Investigating Multimodal Real-Time Patterns of Joint Attention in an HRI Word Learning Task

Chen Yu
Department of Psychological and Brain Sciences
Cognitive Science Program
Indiana University
Bloomington, IN 47405
Email: chenyu@indiana.edu

Matthias Scheutz and Paul Schermerhorn
Human-Robot Interaction Laboratory
Cognitive Science Program
Indiana University
Bloomington, IN 47406
Email: {mscheutz,pscherme}@indiana.edu

***Abstract*—Joint attention – the idea that humans make inferences from observable behaviors of other humans by attending to the objects and events that these others humans attend to – has been recognized as a critical component in human-robot interactions. While various HRI studies showed that having robots to behave in ways that support human recognition of joint attention leads to better behavioral outcomes on the human side, there are no studies that investigate the detailed time course of *interactive joint attention processes*.**

In this paper, we present the results from an HRI study that investigates the exact time course of human multi-modal attentional processes during an HRI word learning task in an unprecedented way. Using novel data analysis techniques, we are able to demonstrate that the temporal details of human attentional behavior are critical for understanding human expectations of joint attention in HRI and that failing to do so can force humans into assuming unnatural behaviors.

***Keywords*-human-robot interaction; joint attention**

I. Introduction

Human social interactions can be very complex and comprise multiple levels of coordination, from high-level linguistic exchanges, to low-level couplings and decouplings of bodily movements. In particular, the temporal patterns of eye-gaze coordination between interacting humans, including mutual eye fixations as well as following gaze shifts to perceivable objects in the environment, play a critical role in establishment of mutual rapport and understanding, a mechanism generally referred to as "joint attention" [1].

The importance of joint attention for HRI has been recognized some time ago [2] and has recently led to interesting findings [3], promising architecture designs [4], and computational models of human development [5]. Yet, there is no HRI study that rigorously investigates the *exact time course of multi-modal human attention behaviors in social interaction tasks*, which is a prerequisite for fully understanding the nature of joint attention in HRI and ultimately for the development of robotic architectures that can lead to *natural HRI* [6].

In this paper, we investigate the temporal characteristics of joint attention processes in human-robot interactions in an experimentally unprecedented way. We use carefully controlled robot behaviors to collect detailed time-stamped multimodal data (including eye-tracking, visual and auditory data in naturalistic human-robot interactions) and apply novel data analysis methods to this rich data set to discover the exact time course patterns of human joint attention behaviors. The goal is to better understand the nature of human multi-modal coordination processes to be able to inform the design of HRI architectures in the future.

The paper is organized as follows. We start with a brief background and related work section making the case for a rigorous experimental study of joint attention behaviors. We then introduce our experimental paradigm and report results from an HRI experiment that demonstrates the complexity of temporal patterns of human eye gaze, motion and verbal behaviors. We conclude with a discussion of our findings and directions for future work.

II. Background and Related Work

The very idea of "joint attention" is that individuals can make inferences from the observable behaviors of others by attending to objects and events that others attend to. In its simplest form, joint attention requires the establishment of eye contact between two people, followed by one person intentionally breaking it to be able to look at an object of interest, followed by a subsequent fixation of that object by the other person. In its most general form, it is a "meeting of minds" that requires shared context and presuppositions (in addition to shared focus) [7]. In developmental psychology, there is convincing evidence that even young children are sensitive to gaze and pointing cues in child-parent interaction, and that they use those cues to follow the adult social partner's attention [8], which facilitates early cognition and learning [9], [10]. This prompted some to study the dynamics of joint attention processes and behavioral couplings in simulations (e.g., [11]).

In HRI, joint attention processes have been recognized as a critical component in human-robot interactions for quite some time. Early efforts focussed on building architectural mechanisms of joint attention [2], [12], [13].

Recently, several HRI studies evaluated joint attention behavior. [14] 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.

[3] 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.

[15] performed experiments with a guide robot designed using 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.

Finally, [16] studied the extent to which eye-gaze behavior of the robot could signal "participant roles" (in a conversation) to human observers and confirmed that subjects' behaviors conformed to the communicated roles.

While the first two studies found overall improvement in human comprehension and the other two studies confirmed typical joint attention behavior of humans in some robot conditions, none of the studies investigated *interactive* attention processes, e.g., how humans react to the robot's reaction to the human shift in eye gaze. For this kind of investigation, the robot's behavior cannot be scripted in advance (as in the above studies). Rather, the robot must be able to generate real-time behaviors based on real-time perception of human behavior.

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 a robot 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 HRI; in the worst case, however, it could lead to human lack of interest and trust, frustration and anxiety, and possibly resentment of the robot.

III. Experiment

The overall goal of our study is to investigate the exact time course of multi-modal interaction patterns that occur naturally as part of joint attention processes in human-robot interactions. We employed a *word learning task* where the human participants were asked to teach the robot the names of a set of objects. We selected this task for five reasons: (1) it has an explicit goal that allows participants to naturally engage with the robot in interactions while being constrained enough to make real-time processing on the robot's actions feasible, which in turn allows for adaptive robot behavior; (2) it has been used successfully in a variety of developmental studies investigating multi-modal human-human interactions (e.g., between parents and their children [17]); (3) it allows us to investigate the fine-grained temporal patterns and relationships between human eye gaze and human speech as part of the larger joint attention processes; (4) beyond modeling human interactions, the task itself has its own merits as it can help shed light on how robots might acquire new knowledge through human-robot social interaction [18];[1] and (5) it can ultimately be used to develop cognitive models of temporal interaction patterns that in an unprecedented way capture the time course of human-human interactions (cp. to [5]).

The experimental paradigm is noteworthy in that it is:

- *Multimodal:* Participants and the robot interact through speech and visual cues (including perceivable information from vision, speech, and eye gaze).
- *Interactive and adaptive:* The robot can follow what the human is visually attending to (based on real-time tracking of human eye gaze) and thus provide visual feedback to human subjects who can (and will) adjust their behavior in response to the robot's response.
- *Real-time:* The robot's actions are generated in real time as participants switch their visual attention moment by moment.
- *Naturalistic:* there are no constraints on what participants should or should not do or say in the task.

We defined two experimental conditions: 1) the *following condition* in which the robot monitors the participant's gaze in real time and then moves its head to look toward the same location that the participant is visually attending to; and 2) the *random condition* in which the robot completely ignores the participant's behaviors and instead generates random movements during the whole interaction. Note that compared with descriptive statistics that one can compute from observing natural human-robot interaction, our experimental manipulation with two conditions provides a more systematic way to study and quantify real-time human-robot interaction.

Will participants in the *random condition* pay *overall more attention* to the robot compared with participants in the *following condition*? Moreover, in addition to a comparison of overall eye movement patterns in the two conditions, the more interesting research question, is *how* participants will differ in the two conditions at a fine-grained level. For example, subjects might spend more time attending to the robot with longer eye fixations in the random condition. They also might spend more time in attracting the robot's attention before naming objects, and therefore, generate fewer naming utterances. Alternatively, they might more frequently monitor the robot's attention with shorter eye fixations and generate more naming utterances in order to attract the robot'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 robot or more at the named object? Will their eye movement patterns change over the course of naming speech production? Will those in the following condition generate different eye movement patterns compared with people in the random condition? And if so, in what ways?

To be able to answer these questions, we collected and

[1]For example, it can help answer the question of how robots should adjust their behaviors in ways that may encourage humans to generate better teaching behaviors to facilitate communication and learning.

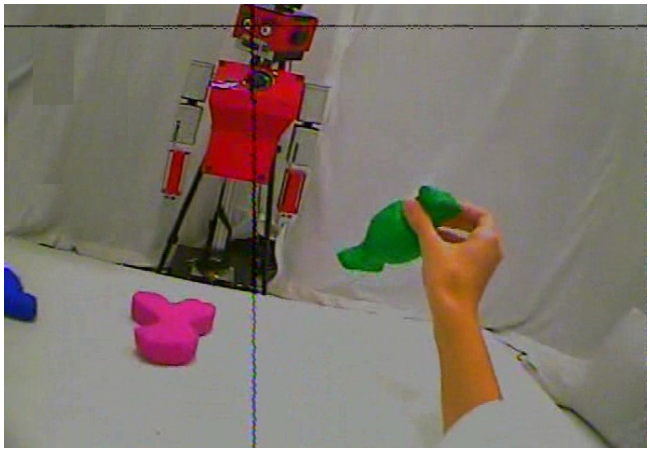

Fig. 1. A snapshot from the participant's first-person view, sitting across a table from a robot, trying to teach the robot object names. The cross-hair indicates the participant's eye gaze at this moment. In this example, the robot was not following the participant's attention ("random condition").

analyzed fine-grained multimodal behavioral data that would allow us to discover the time course of sensorimotor patterns.

A. Participants

25 undergraduate students at Indiana University participated in the study (4 of them were excluded due to technical problems with their eye tracking).

B. Experimental Setup

The experimental setup is depicted in Figure 1, with a human subject sitting across a table from a robot. 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 colors "blue", "pink", and "green". Objects are unique in each set in terms of shapes and names. The object names are displayed on labels that are attached to the drawers containing the objects and are visible to the participant during the experimental run.

The employed robot is humanoid torso with 2 DoF (degrees of freedom) movable head and two 3 DoF arms. The robot's head includes two 2 DoF "eyes" with integrated Firewire cameras, two 1 DoF movable eye-brows and two 2 DoF movable lips. In the present experiment, only the head was actuated. The robot's chest also includes integrated microphones and speakers, but only the microphones were used for recording (no speech recognition was performed). A single Linux PC (Intel Pentium 4, 2.80Ghz) was used for running all robotic control software implemented in our `DIARC` architecture [6] (which has been used successfully for many HRI studies in our lab, e.g., [19]–[21]).

An additional camera was mounted behind the robot to provide a static view of the visual scene regardless of the robot's head orientation or head movement. This static camera was used in the "follow condition" together with real-time eye tracking information from the eye-tracker worn by the human to determine the object the human looked at. Specifically, color blob detection was performed on the image from the first-person view camera mounted on the eye tracker to determine whether the human looked at any of the colored objects or at the robot (based on superimposing the cross-hair human eye fixation data on the image from the head cam). The corresponding colored object could then be determined in the static camera's image based on color correspondence alone (unless the human fixated on the robot). A head motion was initiated to center the head on the fixated object (or the person's face if the human was looking at the robot) based on real-time gaze data from the participant. In the random condition, the robot only executed pre-programmed random sequences of head movements followed by short pauses where the head remained stationary for a randomly determined amount of time. Thus, the robot completely ignored the participant's behaviors.

C. Procedure

To facilitate visual processing, participants were asked to 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 robot. 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 robot 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 robotic architecture was started, and the experimenter instructed the subject to start. The subject was allowed one minute to teach the robot the names of the three objects in a set (see Figure 1). 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=4.21s, SD=0.25).

D. Data 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 robot, 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.

a) 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 skin color to detect hand blobs and red color to detect the robot in view.

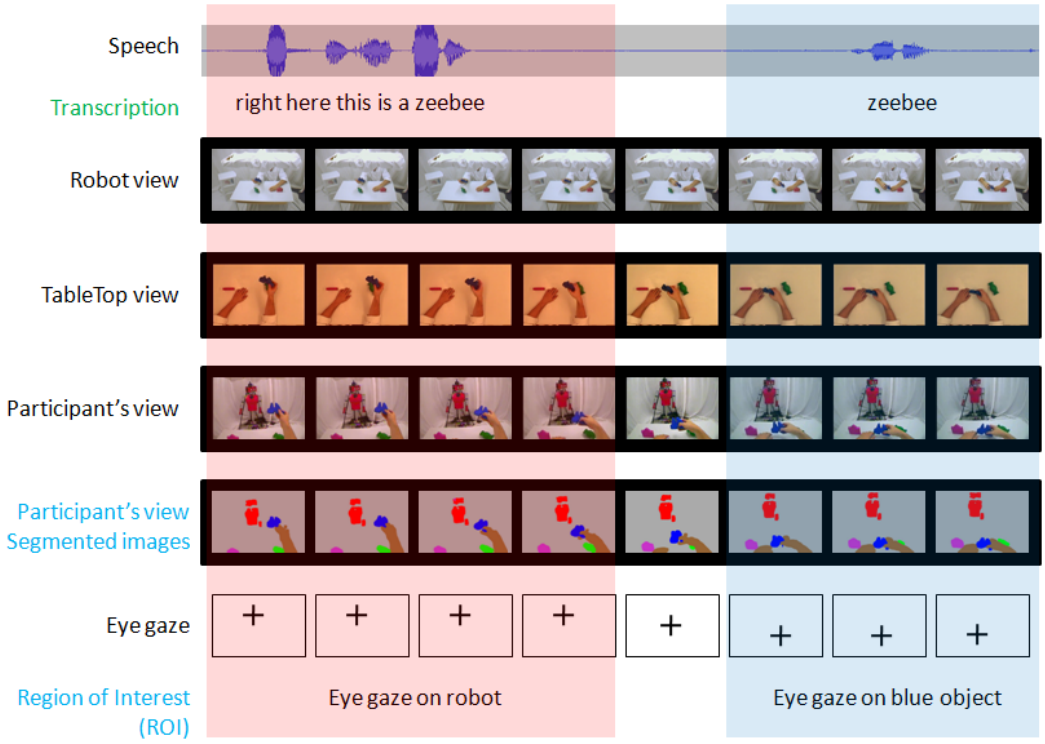

Fig. 2. In our experiment, the collected and processed multi-streaming multi-modal data included speech, multiple video streams from different cameras, and eye gaze. The combination of visual processing and gaze information defined temporal Region-of-Interest (ROI) events (highlighted in the figure).

As shown in Figure 2, five Regions Of Interest (ROIs) were defined and calculated frame by frame: the robot, the three objects, and the subject's hands. In practice, we decomposed the whole video from the participant's view into an image sequence and ran the color detection program on each image to find the 5 ROIs.

b) Gaze data: We computed eye fixations using a velocity-based method to convert continuous eye movement data into a set of fixation events (see details in [10]) . For each fixation, we superimposed (x,y) coordinates of eye gaze onto the image sequence of the first-person view to calculate the corresponding ROIs moment by moment.

c) 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 the following subsections, we first report analyses of eye movement data from participants of both groups and then report analyses of their adaptive behaviors in speech acts. We integrated speech and eye movement data, zoomed into the moments when they named objects and extracted dynamic time-course patterns around those naming moments. All of the following statistics and analyses were based on the whole data set across 21 subjects (11 in the following condition and 10 in the random condition), containing about 50,000 image frames, 190,000 gaze data points, 4590 eye fixations, 2585 spoken utterances, and 1053 naming utterances.

TABLE I
EYE MOVEMENT MEASURES AVERAGED ACROSS TRIALS (APPROXIMATELY 1 MINUTE PER TRIAL, ** INDICATING STATISTICAL SIGNIFICANCE, P<.001, T-TEST, TWO-TAILED)

	following	random
number of attention switches (eye fixations)	53.61	55.8
average fixation length in seconds	0.96	1.16
number of robot looking fixations	22.32	21.75
average length of robot fixations in seconds (**)	1.11	1.72
longest fixation in seconds (**)	3.66	5.92

E. Analyses of Eye Movement Data

As shown in Table I, the overall eye movement patterns from the two experimental conditions are similar in terms of the total number of eye fixations and the average length of those fixations. Moreover, human participants in both groups generated similar gaze behaviors toward the robot. For example, the total number of eye fixations on the robot is comparable between two groups. However, a closer examine of gaze data revealed the differences between the two experimental conditions. Even though the total number of fixations on the robot is similar, participants in the random group generated longer fixations (M=1.72 sec per fixation) than those in the following group (M=1.11sec). Taken together, the results suggest that

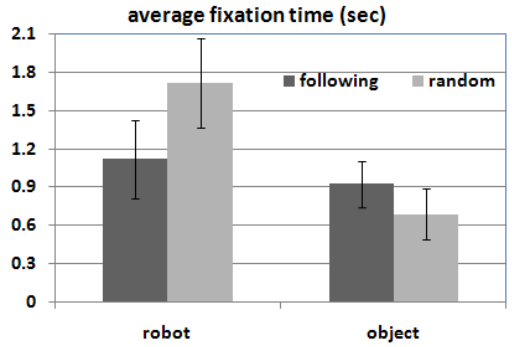

Fig. 3. A comparison of average fixation times on the robot or objects in the two experimental conditions.

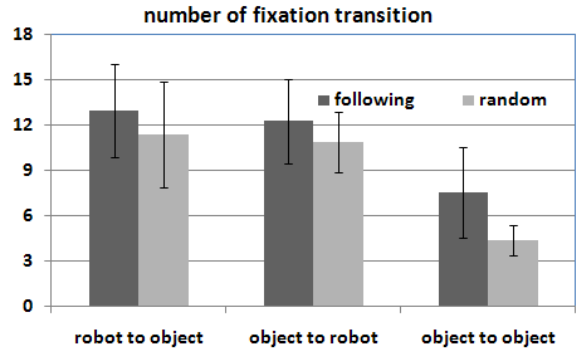

Fig. 4. The average numbers of fixation transitions (saccades, etc.) within a trial. Three transition types in the two conditions were calculated and compared: from robot to object, from object to robot, and from object to object.

participants in the random condition visually attended to the robot significantly longer (through longer eye fixations) than to objects and also longer than those in the following group (see Figure 3). In contrast, there is no significant difference in average fixation of the robot or objects in the following condition. Moreover, participants in the following condition looked slightly longer at visual objects compared with those in the random condition. Hence, the most interesting difference between the random and following groups with respect to eye movement patterns is the increasing attention on the robot in the random group where the robot and participants were not in the joint attention state.

Figure 4 reports transitional eye fixations (saccades, etc.) from one ROI to the next ROI. Three different kinds of transition were measured: 1) from robot to object; 2) from object to robot; and 3) from object to object. We found that participants produced a similar number of attention switches between the robot and objects in the two experimental conditions. However, there were significantly more attention switches between objects in the following condition. Since we also know that the total number of attention switches is similar in the two conditions, this suggests that participants in the following condition were more likely to switch visual attention between objects while those in the random condition needed to check the robot's visual attention very frequently and therefore rarely switch gaze directly between objects without monitoring (going back to) the robot.

In sum, the analyses of eye movement data suggest that 1) humans were sensitive to the differences in the robot's behaviors in the experimental conditions; 2) they adjusted their behaviors accordingly by spending more time on keeping track of the robot's attention in the random condition; and 3) although the overall eye movement patterns were (more or less) similar between two groups, fine-grained data analyses discovered various significant differences, ranging from the duration of the longest fixation, to the proportion of time looking at the robot, to the transitional gaze movements between visual objects.

F. Analyses of Speech Acts

We first calculated a few basic statistics from participant's speech acts (see Table II). On average, participants in the two conditions produced a similar number of distinct word types ($M_{following}$ = 88; M_{random} = 86) and a similar number of spoken utterances ($M_{following}$ = 114; M_{random} = 121). However, participants in the random condition generated more tokens [2] (M_{random} = 459) than those in the following condition ($M_{following}$ = 349) and they also generated longer spoken utterances (3.79 words per utterance) than the other group (3.31 words per utterance). Although these differences are statistically significant (t-test, p<.01), they do not seem useful for extracting general principles for HRI designs. More interesting are naming utterances (containing object names) as participants in the random condition produced significantly more than those in the following group (60 versus 48); and this is not simply because participants in the random produced more speech (as we already stated). In fact, the proportion of naming utterances from participants in the random condition (0.50) is also much higher than from those in the following group (0.42). Together, both the total number of utterances and the proportion of naming utterances contribute to the overall higher number of naming speech acts in the random condition.

TABLE II
OVERALL STATISTICS OF VOCABULARY (** INDICATING STATISTICAL SIGNIFICANCE P<0.001; * INDICATING P<.01, T-TEST, TWO-TAILED).

	following	random
number of words	88	86
number of tokens (*)	394	459
number of utterances	114	121
words per utterance (*)	3.31	3.79
number of naming utterances (**)	48	60
proportion of naming utterances (**)	0.42	0.50

This raises the question of what else participants produced in their speech and what the potential differences were be-

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

TABLE III
AVERAGE OCCURRING FREQUENCIES OF SELECTED WORDS

	following	random
"look"	2.12	6.33
"see"	2.23	7.78
"here"	2.92	11.12
"robot"	0.33	3.44
"hey"	0	1.78
"yes"	1.10	2.83
object names	48.78	60.44

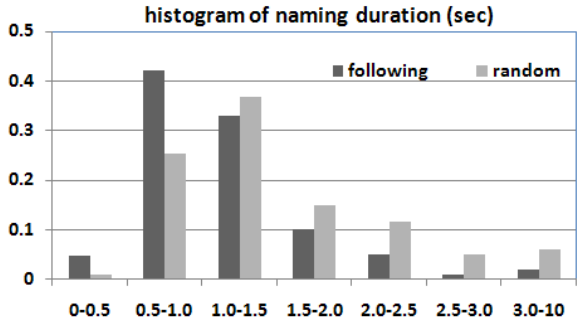

Fig. 5. The histogram of durations of naming utterances in the experimental conditions

tween the two conditions? Hence, we next calculated the histogram of all spoken words in each group separately and compared the two lists to be able to detect words with different frequencies. Table III shows a subset of those words. Participants in the random condition produced words like "look", "see", "hey" and "here" much more frequently, clearly to attract the robot's attention. Taken together with the results from Table II, we can conclude that participants in the random condition generated more attention-attracting utterances and more naming utterances than participants in the following condition, possibly because the latter did not (need to) do so since the robot was always following the participant's visual attention. It is possible that more naming and more attention-attracting utterances in the random condition could benefit communication and learning in at least two ways. First, since participants noticed that the robot was not visually attending to their behaviors and that it was difficult to attract/control the robot's attention through vision and manual action, they attempted to use speech acts from the auditory modality to better attract the robot's attention. Moreover, from the language learning perspective, participants might have realized that learning situations were not ideal since the robot did not pay attention to the objects when they named them. In order to overcome this, they chose to name those objects more frequently with the hope that more (imperfect) naming instances would increase the chance of successful learning even though each individual naming instance was not ideal.

But did people in the two conditions name objects in different ways? Figure 5 shows the histograms of the durations of naming utterances in two conditions. Participants in the random condition tended to generate longer naming utterances while naming utterances from those in the following conditions contained more than 70% of shorter ones (<1.5 sec). Shorter spoken utterances might be due to two factors: fewer words in the utterance or a lower speech rate. We calculated the speech rate and found that there was no difference between two groups ($m_{following}$ = 1.72 word/sec, m_{random} = 1.67 word/sec). However, the number of words in naming utterances in two conditions differ – participants in the following condition generated almost 50% single-word naming utterances (i.e., only object names). If the robot is following the human teacher's attention (as in the following condition), a human teacher can simply utter the object name at the moment that both the robot and the teacher are jointly attending to that object, a straightforward teaching strategy requiring no complicated syntactic structures to attract the robot's attention. Our results confirm that participants in the following condition clearly adapted to this teaching strategy. Meanwhile, participants in the random group realized that the robot was not paying attention and therefore produced more attention-attracting speech, naming those objects using sentence structures instead of single words. In sum, participants in both conditions were sensitive to the robot's behaviors and adjusted their own speech acts accordingly.

G. Analyses of Temporal Dynamics in Multimodal Data

In an effort to better understand the dynamic processes that lead up to naming events, we focused on the moments just before and just after a naming utterance and measured participants' visual attention as a way to integrate speech data with gaze data. The approach is based on what has been used in psycholinguistic studies to capture temporal profiles across a related class of events [22] (in our case, the relevant class of events is a naming event). Such profiles enable one to discern potentially important temporal moments within a trajectory and compare temporal trends across trajectories. Figure 6 shows the average proportions of time across all naming events (which can be viewed as a probability profile) that participants looked at the robot, the target object, the other two objects, or their own hands. Thus, each trajectory in a plot shows the probability that participants looked at one of four identities either for the 10 sec prior to naming events (the left two plots), or during naming events (the middle two plots, a temporal window of 1.5 sec was used since name utterances had different lengths), or the 10 sec after naming events (the right two plots). We observe several interesting patterns. First, participants in the random condition spent around 70% of their time on gazing at the robot both before, during and after naming events. Moreover, there seems to be no difference among those moments, suggesting that those participants looked at the robot all the time. Second, even in the following condition, participants kept track of the robot's attention and as a result, looked more at the robot than the target object. This is a surprising result since we know from previous studies on object naming in face-to-face interaction (see a review in [23]) that speakers are likely to look at the

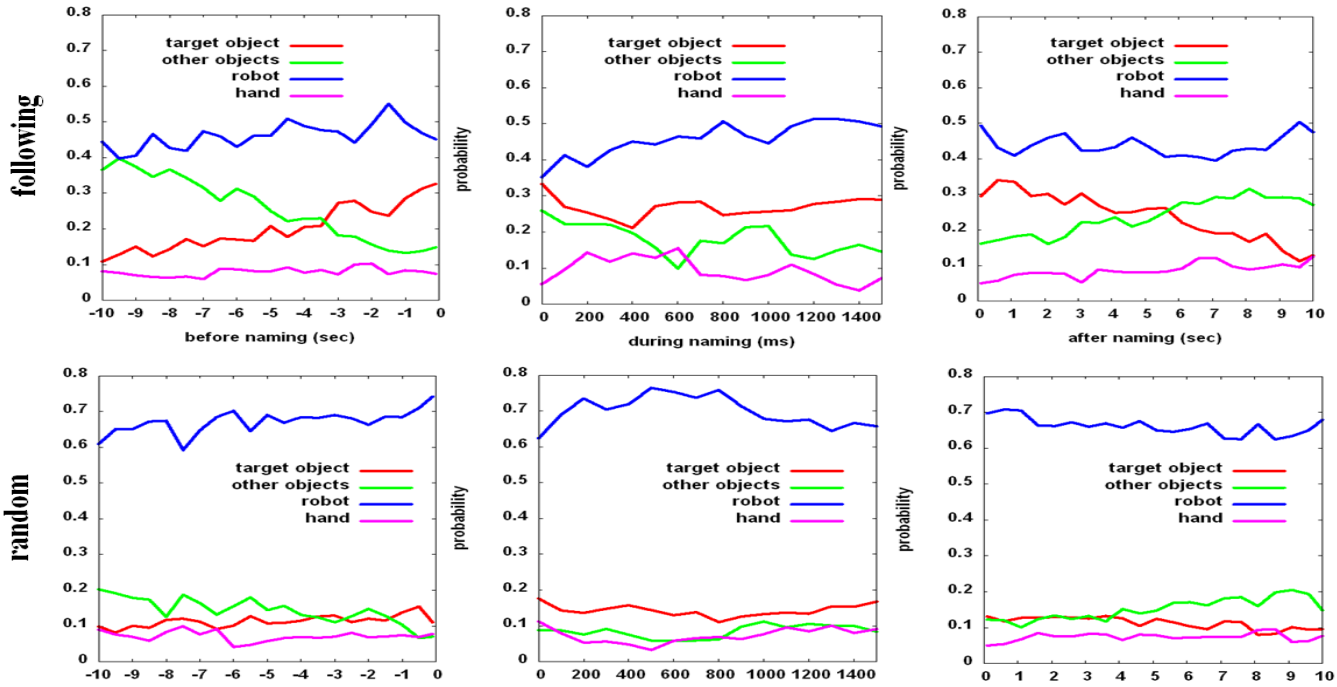

Fig. 6. The proportion of time that participants were looking at the robot, the named target object, the other two objects, or their own hands before, during and after a naming utterance. The top three plots were calculated from naming instances of participants in the following condition and the bottom three were derived from participants in the random condition. The left two plots are before naming events; the middle two during naming events, and the right two after naming events.

target object when they produced the object name. Our results thus show that maintaining and monitoring joint attention in face-to-face human-robot interaction significantly changes participants' visual attention: they looked at the social partner more regularly than face-to-face human-human interaction with referential communication tasks [23], and they did so even at the moments of object naming. Third, also in the following condition, we observe increased gaze at the to-be-named object even before the naming utterance appeared. Similarly, participants' visual attention on the target object decreased after the naming event. Taken together, at those moments around a naming event, they paid more attention on the named object than other objects on the table. This particular gaze pattern between target and other objects is in line with the results from psycholinguistic studies on the coupling between speech and gaze [23]. However, this pattern was *not replicated in the random condition*. Besides the fact that participants in the random condition spent a significant amount of looking time on the robot, the rest of their attention was more or less equally distributed between target and other objects. This suggests that participants perceived the robot'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. This is of importance for HRI and requires further investigation as current robots are likely to violate subtle temporal patterns in joint attention processes and thus the time course of attention processes that humans expect. This finding also serves as justification for pursuing a temporally fine-grained multi-modal analysis of human joint attention processes in HRI.

Finally, we also extracted more fine-grained dynamic patterns around naming events. For example, even 4 seconds prior to a naming event in the following condition (the upper left plot in Figure 6), participants started looking at the target object much more than the combined looking time on the other two objects. Between 2500 and 1500 milliseconds prior to naming, there was an increase of attention toward the robot, and that increasing attention to the robot stopped and then dropped after 1500 ms prior to a naming event. Instead, participants in the following condition paid more attention to to-be-named objects from then on until to the moment right before a naming utterance was produced.

IV. Conclusions and Future Work

In this paper, we demonstrated the utility of analyzing the time course of human-robot interactions in an effort to reveal human moment-to-moment behaviors. The particular focus was on detecting critical phases in human multi-modal joint attention processes. For this purpose, we conducted HRI experiments in a naturalistic word learning task where a human instructor had to teach a robot the names of new objects. In one condition, the robot followed human eye gaze based on real-time processing of human eye fixations implementing a simple form of joint attention behavior, while in the other condition the robot exhibited random head movements that were indicative of distraction or loss of attentional focus.

We collected detailed multi-modal time course data from the experiments and use it to analyze the differences in human reactions to the two conditions.

The results confirmed our general prediction, that humans paid more attention to the robot in the random condition compared to the following condition in an attempt to get the robot's attention so that they could teach the robot the names of the objects. However, the more interesting results were revealed at a finer-grained temporal level of multi-modal interactions investigating eye-fixations before, during, and after naming utterances (among other analyses). We found that humans exhibit similar human joint attention behavior in the following condition to what happens in human-human face-to-face interaction, but unnatural behavior with respect to eye fixations as well as their coupling with naming events in the random condition. This finding is of high significance, especially if the goal is to aim for *natural* HRI.

Future work, thus, can use our experimental findings and apply them to the design of better HRI architectures. For example, we can predict based on our data what object participants will name based on their eye gaze 3 to 4 seconds prior to the actual naming event. Hence, a robot that can detect human eye gaze could be programmed to focus its attention on that object as soon as the human intention is discovered. This will likely result in better coordination between humans and robots, and also reduce human cognitive effort (given that subjects in the following condition were using significantly simpler linguistic expressions). Moreover, by studying the details of eye gaze and naming coupling we will be able to design behavioral scripts for learning robots that will allow robots to assume the role of a human teacher that is intuitive and easy to follow for human learners. Overall, we believe that the kind of empirically-grounded and sensorimotor-based study of human-robot 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 HRI.

ACKNOWLEDGMENT

The authors would like to thank You-Wei Cheah and Amanda Favata for working on data collection, and Thomas Smith, Ruj Akavipat, and Ikhyun Park for working on data preprocessing. This work was supported by a grant from NSF BCS0924248 to the first author.

REFERENCES

[1] C. Moore and P. J. Dunham, Eds., *Joint Attention: Its Origins and Role in Development*. Lawrence Erlbaum Associates, 1995.

[2] B. Scassellati, "Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot," in *Computation for Metaphors, Analogy, and Agents*, C. Nehaniv, Ed. Berlin/Heidelberg: Springer-Verlag, 1999, vol. 1562, pp. 176–195.

[3] M. Staudte and M. Crocker, "Visual attention in spoken human-robot interaction," in *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. ACM New York, NY, USA, 2009, pp. 77–84.

[4] Y. Okuno, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita, "Providing route directions: design of robot's utterance, gesture, and timing," in *HRI '09: Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. New York, NY, USA: ACM, 2009, pp. 53–60.

[5] J. G. Trafton, B. Fransen, A. M. Harrison, and M. Bugajska, "An embodied model of infant gaze-following," in *International Conference on Cognitive Modeling*, 2009.

[6] M. Scheutz, P. Schermerhorn, J. Kramer, and D. Anderson, "First steps toward natural human-like HRI," *Autonomous Robots*, vol. 22, no. 4, pp. 411–423, May 2007.

[7] J. Bruner, "Foreword from joint attention to the meeting of minds: An introduction," in *Joint Attention: Its Origins and Role in Development*, C. Moore and P. J. Dunham, Eds. Lawrence Erlbaum Associates, 1995, pp. 1–14.

[8] M. Carpenter, K. Nagell, M. Tomasello, G. Butterworth, and C. Moore, "Social cognition, joint attention, and communicative competence from 9 to 15 months of age," *Monographs of the society for research in child development*, 1998.

[9] A. Meltzoff, P. Kuhl, J. Movellan, and T. Sejnowski, "Foundations for a New Science of Learning," *science*, vol. 325, no. 5938, p. 284, 2009.

[10] C. Yu, D. Ballard, and R. Aslin, "The role of embodied intention in early lexical acquisition," *Cognitive Science: A Multidisciplinary Journal*, vol. 29, no. 6, pp. 961–1005, 2005.

[11] T. Ikegami and H. Iizuka, "Joint attention and dynamics repertoire in coupled dynamical recognizers," in *Proceedings of AISB*, 2003, pp. 125–130.

[12] M. Imai, T. Ono, and H. Ishiguro, "Physical relation and expression: Joint attention for human-robot interaction," in *Proceedings of the 10th IEEE International Workshop on Robot and Human Communication*, 2001.

[13] Y. Nagai, K. Hosoda, A. Morita, and M. Asada, "A constructive model for the development of joint attention," *Connection Science*, vol. 15, no. 4, pp. 211–229, 2003.

[14] B. Mutlu, J. Forlizzi, and J. Hodgins, "A storytelling robot: Modeling and evaluation of human-like gaze behavior," in *6th IEEE-RAS International Conference on Humanoid Robots*, 2006, pp. 518–523.

[15] A. Yamazaki, K. Yamazaki, Y. Kuno, M. Burdelski, M. Kawashima, and H. Kuzuoka, "Precision timing in human-robot interaction: coordination of head movement and utterance," in *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*. New York, NY, USA: ACM, 2008, pp. 131–140.

[16] B. Mutlu, T. Shiwa, T. Kanda, H. Ishiguro, and N. Hagita, "Footing in human-robot conversations: how robots might shape participant roles using gaze cues," in *HRI '09: Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. New York, NY, USA: ACM, 2009, pp. 61–68.

[17] C. Yu, L. Smith, H. Shen, A. Pereira, and T. Smith, "Active Information Selection: Visual Attention Through the Hands." *IEEE Transactions on Autonomous Mental Development*, in press.

[18] C. Breazeal, G. Hoffman, and A. Lockerd, "Teaching and working with robots as a collaboration," in *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 3*. IEEE Computer Society Washington, DC, USA, 2004, pp. 1030–1037.

[19] P. Schermerhorn, M. Scheutz, and C. R. Crowell, "Robot social presence and gender: Do females view robots differently than males?" in *Proceedings of the Third ACM IEEE International Conference on Human-Robot Interaction*, Amsterdam, NL, March 2008.

[20] T. Brick and M. Scheutz, "Incremental natural language processing for HRI," in *Proceedings of the Second ACM IEEE International Conference on Human-Robot Interaction*, Washington D.C., March 2007, pp. 263–270.

[21] M. Scheutz, P. Schermerhorn, J. Kramer, and C. Middendorff, "The utility of affect expression in natural language interactions in joint human-robot tasks," in *Proceedings of the 1st ACM International Conference on Human-Robot Interaction*, 2006, pp. 226–233.

[22] P. Allopenna, J. Magnuson, and M. Tanenhaus, "Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models," *Journal of Memory and Language*, vol. 38, no. 4, pp. 419–439, 1998.

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