agent detected the human's visual attention moment-by-moment in real time, which allowed us to systematically manipulate the ways in which the robot reacted to the human's real-time gaze behavior, such as following the human participant's attentional switches immediately, wherever they were looking. To do so, the robot needed to look at various locations to follow the target object moment by moment. In the following, we describe several key components in the real-time system. A Nao humanoid robot by Aldebaran Robotics was used for the experiment. The Nao

robot has 35 Degree-Of-Freedoms (DOFs) as a whole. His eye unit is made of a Complementary Metal-Oxide Semiconductor (CMOS) camera with an image resolution of 640×480 at a sample rate of 30 frames per second. The camera's field of view is 58°. The Nao robot used here becomes a popular platform in various applications, such as RoboCup [Gouaillier et al. 2008], human-robot interaction therapy with autistic children [Niemüller et al. 2011; Shamsuddin et al. 2012], and gaze-based human-robot interaction [Andrist et al. 2014; Csapo et al. 2012; Jokinen and Wilcock 2014; Meena et al. 2012].

2.1. Visual Processing

As shown in Figure 1 (left), the key idea of the real-time control system is to detect the human's attention based on real-time eye tracking, and then to generate gaze-contingent responsive behaviors in the robot. We used an ASL head-mounted eye tracker (Applied Science Laboratories, LLC) to detect the human's gaze direction and integrated that information with the first-person view images captured from the head-mounted camera attached to the forehead of the human participant. The human attention detection system processed first-person view video and human gaze data to detect the attended objects moment by moment at the rate of 30fps (33ms). Since the interaction environment was covered with white curtains and visual objects were made with unique colors, detecting objects in the first-person view was done reliably based on color blobs. There are two steps involved to implement gaze-contingent interaction: (1) on the human side as shown in Figure 2(a), the system tracks and detects the human participant's eye gaze, determines the reliable moment that the participant switches attention from one target to another (e.g., a switch from the robot's face to the red object), and sends the gaze switch command to the robot's side; (2) on the robot side, as shown in Figure 2(b), the robot locates the target object in its first-person view and starts following the target object. In the Appendix, we explain the complete process of object segmentation and detection on both the participant's and the robot's sides, as well as the detailed information on the reliability test.

2.2. Attention Detection and Robot Control

The eye tracking and object detection system provides real-time data at the frequency of 30Hz as a sequence of Region-Of-Interest (ROI) derived from human eye gaze. In the present context, there are four ROIs (the social partner's face and three objects). The robot control system needs to send control commands about where to look based on the human's looking behaviors. A key challenge here is the stability of the control system because it is driven by momentary human gaze, which can be sporadic from time to time. In practice, participants might briefly look at one location for a very short period of time before quickly shifting to another location. In this case, if the robot starts executing a motor command to follow the first look, even before the robot fixates on the target, the human attention may already switch to the next target, and therefore, the robot would fail to follow the human's attention at the moment. In addition, the eye tracking system might sometimes lose track of human eyes, missing data in real-time control. Therefore, to build a control system that can reliably respond to human gaze, we designed the system in a way in which the robot determined to switch its attention only after it detected a stable and sustained look from the human. In implementation, the system kept track of not only the current gaze data point, but also a buffer of 30 data points in the past second. Only after more than 50% of data points in the buffer indicated the same target object (either the robot face or one of the objects), the control system would use this ROI to be the new target for the robot to follow. And if the new target didn't match the last detected one, this meant that the participant generated an attention switch. Only at this moment, a command was sent to the motor system in the robot's head to switch gaze and follow the new gaze direction from the human.

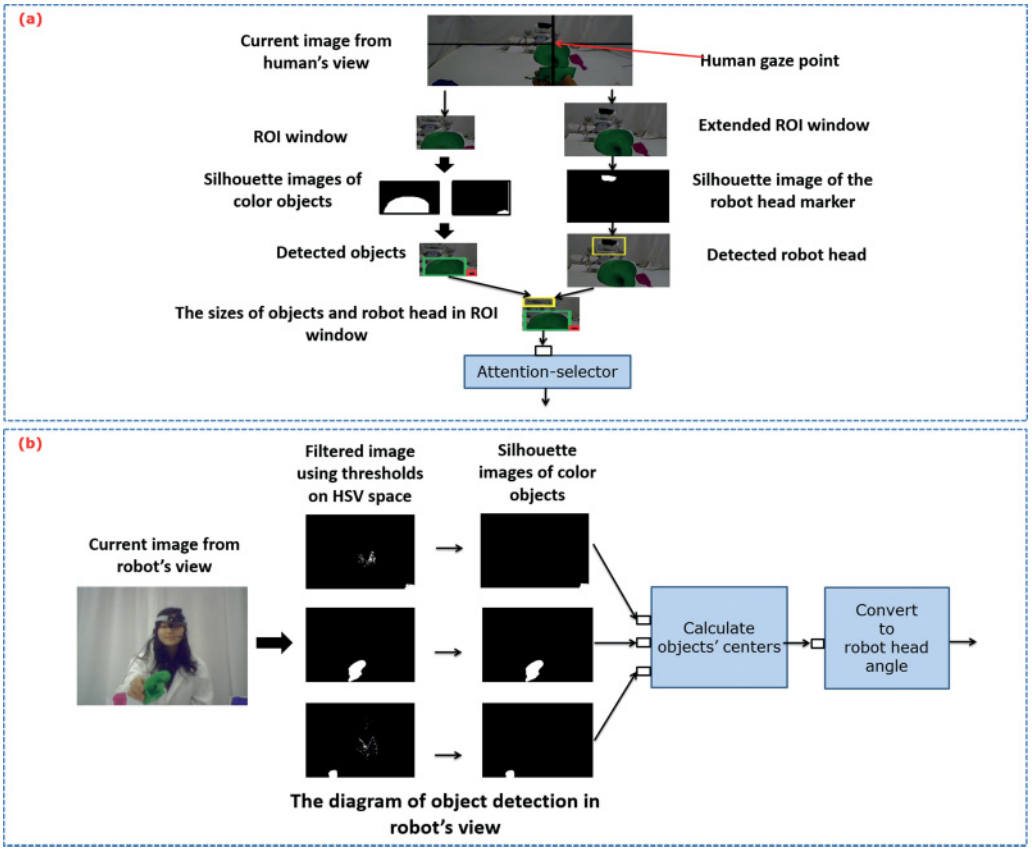

Fig. 2. The complete procedure of object segmentation and detection on both the participant's side (a) and the robot's side (b). More details concerning each step will be explained in the Appendix.

Figure 3 illustrates the detailed mechanism in determining a gaze switch in human participants. With multiple test runs, we found this mechanism is a good balance between being responsive and not being hypersensitive to short glances, which makes the entire interaction smooth and naturalistic.

2.3. Timing in Real-Time Control

After multiple rounds of system testing via pilot experiments, the attention switch detection mechanism described above added an additional 350ms lag on average. As explained in the section above, with this mechanism (Figure 3, p1), the robot's gaze-following behavior was smoother, and it didn't get trapped in switching and searching mode constantly during interaction. To gaze at a target object, the robot's visual system was then triggered to find the location of the target from the robot's camera view. The object location detection on the robot's side took about 50ms in real time. In the cases that the robot decided to turn its head toward the target object location, this motor command on average took 250ms to be executed. Taken together, the total system lag in principle was supposed to be about 633ms. We've performed a thorough test on the real-time system and the empirical result obtained was around 657ms. Thus, during an experiment, the robot system can follow the human's gaze direction on average within 657ms in our current implementation (close to the theoretical estimate of 633ms). Based on psychophysics literature, human adults generate three eye fixations per

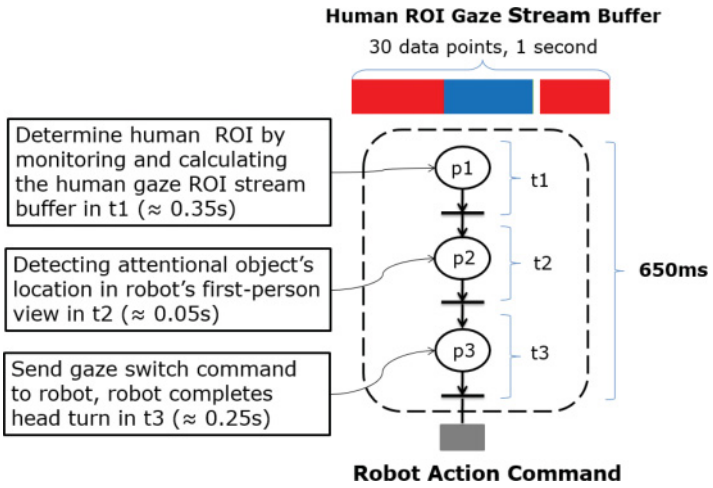

Fig. 3. The illustration of detailed timeline (three different phases) for human attention detection and robot gaze-contingent action control process. Only when more than 50% of data points in the Human ROI Gaze Stream Buffer indicated the same Region-Of-Interest (in this example, the red object), the control system would conclude this ROI to be the new attention target and send the control command to the robot to follow the new target attended by the human participant.

second on average (250–400ms per fixation) [Matin 1974; Volkmann 1986], and it takes about 300 to 400ms to plan and execute a saccade [Hyönä and Olson 1995; Liversedge and Findlay 2000; Sereno and Rayner 2003]. The total lag in robot following is, in general, in line with human real-time gaze behavior.

More generally, in addition to the above timing to execute a following command, there are a few more decisions needed to be made in the system (which will be explained in detail in Section 3.1) because the timing information is critical in gaze-contingent design and needs to be explicit and precise (for example, how long a look from the human needs to be for the robot to reply, and how long the robot should look at the human's face). As the first study at this micro level, we determined timing parameters in the system based on two general principles. First, we relied on the previous literature in psychophysics to closely emulate human behaviors as much as possible. Second, we did multiple test runs to examine what parameters lead to overall smooth interactions without eliciting abnormal and unexpected behaviors from participants. For example, response times that are too long or too short would make participants feel strange and uncomfortable about the robot.

3. EXPERIMENT

3.1. Experiment Design

We designed a word learning task in which a human teacher was instructed to teach the robot a set of object names in a shared environment. In the experiment, human participants needed to engage the robot in the task, attract the robot's attention to the target object to create joint attention moments, and then label object names for the robot learning (pseudo-English words were used, e.g., "bosa"). The word learning task was chosen for two reasons. First, the task was inspired from similar developmental psychology experiments showing how young children learn to associate visual objects with novel words through child-parent social interactions [Baldwin 1993; Estes et al. 2007; Yu and Smith 2012]. Second, there is a trend in developmental robotics in which users teach robots human languages through human-robot interactions [Lyon et al. 2012; Marocco et al. 2010; Tanaka and Matsuzoe 2012; Xu et al. 2013; Yu et al. 2010].

Moreover, this joint task naturally engaged participants to interact with the robot without any constraint on what they had to do or say. Participants were told to teach a baby robot the names of novel objects. They actively played the teacher's role and freely generated multimodal behaviors to attract the robot's attention, including eye contact, pointing to and manipulating objects in the shared environment, as well as describing and naming the objects in various ways. Like a human learner, the robot may or may not respond by switching its attention toward an object of interest. Thus, the interaction itself was free flowing and unscripted, allowing participants to generate naturalistic behaviors.

There were three experimental conditions, in all of which the robot followed the human participants' attention to establish and maintain joint attention on target objects with the human. That is, by default, the robot spent a large proportion of time on following the human's attention. The differences between the three conditions lied in what triggered the robot to look at the human's face and how long a face look lasted:

—**Responsive**: The robot looked at the participant's face as a response to the participant's look at the robot's face. When the participant looked away from the robot's face and moved on to a new target object, the robot also switched its attention to follow the human's attention on the same target. Thus, the robot in this condition always followed the human's attention either on the face or on one of the three objects. Eye contact between the dyad was both initialized and terminated by the human.
—**Extended responsive**: Similar to the above condition, the robot looked back at the human's face whenever the human was looking at the robot's face. Thus, just like in the responsive condition, the robot's face looks were still triggered by the human's looks. However, the robot in this condition continued looking at the human's face for 1.5 seconds even after the participant's face look to the robot was terminated. The timing of 1.5 seconds was chosen based on empirical data we gathered from test runs, as it gave participants enough time to respond to an extended face look if they wanted to do so. In this condition, eye contact was still initially established by the human, but the robot's extended face looks back on the human's face continued even after the human's gaze moved away from the robot, which could make the human generate a second look back to the robot's face, and by so doing, create a second eye contact.
—**Responsive and eliciting**: In addition to responding to the participant's face looks, just like in the above two conditions, the robot in this condition also generated additional looks toward the human's face to initialize eye contact at the moment when the participant was looking at objects for more than 3 seconds, which was a relatively long time without face looks between the two partners. In this condition, either the human or the robot could establish eye contact. More specifically, if the human didn't do so for 3 seconds, then the robot would initiate one instead.

The first hypothesis to test was whether the participants would be sensitive to the behavioral-design manipulations among the above three different conditions, and as a result, behave differently. Specifically, more mutual gaze would be created in the extended responsive condition and responsive and eliciting condition since the participants would notice and react to more frequent and longer face looks from the robot, and thus, generate more face looks in return. Secondly, we also hypothesized that more eye contact would make participants more engaged, resulting in a smoother interaction which would be revealed by more naturalistic and coordinated behaviors on the human side at the micro-behavioral level. Both hypotheses can be tested by collecting and analyzing the participants' micro-level behavioral data during experiments and how their responsive gaze and speech patterns differ across three conditions.

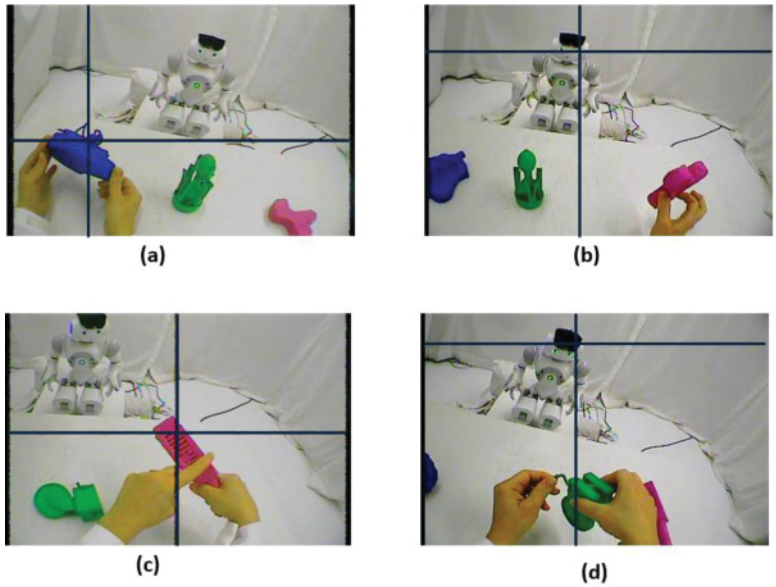

Fig. 4. Four joint attentional states in the interaction from the participant's first-person view (human gaze indicated by a black crosshair): (a) the human was looking at a target object while the robot was looking at the human's face; (b) mutual gaze: both the human and the robot looked at each other's face; (c) both the robot and the human jointly attended to the same object; and (d) the robot was attending to an object in the human's hands while the human gazed at the robot's face.

3.2. Experiment Procedure

Eighteen students at Indiana University participated in the study (five additional participants were excluded due to low eye tracking rates). Participants were given three sets of three novel objects, with each set used in one experimental condition. More specifically, each set contained one blue, one green, and one pink object. Each object was given an artificial two-syllable name, i.e., kaki, regli, or gasser. Participants were asked to teach the robot in three trials with trial orders randomized across participants. The eye tracker on the human's head was calibrated before the experiment started. The participants were then instructed to teach the robot about three objects in a set for the full duration of 2 minutes for each trial. At the end of each trial, an experimenter signaled participants to stop and asked participants to take a voluntary break before starting a new trial with a new set of three objects. Figure 4 shows different coupled gaze patterns during the experiment from the human's first-person view. A demo video from the responsive condition can be accessed via this link (https://www.youtube.com/watch?v=vGYl7tDe5pM).

4. RESULTS

Our data analyses focused on gaze and speech data from human participants in three experimental conditions, with respect to the robot's gaze following and initiating behaviors. For gaze data, we recorded where the human and the robot attended, moment-by-moment. Figure 5 showed three example ROI gaze streams from both the human and the robot in the three conditions. For speech data, we transcribed speech into text and, additionally, we categorized spoken utterances into four speech act types: naming, describing, attention-getting, and confirming. The present study used two types of speech acts: *naming* when participants uttered object names in speech and *describing* when participants described properties of visual objects. Table I shows the complete

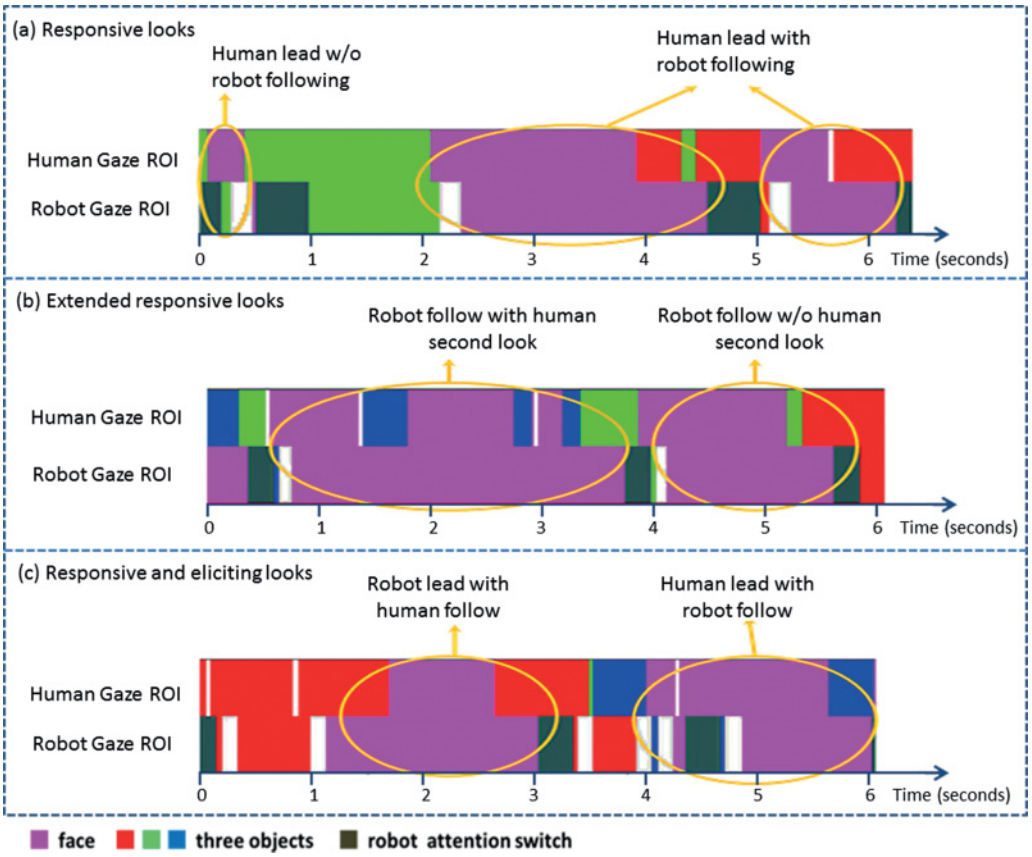

Fig. 5. Examples of the robot's and the human's gaze data streams from three experimental conditions. (a) Responsive looks: the robot ignored brief face looks from the human and exactly copied human gaze behaviors with a short delay. (b) Extended responsive looks: the robot responded to the human's looks by looking back at the human's face, and the robot continued looking at the human's face for another 1.5 seconds even after the human looked away. Humans may or may not generate a second face look to respond. (c) Responsive and eliciting looks: the robot not only followed the human's face looks as it did in the other two conditions, but it also attempted to initiate eye contact by looking at the human's face when the human's attention was not on the robot's face.

transcripts from the demo video (with naming instances highlighted in bold and italic font style). In the following, we will report two sets of results—the first focuses on face look and mutual gaze (Section 4.1) and the second focuses on multimodal behaviors between gaze and speech (Section 4.2).

4.1. Gaze Behaviors from Human Participants

This section reports patterns derived from human gaze behaviors and in particular, how participants adjusted where they looked as a response to the robot's face looks. We first report several general statistics from three individual conditions, and then summarize and compare gaze patterns across the three conditions.

4.1.1. Gaze Behavior in the Responsive Condition. Figure 5(a) shows an example of raw gaze data from the responsive condition in which the robot exactly followed the human's attention moment by moment, either on the human's face or on the same target object, creating a sequence of joint-attention moments. A closer examination of gaze data revealed that the mean duration of the human's face looks was 1.12 seconds when the

Table I. An Example Transcript of Human Participant's Speech

Participant	You can trust me on this being a [***wawa***]. It looks kinda of. . . you could imagine it being some sort of brush or a scoop (manipulating the object in a brushing-like motion)
Robot	(gazing at the pink object and following the target object as it moved)
Participant	You could use it to pick things up. (putting green object and pink object together) You could not pick these things up. (trying to use pink object to scoop up the green object) They are too heavy. Ummm, I think we will move back to the [***blicket***]. (picking up the green object and holding it in the center of his visual field)
Robot	(switching to look at the green object, lifting head up and down to keep gazing at the green object since the participant was manipulating the object with his hands)
Participant	The [***blicket***] is green. This is a [***blicket***]. You can turn the crank but it doesn't do anything. And you can movie it. . . like that. (putting down the green object and picking up the blue object) And the [***mobit***] is a gigantic blue blob.
Robot	(looking up toward the blue object as the participant held the object in his hand and in the center of the participant's first-person view)
Participant	And the [***mobit***] is a gigantic blue blob. <video end>

Note: This is from the demo video in the responsive condition (video link: https://www.youtube.com/watch?v=vGYl7tDe5pM). The object names in speech were highlighted in bold and italic fonts. It showed that the participant used different sentence structures to describe the three objects to the robot and actively manipulated the objects during his verbal descriptions, while the robot acted very responsively by following the participant's attention.

robot followed the human's face looks, while the gaze duration dropped to 0.31 seconds when the robot didn't follow the human's face gaze. There are two possible reasons to explain the huge difference. One is that the robot's responsive face looks made the human look longer to the robot's face. The other possibility is that the difference may have nothing to do with the human's responsive actions to the robot's gaze behaviors; but instead, the human pre-determined how long to look at the robot's face, and those longer looks led to responses from the robot, while shorter ones didn't get responses. The results from the other two conditions allowed us to determine which explanation is more plausible.

4.1.2. Gaze Behavior in the Extended Responsive Condition. The results from the extended responsive condition provided further evidence about the human's reaction to the robot's face looks. As shown in Figure 5(b), the robot in this condition continued looking at the human's face for another 1.5 seconds after the human's gaze moved away from the robot's face. Two consequential patterns appeared based on this manipulation. First, when the robot continued looking at the human's face even after the human looked away from its own face (around 47.01% of those instances) the human looked back to the robot's face as a response to the robot's continued face looks (see Figure 5(b)), and the mean duration of the human's second face look was 0.76sec, which was significantly longer than their average duration of face looks in the responsive condition ($M_{responsive} = 0.59$, $t(17) = 5.88$, $p < 0.001$). Second, since the human looked back by taking the robot's bid on his attention, in turn, this made the robot keep looking at the human's face even more, which formed a feedback loop on mutual gaze until the human decided

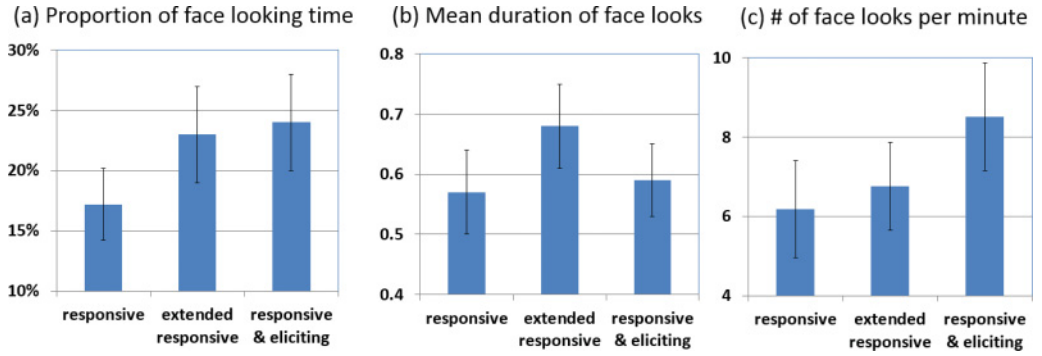

Fig. 6. Different measures of the human participants' gaze behaviors across three conditions. (a) Proportion of face-looking time across the three experimental conditions. (b) Gaze duration on the robot's face. (c) Number of face looks per minute.

to look away and didn't look back again (see Figure 5(b)). As a result, the mean duration of overall human face looks was much longer ($M = 0.68$) than that in the responsive condition ($M_{responsive} = 0.59$; $t(17) = 8.67$, $p < 0.001$). In addition, the robot's face looks were also significantly longer ($M = 3.54$) compared with those in the responsive condition ($M = 1.29$). Note that the dramatic difference cannot be simply accounted by adding 1.5 seconds ($t(17) = 6.47$, $p < 0.005$), suggesting that the robot's face looks led to the human's second face look, which in turn made the robot look more on the human's face.

4.1.3. Gaze Behavior in the Responsive and Eliciting Condition. In both responsive and extended responsive conditions, the robot's gaze at the human's face was always triggered by the human's face looks. As shown in Figure 5(c), the robot in the present condition elicited the human's attention by looking at the human's face first when the human was looking at one of the objects (thus, not looking at the robot's face). We found that 29% of the time, when the robot attempted to lead by initiating face looks, participants took the robot's bids by switching their attention from a previously attended object to the robot's face. In particular, the average timing between the onset of the robot's face look and the onset of the human's responsive face look was 691ms. Given that it took roughly 300ms to plan and execute a saccade in the human [Frischen et al. 2007; Posner 1980] (switching eye gaze from one spatial location to another) and there was also a timing for the human to detect the robot's face look in the human's peripheral vision, we concluded that participants were capable of immediately detecting the robot's face looks, and then promptly switching their attention to the robot's face as a response if they decided to do so.

4.1.4. Comparison of Gaze Behaviors across Three Experimental Conditions. In this section, we compare gaze patterns across the three experimental conditions and report both shared and different gaze patterns across the three. Since the key manipulation of this gaze-contingent paradigm was the robot's face-look behavior responding to the human's face look, a critical question is how the robot's behavior may influence the human's face looks toward the robot. The potential differences between the three experimental conditions can be captured by three measures of looking behavior: (1) proportion of time looking at the robot's face; (2) looking duration; and (3) frequency of face looks. As shown in Figure 6(a), there was a statistically significant difference between three conditions by one-way ANOVA ($F(2, 53) = 15.258$, $p < 0.0001$). And post hoc tests revealed that both longer face looks from the robot in the extended responsive and more initiative face looks in the responsive and eliciting condition made participants look more toward

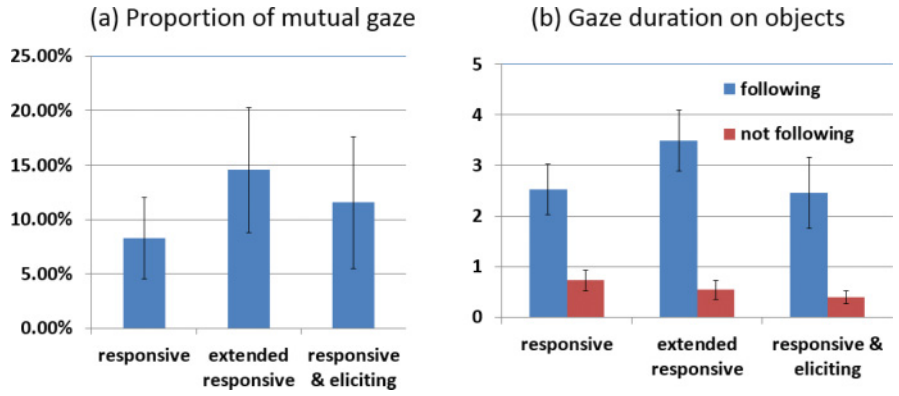

Fig. 7. The proportions of mutual gaze time in which both the robot and the participant looked at each other's faces.

the robot than what they did in the responsive condition as a baseline ($Prop_{(responsive)} = 17.12 \pm 3\%$, $Prop_{(extended\_responsive)} = 22.25 \pm 4\%$, $p < 0.01$; $Prop_{(responsive\&eliciting)} = 23.56 \pm 4\%$, $p < 0.01$). Further, we found two different pathways to an overall larger proportion of face look time shown in Figure 6(a). As shown in Figure 6(b), with strong group difference ($F(2,53) = 13.836$, $p < 0.0001$), participants in the extended responsive condition (0.68 ± 0.07 sec) generated longer face looks compared with the responsive condition (0.56 ± 0.07 sec, $p < 0.01$) and responsive and eliciting condition (0.59 ± 0.07 sec, $p < 0.05$). Also, from Figure 6(c), participants in the responsive and eliciting condition produced more face-looks toward the robot (8.15 ± 1.23 looks/min) than the responsive condition (6.18 ± 1.11 looks/min, $p < 0.005$). Thus, longer face looks from the robot led to longer looks from the human, and more looks from the robot elicited more looks from the human. This can be caused by spontaneous reactions to the robot's behavior or by self-conscious social control [Chartrand and Bargh 1999]. In either case, it is an important demonstration that the robot's behavior can influence the human's responsive behavior in real-time, and in a rhythmic and systematical way. Also note that in the extended responsive condition, participants generated more face looks (6.76 ± 1.11 looks/min, $p < 0.01$) compared with the baseline condition due to the additional second face looks during the extended time, as described in Section 4.1.2.

4.1.5. Mutual Gaze as Coupled Behavior. As shown above, various gaze behaviors from the robot were noticed and responded to by participants, which also changed joint gaze patterns between the two social partners. Here, we focus on mutual gaze, which has been demonstrated to be important in face-to-face interaction [Goodwin 1980; Nakano et al. 2003]. With more face looks between the human and the robot, there was a dramatic increase of mutual gaze moments with a higher proportion of total time in the extended condition ($14.55 \pm 3.7\%$, $p < 0.001$) and responsive and eliciting condition ($11.55 \pm 6\%$, $p < 0.01$), compared with the responsive condition ($8.27 \pm 5.7\%$; $F(2, 53) = 6.408$, $p = 0.003$), as shown in Figure 7. Mutual gaze as coordinated behavior had to be created and maintained by joint activities between two social partners. In the extended responsive condition, the robot generated longer looks, which led people to generate more looks toward the robot. In the responsive and eliciting condition, the robot initiated additional face looks that were responded to by participants. In both cases, reciprocal behaviors from both social partners jointly created more mutual gaze between the two. In human-human interactions, people (even infants) prefer to look at faces that engage them in mutual gaze with other social partners [Cohn and Tronick

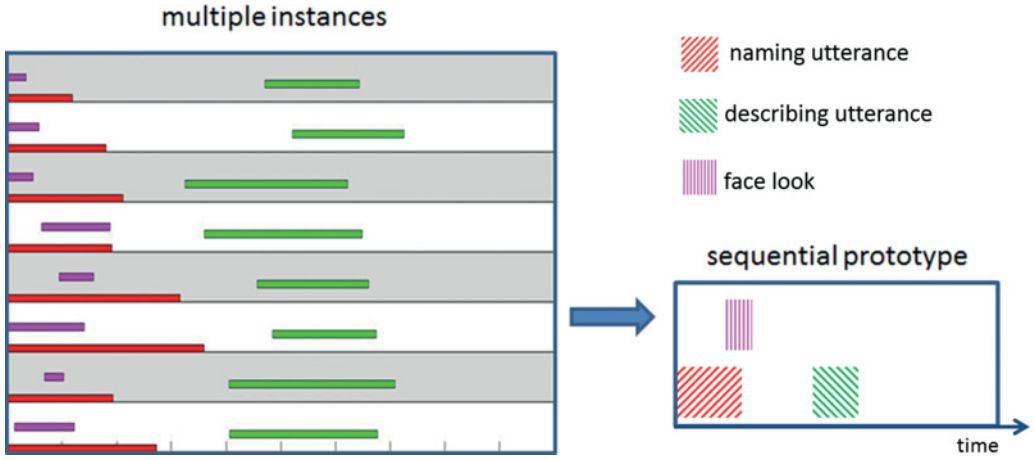

Fig. 8. Given multiple temporal instances, the algorithm computes the most probable sequential prototype based on comparing and matching individual instances.

1987; Fogel 1977]. More mutual gaze can be interpreted as participants cared about the robot by responding to the robot's face looks [Boucher et al. 2012; Sidner et al. 2005]. Therefore, the robot's elicitation of mutual gaze with human users has the potential to be a powerful reinforcement to strengthen both the robot's social influence and the user's social engagement in the interaction. In light of this, the present results show that both lengthened face looks and more face looks from the robot can achieve this goal.

So far, our results showed that participants were sensitive to real-time gaze behaviors generated by the robot, and in particular, they were more likely to respond to the robot's bids of eye contact and look back to the robot's face. Next, we are going to investigate whether more eye contact made participants more engaged in the teaching task, and by doing so, they would demonstrate more coordinated behaviors to create a better teaching environment for the robot learner.

4.2. Coordinated Multimodal Behaviors

In the present task, the goal of participants as a language teacher has been to attract the robot learner's attention and then name and describe to-be-learned objects. A typical speech act that appeared with the highest frequency from speech transcriptions was a naming event (e.g., "this is a bosa", "look, bosa") followed by one or several describing utterances (e.g., "bosa is red with a round shape"). A sample recording of responsive condition with the human participant's speech transcriptions and robot responses was provided in Table I. In total, participants generated 282 naming-and-describing speech sequences with approximately the same number of instances in each of three experimental conditions. We considered these moments as critical for teaching in social interaction, and zoomed into the moments to analyze the coordination between speech (verbal) and gaze (non-verbal) behaviors by examining where participants looked when they produced naming and describing utterances.

We used a sequential pattern-mining algorithm that was developed to extract exact timings and durations of sequential patterns from multiple temporal event streams [Fricker et al. 2011]. As shown in Figure 8, the first step in this method is to segment and decompose continuous data streams into multiple local instances, wherein each instance consists of a set of temporal events from multiple event utterances. The onsets of these events were used to segment and align individual instances (e.g. red events in Figure 8). As a result, a set of multiple sequential instances were temporally aligned,

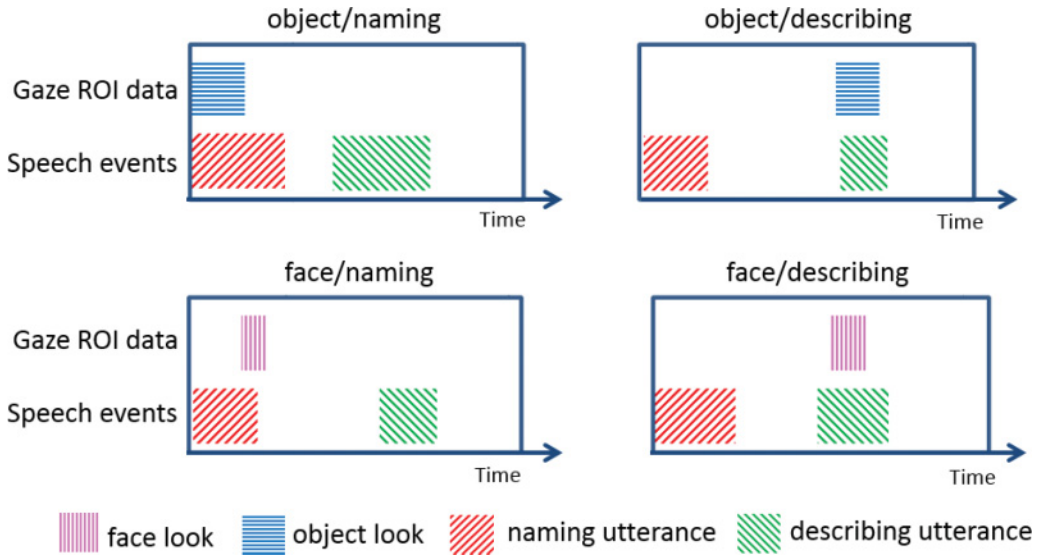

Fig. 9. Four speech-gaze temporal patterns: object/naming – looking at the target object while naming it; object/describing – looking at the target object while verbally describing it; face/naming – looking at the robot-learner's face while naming an object; and, face/describing – looking at the robot's face while describing an object. Note that these patterns contain the information about not only the temporal order of these multimodal events, but also the timings and durations between and within these events.

each starting with the onset of a naming event and lasting for 5 seconds after. Next, given multiple sequential instances aligned in time, the algorithm calculates a set of sequential event prototypes by aggregating and statistically matching these instances. Each prototype is defined by a set of multimodal events with specific durations and timings. For example, in Figure 8, 50% of the instances contain a face look (pink) followed by a naming utterance (red) with a lag time of 500ms, and the remaining 50% contain the same sequential pattern, but with a lag of 700ms between a face look and the initial naming utterance. Then, in this case, two prototypes will be derived, with a specific time lag for each. In addition, the algorithm generates and sorts derived prototypes based on a frequency score, indicating how many instances from raw data support a particular prototype. Technical details of this algorithm can be found in Fricker et al. [2011].

As the first steps of analyzing micro-level speech and gaze data collected from the free-flowing human-robot interaction, we focused on extracting sequential patterns from four data streams: two speech events (naming and describing, etc.) and two gaze event types (looking at the robot's face or looking at the target object, etc.). Thus, we fed four data streams to the sequential event detection algorithm and Figure 9 shows the top four statistically reliable sequential prototypes extracted. The top two in Figure 9 are composed of speech and object-gazing events, and the bottom two are composed of speech and face-gazing events. Within each event type, the difference between two patterns lies in the timing of looking at the target. For example, the two object-gazing patterns capture two different kinds of coordination between speech and gaze: participants looked at the target at a naming moment in one case versus they did so at the describing moment in the other case. Similarly, the two face-look patterns shown in the bottom two plots in Figure 9 were made of two different timings to look at the robot's face. In one case, participants looked at the robot's face while naming the target object; in the other case, they did so when describing the target object. These speech-gaze patterns may serve different roles in the interaction and reflect different processes and

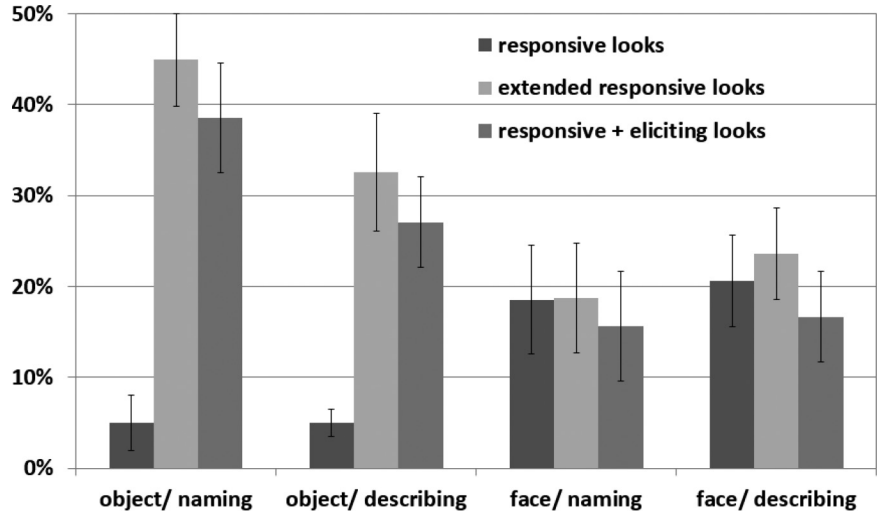

Fig. 10. A comparison of the proportions of instances that participants in three experimental conditions exhibited the four coordinated sequential patterns. Compared with those in the responsive condition, participants in both extended responsive and responsive and eliciting conditions generated more synchronized gaze-speech patterns toward the target object when naming and describing that object. In all of the three conditions, they attended to and checked the robot's face during naming and describing events with no difference among the three conditions.

internal states in human cognitive systems. For instance, it has been shown that speakers tend to look at the target object when they name that object (pattern *object/naming* in Figure 9) and describe it in speech (pattern *object/describing* in Figure 9) [Griffin and Bock 2000]. In addition, monitoring and checking the robot learner's face while naming or describing the target object (*face/naming* and *face/describing* in Figure 9) also seemed to be a necessary behavioral component in a smooth interaction, as the teacher needed to make sure the robot learner was actually paying attention the target object while hearing its name. Indeed, these two speech-and-face-look patterns have been demonstrated by parents teaching their children to learn object names in which parents frequently assessed the child's attention to ensure more joint attention moments [Mundy and Newell 2007; Yu and Smith 2012]. Hence, we argue that all of the four speech-gaze patterns revealed naturalistic and coordinated human behaviors and can be treated as good indicators of smooth interaction. Therefore, the more participants generated these behavioral patterns and demonstrated multimodal regularities and synchrony between speech and gaze, the smoother their interaction with the robot went. Moreover, in the context of language teaching, more multimodal coordination from the teacher creates better teaching signals that can facilitate learning through social interactions—a research finding that has already been convincingly shown in both human learning [Nagai and Rohlfing 2007; Pitsch et al. 2009] and robot learning [Nagai et al. 2008].

The proportions of instances that matched with each of four sequential patterns were calculated in three experimental conditions. Note that these four patterns are not mutually exclusive. For example, participants may look at the robot's face during both naming and describing speech acts, which created two different instances—once as object/naming and once as object/describing. As shown in Figure 10, for two face-look patterns, there is no significant difference among three conditions ($F(2, 53) = 2.51$, $p = 0.12$), suggesting that in all of the three conditions, participants paid attention to the robot's face when producing either naming or describing utterances. However,

for two object-gazing patterns, participants in "extended responsive" and "responsive and eliciting" conditions demonstrated more coordinated behaviors compared with those in the responsive condition. They were more likely to look at the target object while naming it ($M_{responsive} = 5.22\%$; $M_{extended\_responsive} = 44.94\%$, $p < 0.001$; $M_{responsive\&eliciting} = 38.54\%$, $p < 0.001$) or describing it ($M_{responsive} = 5.56\%$; $M_{extended\_responsive} = 32.58\%$, $p < 0.001$; $M_{responsive\&eliciting} = 38.54\%$, $p < 0.001$) in the two experimental conditions compared with the responsive condition. As mentioned earlier, looking at the target object while naming or describing it is evidenced in psycholinguistics studies in language production [Griffin and Bock 2000]. In addition, the synchrony between speech and gaze provides better teaching signals for language learning [Gogate 2010; Gogate et al. 2000; Rolf et al. 2009]. Putting this together with the results from the previous sections, we concluded that the robot was able to influence the human's gaze behaviors through either increasing the amount of face-look time or initializing eye contact by itself. As a response to the robot's face looks, participants looked back more and created more mutual gaze. Further, more eye contact seemed to make participants more engaged in the interaction, and therefore, they demonstrated more coordinated speech-gaze behaviors, which led to both a smoother interaction and better teaching signals for the robot in the language learning task.

5. GENERAL DISCUSSION

The overall idea that motivated the present work is that a deep understanding of human-robot interactions requires a level of analysis that concentrates on sensory-motor behaviors in which the behaviors of social partners continuously adjust to and influence each other in real time [Marsh et al. 2009]. Since those behaviors happen in fractions of a second, they have to operate at a sensory-motor level and be supported by low-level sensory-motor processes. Nonetheless, the same sensory-motor behavioral exchanges between the human and the robot may shape social interaction in a profound way. The present article focuses on one particular looking behavior—gazing at each other's face. Face look and mutual gaze between two social partners are critical in smooth human-human interactions [Argyle 1988; Clark and Brennan 1991]. The general impression formed from looking at a person's face is that the other is interested and wishing to initiate interaction. Therefore, investigating at what moments and in what ways the robot should look at the human's face in the context of the human's spontaneous gaze behavior is an important topic [Tapus et al. 2007].

5.1. Adaptive Behaviors at the Micro Level

With our system design and gaze-contingent platform implementation, the results showed that the participants were very sensitive and responsive to the robot's face looks and they rapidly adjusted their behaviors based on their perception of the robot's behavior. For example, in the extended responsive condition, longer face looks from the robot not only made people look back to the robot's face but also made their second face looks longer (0.76 seconds) than their average looking duration (0.59 seconds). Recent studies in human-human interaction have shown nonconscious responses of postures, mannerisms, facial expressions, and other behaviors of one's interaction partners, such that people may passively and unintentionally change their behaviors to match the behaviors of others [Chartrand and Bargh 1999]. An interesting question is whether participants in the present study responded to the robot's looks intentionally. Alternatively, just like mimicry effects, their responses could be unintentional and nonconscious. One relevant piece of evidence for this question is derived from response time—participants' responsive gaze behaviors were prompt. In the responsive and eliciting condition, among all the instances that the robot initialized eye contact, more than 29% of the time, participants responded to look back to the robot's face within a window of

700ms after the onset of the robot's initial look. It is unlikely that these fast responses can be generated by high-level cognitive and social processes, but instead more relied on sensory-motor processes to take immediate actions. Whether operating on high-level or low-level processes, participants in the present study seemed to demonstrate prompt and contingent responses that are similar to what they do in human-human interactions [Goldstein et al. 2009; Kuhl et al. 2003]. Prompt responses from humans pose the requirement of real-time adaptive actions from interactive robots that are expected to interact with humans in a human-like way.

5.2. Engaging Human Participants through Responsive and Contingent Actions

More responsive looks on the robot's face suggested the human's interests to the robot and more eye contact between the two may serve as social reinforcement to influence and encourage participants to interact more with the robot and treat the robot as a social partner. As follow-up studies are necessary to further investigate how to take advantage of human-robot eye contact to facilitate smooth interaction and how to make eye contact more human-like, the results here are encouraging as they clearly show that more eye contact between the human and the robot can be accomplished by controlling the robot to direct its attention to the human face.

In a recent study reported in Freeth et al. [2013], participants were asked to answer questions from an experimenter. Interactions were conducted either live or via video. In the live condition, participants and an experimenter completed a one-on-one, face-to-face interaction. In the video condition, an experimenter was videotaped from a distance and the video was displayed on a monitor. Only in the live interaction condition, modifications of the experimenter's eye contact influenced participants' eye movements. They looked more at the experimenter's face when eye contact was made, but the direction of the experimenter's gaze had no influence on participants' viewing behavior in the video-based condition. Taken together with our finding that the changes of gaze direction from a physical robot influence participants' gaze behaviors, people are more responsive toward a physical robot compared with another human on a computer screen. This indicates the potential feasibility of creating human-robot interaction through real-time micro-level behavioral exchanges to approach smoothness and spontaneity in natural human-human interaction. For example, if a robot learner intends to re-direct the human's attention, one effective way is to initialize eye contact first, wait for the human's look back, and then at the exact moment when the human looks back, the robot can direct his attention to the target right after the establishment of initial eye contact. Based on our findings here, this strategy has a better chance to redirect the human's attention to share with the robot's attention. More generally, eye contact can serve the function of attention-getting in human-robot interaction, which is a building block of social interaction and social learning.

We also note that even mutual gaze seems to play a positive role in the present study, "too much" mutual gaze, wherein two partners just stare at each other's face may hinder interaction. In a recent study [Wang and Gratch 2010], the authors found negative evaluations when an interacting virtual agent was not responsive but just staring at the participants 100% of the time. Instead, if mutual gaze was accompanied by other responsive behaviors, such as eye blinking and nodding, participants became more positive about the interaction. Combined with our results, building smooth human-robot interaction—whether it is through mutual gaze or other behaviors—needs to incorporate contingency and responsiveness in real-time coupled behaviors.

5.3. Limitations and Future Work

Our experimental findings can be used to guide the development of a gaze-control system in embodied conversational agents. For example, we can cast a general prediction,

based on our data, when participants may respond to the robot's face look and look back to the robot's face. In this way, our experiments and collected results can provide insights on how a gaze-contingent robotic system creates better engagement to facilitate coordination in real time by reducing human users' cognitive loads during face-to-face interaction. Furthermore, by studying the temporal coupling of eye gaze and speech, we will be able to design behavioral scripts that will allow artificial agents to assume the role of a learner that is intuitive and easy to follow for human teachers. Incorporating those findings into a gaze model and implementing them in robot systems will allow us to explicitly test the real-time mechanism.

The gaze-contingent system and experimental design in the present study require us to include multiple temporal variables in the robotic system. Thus, the detailed reported results, such as gaze duration, were conditional on certain thresholds used in our system and led to follow-up questions. For example, it took the robot 633ms on average, empirically, to follow the human's attention switch; the robot looked at the human's face for 1.5 more seconds after the human face look was terminated; and, the robot initialized a face look if there was no face look in the past 3 seconds. One consistent observation shared among various findings reported in the present study is that people are very sensitive to real-time behaviors from the robot. This leads to an important question—whether different timings programmed in the robot would lead to different kinds of responses. For example, instead of using 1.5 secs in the extended condition, would using a 2.5-second threshold completely change people's responses? More generally, would some delay or speed-up in the robot's responses change how people respond? As the first steps to understand real-time gaze-contingent behaviors in human-robot interaction, the results here are informative about the trend and nature of the human behavior in these particular experimental conditions. However, given spontaneous responses observed in the present study with our current threshold settings, it is critical to further examine different timing factors and detailed ways in which they may influence people's responses.

We also note that the present findings are derived from one joint-attention task with a specific experimental setup, and therefore, we are interested in further testing and extending these results with different kinds of collaborative tasks to generalize and infer fundamental principles in human-robot interaction. In addition, studies on human-human and human-robot interactions also show individual differences due to subjective judgements of the current task and previous personal experiences [Fischer 2011]. Along this line, we did not include any survey data in the present study that can be critically informative of participants' overall assessment of how well they interacted with the robot. Such information can be integrated together with micro-level behavioral data, such as speech and gaze, so that we can link and compare what participants feel with what momentary multimodal behaviors they generate. Thus, combining and integrating survey data and micro-level behavioral data will provide a unique opportunity to better understand multimodal human-robot interaction.

APPENDIX

In this appendix, we explain the complete process of object segmentation and detection on both the participant's and the robot's sides. Figure 2(a) shows the steps of object detection in human participant's first-person view. By attaching a visual marker on Nao's forehead, the module of "Human attention recognition" detects the face of Nao and the three single-colored objects from human's first-person view image (320×240 pixels). The basic idea is to compare the four different target ROIs with a 70×0 window centered at the human gaze point. It consists of three steps:

- Step 1: Sub-images in the ROI window are used to generate silhouettes of the objects or the robot's head. We converted the current image frame into an HSV space (Hue,

Saturation, and Value) and used thresholds in the HSV space to generate binary images that contain candidates of the silhouettes of the single-colored objects and the visual patch on Nao's head. Then, we filtered out noise pixels by a weighted median filter incorporating both the size and shape of each blob with regard to the entire frame.

- Step 2: A sub-image in an extended ROI window (140×140) is used to detect the robot's head around the human's eye gaze. The reason to use an extended ROI window is that the black marker put on the robot's head can sometimes be out of the ROI window while the lower-half part of the robot's head is still within the ROI window. After the marker is detected, the program generates a rectangle to cover the robot's head and neck according to the marker's position and size.
- Step 3: The third step is to find the largest target ROI among the robot's head region and the color objects within the above-mentioned window, which will be identified as the ROI. For instance, in Figure 2(a), the program would detect that the human is looking at the green object.

A similar process of color-blob detection and segmentation was done in the robot's view (Figure 2(b)) with different thresholds to fit with image properties in Nao's forehead camera. In order to ensure the accuracy of object detection, we randomly generated about 50 frames per participant and asked human coders to manually label and segment individual objects and faces, frame-by-frame. The object/face detection accuracy is above 98% because of our simplified white room set-up and customized HSV thresholds for each target.

ACKNOWLEDGMENTS

We thank Hong-Wei Shen and Amanda Favata for help with running experiments. Henry Choi and Hong-Wei Shen were involved with developing image processing and data analysis programs.

REFERENCES

Henny Admoni, Christopher Datsikas, and Brian Scassellati. 2014. Speech and gaze conflicts in collaborative human-robot interactions. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society (CogSci'14)*.

Henny Admoni, Bradley Hayes, David Feil-Seifer, Daniel Ullman, and Brian Scassellati. 2013. Are you looking at me?: Perception of robot attention is mediated by gaze type and group size. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE Press, 389–396.

Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. 2014. Conversational gaze aversion for humanlike robots. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 25–32.

M. Argyle. 1988. *Bodily Communication*. Methuen, New York, NY.

Michael Argyle and Jean Ann Graham. 1976. The central Europe experiment: Looking at persons and looking at objects. *Journal of Nonverbal Behavior* 1, 1 (1976), 6–16.

Dare A. Baldwin. 1993. Infants' ability to consult the speaker for clues to word reference. *Journal of Child Language* 20, 02 (1993), 395–418.

Jean-David Boucher, Ugo Pattacini, Amelie Lelong, Gerard Bailly, Frederic Elisei, Sascha Fagel, Peter Ford Dominey, and Jocelyne Ventre-Dominey. 2012. I reach faster when I see you look: Gaze effects in human–human and human–robot face-to-face cooperation. *Frontiers in Neurorobotics* 6, 3 (2012), 1–11.

Tanya L. Chartrand and John A. Bargh. 1999. The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology* 76, 6 (1999), 893.

H. H. Clark and S. E. Brennan. 1991. Grounding in communication. *Perspectives on Socially Shared Cognition* 13 (1991), 127–149.

Jeffrey F. Cohn and Edward Z. Tronick. 1987. Mother–infant face-to-face interaction: The sequence of dyadic states at 3, 6, and 9 months. *Developmental Psychology* 23, 1 (1987), 68.

Adam Csapo, Emer Gilmartin, Jonathan Grizou, Jingguang Han, Raveesh Meena, Dimitra Anastasiou, Kristiina Jokinen, and Graham Wilcock. 2012. Multimodal conversational interaction with a humanoid robot. In *Proceedings of the 2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)*. IEEE, 667–672.

K. De Barbaro, C. M. Johnson, D. Forster, and G. O. Deak. 2013. Methodological considerations for investigating the microdynamics of social interaction development. *IEEE Transactions on Autonomous Mental Development* 5, 258–270.

Katharine Graf Estes, Julia L. Evans, Martha W. Alibali, and Jenny R. Saffran. 2007. Can infants map meaning to newly segmented words? *Statistical Segmentation and Word Learning. Psychological Science* 18, 3 (2007), 254–260.

Kerstin Fischer. 2011. Interpersonal variation in understanding robots as social actors. In *Proceedings of the 6th International Conference on Human-Robot Interaction*. ACM, 53–60.

Alan Fogel. 1977. Temporal organization in mother-infant face-to-face interaction. In *Studies in Mother-Infant Interaction*. 119–152.

Megan Freeth, Tom Foulsham, and Alan Kingstone. 2013. What affects social attention? Social presence, eye contact and autistic traits. *PLoS One* 8, 1 (2013), e53286.

D. Fricker, H. Zhang, and C. Yu. 2011. Sequential pattern mining of multimodal data streams in dyadic interactions. In *Proceedings of the IEEE Conference of Development and Learning*. 1–6.

Alexandra Frischen, Andrew P. Bayliss, and Steven P. Tipper. 2007. Gaze cueing of attention: Visual attention, social cognition, and individual differences. *Psychological Bulletin* 133, 4 (2007), 694.

Lakshmi J. Gogate. 2010. Learning of syllable–object relations by preverbal infants: The role of temporal synchrony and syllable distinctiveness. *Journal of Experimental Child Psychology* 105, 3 (2010), 178–197.

Lakshmi J. Gogate, Lorraine E. Bahrick, and Jilayne D. Watson. 2000. A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures. *Child Development* 71, 4 (2000), 878–894.

Michael H. Goldstein, Jennifer A. Schwade, and Marc H. Bornstein. 2009. The value of vocalizing: Five-month-old infants associate their own noncry vocalizations with responses from caregivers. *Child Development* 80, 3 (2009), 636–644.

Charles Goodwin. 1980. Restarts, pauses, and the achievement of a state of mutual gaze at turn-beginning. *Sociological Inquiry* 50, 3–4 (1980), 272–302.

David Gouaillier, Vincent Hugel, Pierre Blazevic, Chris Kilner, Jerome Monceaux, Pascal Lafourcade, Brice Marnier, Julien Serre, and Bruno Maisonnier. 2008. The Nao humanoid: A combination of performance and affordability. *CoRR Abs/0807*.3223.

Z. M. Griffin and K. Bock. 2000. What the eyes say about speaking. *Psychological Science* 11, 4 (2000), 274–279.

Koh Hosoda, Hidenobu Sumioka, Akio Morita, and Minoru Asada. 2004. Acquisition of human-robot joint attention through real-time natural interaction. In *Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'04)*. IEEE, 2867–2872.

Jukka Hyönä and Richard K. Olson. 1995. Eye fixation patterns among dyslexic and normal readers: Effects of word length and word frequency. J*ournal of Experimental Psychology: Learning, Memory, and Cognition* 21, 6 (1995), 1430.

Kristiina Jokinen and Graham Wilcock. 2014. Multimodal open-domain conversations with the Nao robot. In *Natural Interaction with Robots, Knowbots and Smartphones*. Springer, 213–224.

Masayuki Kamashima, Takayuki Kanda, Michita Imai, Tetsuo Ono, Daisuke Sakamoto, Hiroshi Ishiguro, and Yuichiro Anzai. 2004. Embodied cooperative behaviors by an autonomous humanoid robot. In *Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2506–2513.

A. Kendon and M. Cook. 1969. The consistency of gaze patterns in social interaction. *British Journal of Psychology* 60, 4 (1969), 481–494.

Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta Psychologica* 26 (1967), 22–63.

C. L. Kleinke. 1986. Gaze and eye contact: A research review. *Psychological Bulletin* 100, 1 (1986), 78–100.

Patricia K. Kuhl, Feng-Ming Tsao, and Huei-Mei Liu. 2003. Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences* 100, 9096–9101.

Chaoran Liu, Carlos T. Ishi, Hiroshi Ishiguro, and Norihiro Hagita. 2012. Generation of nodding, head tilting and eye gazing for human-robot dialogue interaction. In *Proceeding of the 2012 7th ACM/IEEE International Conference Human-Robot Interaction (HRI)*. IEEE, 285–292.

Simon P. Liversedge and John M. Findlay. 2000. Saccadic eye movements and cognition. *Trends in Cognitive Sciences* 4, 1 (2000), 6–14.

Caroline Lyon, Chrystopher L. Nehaniv, and Joe Saunders. 2012. Interactive language learning by robots: The transition from babbling to word forms. *PLoS One* 7, 6 (2012), e38236.

Davide Marocco, Angelo Cangelosi, Kerstin Fischer, and Tony Belpaeme. 2010. Grounding action words in the sensorimotor interaction with the world: Experiments with a simulated iCub humanoid robot. *Frontiers in Neurorobotics* 4 (2010), 7.

K. L. Marsh, M. J. Richardson, and R. C. Schmidt. 2009. Social connection through joint action and interpersonal coordination. *Topics in Cognitive Science* 1, 2 (2009), 320–339.

Ethel Matin. 1974. Saccadic suppression: A review and an analysis. *Psychological Bulletin* 81, 12 (1974), 899.

Raveesh Meena, Kristiina Jokinen, and Graham Wilcock. 2012. Integration of gestures and speech in human-robot interaction. In *Proceedings of the 2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)*. IEEE, 673–678.

P. Mundy and L. Newell. 2007. Attention, joint attention, and social cognition. *Current Directions in Psychological Science* 16, 5 (2007), 269–274.

B. Mutlu, T. Shiwa, T. Kanda, H. Ishiguro, and N. Hagita. 2009. Footing in human-robot conversations: How robots might shape participant roles using gaze cues. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction.* ACM, 61–68.

Bilge Mutlu, Fumitaka Yamaoka, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2009. Nonverbal leakage in robots: Communication of intentions through seemingly unintentional behavior. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction.* ACM, 69–76.

Yukie Nagai, Claudia Muhl, and Katharina J. Rohlfing. 2008. Toward designing a robot that learns actions from parental demonstrations. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'08).* IEEE, 3545–3550.

Yukie Nagai and Katharina J. Rohlfing. 2007. Can motionese tell infants and robots what to imitate. In *Proceedings of the 4th International Symposium on Imitation in Animals and Artifacts*. 299–306.

Yukiko I. Nakano, Gabe Reinstein, Tom Stocky, and Justine Cassell. 2003. Towards a model of face-to-face grounding. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, -Volume 1.* Association for Computational Linguistics, 553–561.

Tim Niem Ller, Alexander Ferrein, Gerhard Eckel, David Pirro, Patrick Podbregar, Tobias Kellner, Christof Rath, and Gerald Steinbauer. 2011. Providing ground-truth data for the Nao robot platform. In *RoboCup 2010: Robot Soccer World Cup XIV.* Springer, 133–144.

Karola Pitsch, Anna-Lisa Vollmer, Jannik Fritsch, Britta Wrede, Katharina Rohlfing, and Gerhard Sagerer. 2009. On the loop of action modification and the recipient's gaze in adult-child interaction. In *Proceedings of the International Conference on Speech and Gesture in Interaction (GESPIN)*.

Michael I. Posner. 1980. Orienting of attention. *Quarterly Journal of Experimental Psychology* 32, 3–25.

Charles Rich, Brett Ponsler, Aaron Holroyd, and Candace L. Sidner. 2010. Recognizing engagement in human-robot interaction. In *Proceedings of the 2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 375–382.

M. Rolf, M. Hanheide, and K. J. Rohlfing. 2009. Attention via synchrony: Making use of multimodal cues in social learning. *IEEE Transactions on Autonomous Mental Development* 1, 55–67.

Brian Scassellati. 1999. Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot. *Computation for Metaphors, Analogy, and Agents*, 176–195.

Atsushi Senju and Mark H. Johnson. 2009. The eye contact effect: Mechanisms and development. *Trends in Cognitive Sciences* 13, 3 (2009), 127–134.

Sara C. Sereno and Keith Rayner. 2003. Measuring word recognition in reading: Eye movements and event-related potentials. *Trends in Cognitive Sciences* 7, 11 (2003), 489–493.

Syamimi Shamsuddin, Hanafiah Yussof, Luthffi Ismail, Fazah Akhtar Hanapiah, Salina Mohamed, Hanizah Ali Piah, and Nur Ismarrubie Zahari. 2012. Initial response of autistic children in human-robot interaction therapy with humanoid robot NAO. In *Proceedings of the 2012 IEEE 8th International Colloquium on Signal Processing and Its Applications.* IEEE, 188–193.

Candace L. Sidner, Christopher Lee, Cory D. Kidd, Neal Lesh, and Charles Rich. 2005. *Explorations in Engagement for Humans and Robots. Artificial Intelligence* 166, 140–164.

Maria Staudte and Matthew W. Crocker. 2011. Investigating joint attention mechanisms through spoken human–robot interaction. *Cognition* 120, 2 (2011), 268–291.

Fumihide Tanaka and Shizuko Matsuzoe. 2012. Children teach a care-receiving robot to promote their learning: Field experiments in a classroom for vocabulary learning. *Journal of Human-Robot Interaction* 1, 1 (Aug. 2012).

Adriana Tapus, Maja J. Mataric, and B. Scasselati. 2007. Socially assistive robotics [grand challenges of robotics]. *Robotics & Automation Magazine, IEEE* 14, 1 (2007), 35–42.

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

Frances C. Volkmann. 1986. Human visual suppression. *Vision Research* 26, 9 (1986), 1401–1416.

Ning Wang and Jonathan Gratch. 2010. Don't just stare at me! In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1241–1250.

Tian Xu, Hui Zhang, and Chen Yu. 2013. Cooperative gazing behaviors in human multi-robot interaction. *Interaction Studies* 14, 3 (2013), 390–418.

A. Yamazaki, K. Yamazaki, Y. Kuno, M. Burdelski, M. Kawashima, and H. Kuzuoka. 2008. Precision timing in human-robot interaction: Coordination of head movement and utterance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 131–140.

Yuichiro Yoshikawa, Kazuhiko Shinozawa, Hiroshi Ishiguro, Norihiro Hagita, and Takanori Miyamoto. 2006. The effects of responsive eye movement and blinking behavior in a communication robot. In *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 4564–4569.

C. Yu, M. Scheutz, and P. Schermerhorn. 2010. Investigating multimodal real-time patterns of joint attention in an HRI word learning task. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 309–316.

C. Yu and L. B. Smith. 2012. Embodied attention and word learning by toddlers. *Cognition* 125, 2 (2012), 244–262.

Received December 2014; revised December 2015; accepted January 2016