# EXPLORATIONS IN ENGAGEMENT FOR HUMANS AND ROBOTS

CANDACE SIDNER,[1] CHRISTOPHER LEE,[1] CORY KIDD,[2] and NEAL LESH[1]

[1] *Mitsubishi Electric Research Laboratories*
*201 Broadway, Cambridge, MA 02139 USA*
*{sidner,lee,lesh}@merl.com*

[2] *The Media Laboratory*
*Massachusetts Institute of Technology*
*77 Massachusetts Avenue*
*Cambridge, MA 02139 USA*
*coryk@media.mit.edu*

This paper explores the concept of engagement, the process by which individuals in an interaction start, maintain and end their perceived connection to one another. The paper reports on one aspect of engagement among human interactors—the effects of tracking faces during an interaction. It also provides details for an architecture of a robot that can participate in conversational, collaborative interactions with engagement gestures. Finally the paper reports on findings of the effects on human participants who interacted with a robot when it either performed or did not perform engagement gestures. Results of the human-robot studies indicate that people become engaged with robots: they direct their attention to the robot more often in interactions where these gestures are present, and they find these gestures more appropriate than when they are not present.

*Keywords*: engagement; human-robot interaction; conversation; gestures

## 1. Introduction

When individuals meet face-to-face to interact with one another, they do so by means of gestures and conversation to begin their interaction, to maintain and accomplish something during it, and to end it. Engagement is the process by which interactors start, maintain and end their perceived connection to each other during an interaction. It combines verbal communication with non-verbal behaviors, all of which support the perception of connectedness between interactors. While the verbal channel provides detailed and rich semantic information as well as social connection, the non-verbal channel can be used to provide information about what has been understood so far, what the interactors are each, or together, attending to, evidence of their waning connectedness, and evidence of their desire to disengage. Evidence of the significance of engagement can be seen when engagement behaviors conflict, as for example happens when the dialogue behavior indicates that interactors are engaged through turn taking, conveying intentions and the like, but

when one or more of them looks away for long periods to free space or objects that have nothing to do with the dialogue. This paper explores the idea that engagement is central to human-robot interaction just as it is for human-human interaction.

Engagement is not well understood in the human to human context, in part because it has only been pursued in the sociological and psychological communities as part of general communication studies, and in part because in artificial intelligence, much of the focus has been on language understanding and production, rather than on the fundamental problems of how to get started and stay connected and the role of gesture in connecting. Only with the advent of conversational 2D characters and better vision technology have these issues begun to come forward (see Traum[32] and Nikano[23] for examples of 2D agents where these issues are relevant).

The methodology applied in this work has been to study human-human interaction and then apply the results to human-robot interaction, with a focus on hosting activities. Hosting activities are a class of collaborative activity in which an agent provides guidance in the form of information, entertainment, education or other services in the user's environment and may also request that the user undertake actions to support the fulfillment of those services. Hosting activities are situated or embedded activities, because they depend on the surrounding environment as well as the participants involved. Hosting activities are modelled using the collaboration and conversation models of Grosz and Sidner,[13] Grosz and Kraus,[12] and Lochbaum,[19] and distinguished from interactions between competitors, enemies or agents who are only cooperating (where shared goals are not held). This work defines interaction as an encounter between two or more individuals during which at least one of the individuals has a purpose for encountering the others. Interactions often include conversation although it is possible to have an interaction where nothing is said linguistically. Collaborative interactions are those in which individuals come to have shared goals and intend to carry out activities to attain the shared goals. This work focuses on interactions between two individuals.

Our hypothesis for this work concerned the effects of engagement gestures during conversation. We predicted that human partners would respond with corresponding looking gestures when the robot looked at and away from the human partner in appropriate ways. That is, a robot with looking gestures and one that had no such gestures would differentially affect how the human judged that experience. The first part of this paper investigates the nature of looking gestures in human-human interactions. The paper then explains how we built a robot to approximate the human behavior for engagement in conversation. Finally, the paper reports on an experiment where human partners interacted with a robot with looking gestures and one where no such gestures were present. A part of that experiment has been to discover measures for evaluating the behavior of the human partner.

| | Count | *Percentage of:* | |
|---|---|---|---|
| | | Tracking failures | Total host looks |
| Quick looks | 11 | 30% | 13% |
| Nods | 14 | 38% | 17% |
| Uncategorized | 12 | 32% | 15% |

Table 1. Failures of Visitor to track changes in Host's looking during a conversation.

## 2. Human-human engagement: results of video analysis

This section summarizes our work on human-human engagement.[31] Using videotaped interactions of two people in a hosting situation, we transcribed portions of the video for all the utterances made and some of the gestures (head, body position, body addressing) that occurred. We then considered one behavior in detail, namely mutual face tracking of the participants, as evidence of their focus of interest and engagement in the interaction. The purpose of the study was to determine how well the visitor (V) in the hosting situation tracked the head motion of the host (H), and to characterize the instances when V failed to track H.[a]

While it is not possible to draw conclusions about all human behavior from a single pair interaction, even a single pair provides an important insight into the kinds of behavior that can occur.

In this study we assumed that the listener would track the speaker almost all the time, in order to convey engagement and use non-verbal as well as verbal information for understanding. There was no literature to suggest what in fact was the case. In our study the visitor is the listener in more than 90% of the interaction (which is not the normal case in conversations).[b]

To summarize, there are 82 instances where the (male) person acting as host (H) changed his head position, as an indication of changes in looking, during a five minute conversational exchange with the person who was the (female) visitor (V). Seven additional changes in looking were not counted because it was not clear to where the host turned. Of his 82 counted changes in looking, V tracks 45 of them (55%). The remaining failures to track looks (37, or 45% of all looks) can be subclassed into 3 groups: "quick looks" (11), "nods" (14), and uncategorized failures (12), as shown in Table 1. The "quick look" cases are those for which V fails to track a look that lasts for less than a second, and the "nod" cases are those for which V nods (e.g., as an acknowledgement of what is being said) rather than tracking H's look.

The quick look cases happen when V fails to notice H's look due to some other

[a]We say that V "tracks H's changes in looking" if: when H looks at V, then V looks back at H; and when H looks elsewhere, V looks toward the same part of the environment as H.

[b]The visitor says only 15 utterances other than 43 backchannels (for example, ok, ah-hah, yes, and wow) during 5 minutes and 14 seconds of dialogue. Even the visitor's utterances are brief, for example, *absolutely, that's very stylish, it's not a problem.*

activity, or because the look occurs in mid-utterance and does not seem to otherwise affect his utterance. In only one instance does H pause intonationally and look at V. One would expect an acknowledgement of some kind from V here, even if she doesn't track H's look, as is the case with nod failures. However, H proceeds even without the expected feedback.

The nod cases can be explained because they occur when H looks at V even though V is looking at something else. In all these instances, H closes an intonation phase, either during his look or a few words after, to which V nods and often articulates with "Mm-hm," "Wow" or other phrases to indicate that she is following her conversational partner. In grounding terms,[7] H is attempting to ascertain by looking at V that she is following his utterances and actions. When V cannot look, she provides feedback by nods and comments. She is able to do this because of linguistic (that is, prosodic) information from H indicating that her contribution is called for.

Of the uncategorized failures, the majority (8 instances) occur when V has other actions or goals to undertake. In addition, all of the uncategorized failures are longer in duration than quick looks (2 seconds or more). For example, V may be finishing a nod and not be able to track H while she's nodding. Of the remaining three tracking failures, each occurs for seemingly good reasons to video observers, but the host and visitor may or may not have been aware of these reasons at the time of occurrence. For example, one failure occurs at the start of the hosting interaction when V is looking at the new (to her) object that H displays and does not track H when he looks up at her.

Experience from this data has resulted in the *principle of conversational tracking*: participants in a collaborative conversation track the other's face during the conversation in balance with the requirement to look away to: (1) participate in actions relevant to the collaboration, or (2) multi-task activities unrelated to the collaboration at hand, such as scanning the surrounding scene for interest, avoidance of damaging encounters, or personal activities.

The above results and the principle of conversational tracking have been put to use in robot studies via two different gesture strategies, one for behavior produced by the robot and one for interpreting user behavior. The robot's default behavior during a conversation is to to attend to the user's face (i.e., to keep its head oriented toward the user's face). However, when called upon to look at objects in the scene during its own conversational turn, the robot will turn to objects (either to point or indicate that the object is being reintroduced to user attention). Because the robot is not mobile and cannot interpret (via vision) other activities going on around it, the robot does not "look around" in the scene.

A portion of the robot's verbal behavior is coordinated with gestures as well. The robot converses about the task and obeys what is known about turn taking in conversation. The robot always returns to face the user when it finishes its conversational turn, if it had been directed elsewhere. It also awaits responses not only to questions, but to statements and requests, to determine user understanding be-

fore it continues the dialogue. The robot's collaboration and conversation abilities are based on the use of a tool for collaborative conversation.[26,27] An instantiated conversation for a hosting activity is discussed in Section 3.

In interpreting human behavior, the robot does not adhere to the expectation that the user will look at the robot much of the time. Instead it expects that the user will look around at whatever the user chooses. This expectation results from an intuition that users might not view the robot as a typical conversational partner. Only when the robot expects the user to view certain objects will it respond if the user does not do so. In particular, it will use verbal statements to direct the user to the object. However, just as the human-human data indicates, the robot interprets head nods as an indication of grounding.[c] Recognition of user head nodding is a probabilistic classification of sensed motion data, and its interpretation depends on the dialog context where it occurs. Only head nods that occur when or just before the robot awaits a response to a statement or request (a typical grounding point) are interpreted as acknowledgement of understanding.

The robot does not require that the user look at it when the user takes the conversational turn (as is prescribed by Sacks et al[28]). However, as we discuss later, that response is typical in a majority of the user interactions. The robot *does* expect that the user will take a turn when the robot signals its end of turn in the conversation. The robot takes the failure to do so as an indication of disengagement, which it follows up on by determining whether the user wishes to end the interaction.

While not based upon the results reported in the previous section, the robot makes use of opening and closing engagement behaviors in limited ways. The robot searches out a face while offering greetings and then uptakes engagement once it has some certainty (either through user speech or close proximity) that the user wants to engage (see the discussion in Section 3 for details on how this is accomplished). Disengagement occurs by offering to end the interaction, followed by standard (American) good-bye rituals,[29] including the robot's looking away from the user at the close.

## 3. Architectures to support human-robot engagement, collaboration and conversation.

Successful interaction between the human and robot requires the robot behave so as to express engagement in collaborative conversation, and to interpret the human's engagement from the human's behavior. This section reports on an architecture and its components to support engagement in collaborative interactions.

The robot's interaction abilities have been developed and tested using a target task wherein the robot collaboratively demonstrates a hardware invention[8] to a human interlocutor (Figure 1). The robot is designed to resemble a penguin wearing

[c]We view grounding as a backward looking engagement behavior, one that solidifies what is understood up to the present utterance in the interaction. Forward looking engagement tells the participants that they continue to be connected and aware in the interaction.

Fig. 1. Melvin demonstrates IGlassware to a visitor.

glasses, and is stationary (it is attached to a table). Because the robot has only wings but no hands, it relies on the human interlocutor to perform the physical manipulations necessary for this demonstration. The human must thus agree to collaborate for the demo to succeed. A typical interaction lasts about 3.5 minutes and a successful interaction about the demo is illustrated in Figure 2. Robot beat gestures,[d] head turns, head nods, and typical human gestures are not given in the figure. Should the human not agree to participate in the demo, the robot engages in brief, basic social "chit-chat" before closing the conversation.

The robot's hardware consists of

- 7 servos (two 2 DOF shoulders, 2 DOF neck, 1 DOF beak)
- Stereo camera (6 DOF head tracking software[21,33])
- Stereo microphones (with speech detection and direction-location software)
- Far distance microphone for speech recognition
- 3 computers: one for sensor fusion and robot motion, one for vision (6 DOF head tracking and head-gesture recognition), one for dialog (speech recognition, dialog modeling, speech generation and synthesis).

Our current robot is able to:

- Initiate an interaction by visually locating a potential human interlocutor and generating appropriate greeting behaviors,
- Maintain engagement by tracking the user's moving face, and judging the user's engagement based on head position (to the robot, to objects necessary for the collaboration, to other objects or to empty space),
- Reformulate a request upon failure of the user to respond to robot pointing,

[d]Beat gestures are hand or occasionally head movements that are hypothesized to occur to mark new information in an utterance.[4,6]

M: Hi, I'm Mel a robotic penguin.
U: *Hi.*
M: What's your name?
U: *Sam.*
M: Sam, I'd like to show you a demo. OK?
U: *No.*
M: But it's really interesting. Come on. Try it!
U: *Okay.*
M: It would be really nice to know what kind of progress people are making in their dining. So Paul Dietz created a new product called IGlassware for this. By the way, Sam, do you know Paul?
U: *No.*
M: Well, Paul is a member of MERL. This demo is one of his creations.
U: *Okay.*
M: IGlassware stands for Instrumented glassware! Sam, right there <*robot turns head to table*> is the IGlassware cup <*robot turns head to user*>. And near it, is the table readout. The word re-fill on the readout means that it is empty.
U: *Ok.*
M: All right, now we'll see how to use the cup.
U: *All right.*
M: First you should pour enough water from the pitcher into the glass to fill it up. Then make sure the glass is on the IGlassware table.
U: *Okay.* <*User pours water*>
M: Good, notice, the bar on the readout <*robot turns head to readout display on table*> shows that the glass is full <*robot turns back to user*>.
U: *Right.*
M: Okay, pour the water back in the pitcher.
U: *Okay.* <*User pours water*>
M: I'm waiting for a reading from the table for the glass.
M: Good. See, it registers needing a re-fill!
U: *Yes.*
M: Would you like me to explain how this works?
U: *Sure.*
M: The copper in the glass <*robot turns head to readout display*> transmits to the readout display by inductance with the surface of the table. The readout then displays the information coming to the table <*Robot turns head back to user*>.
U: *Cool.*
M: Would you like to know how this technology might be used in restaurants?
U: *Yes.*
M: The glass tells the restaurant when the customer needs a refill. In restaurants drinks mean profit, so this gadget makes it easier for restaurants to sell more drinks and make more money.
U: *Okay.*
M: Well, Sam, that's about all. Go see Paul Dietz, for more about IGlassware. So long!
U: *Good-bye.*

Fig. 2. Example conversation (M: Mel, U: human user).

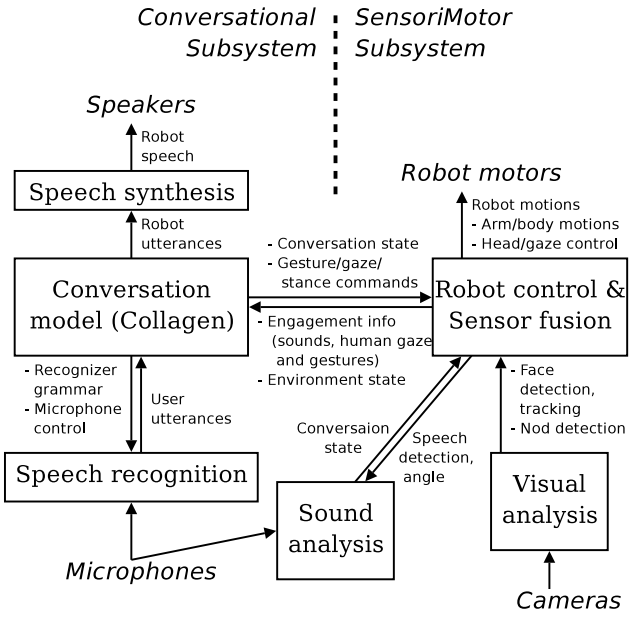


Fig. 3. Robot software architecture

- Point and look at objects in the scene,
- Interpret nods as backchannels and agreements in conversation,
- Understand limited spoken utterances and produce rich verbal spoken conversation, for invention demonstration and social "chit-chat,"
- Accept appropriate spoken responses from the user and make additional choices based on user comments,
- Disengage by verbal interaction and closing comments, and simple gestures, and
- Interpret user desire to disengage (through gesture and speech evidence).

Verbal and non-verbal behavior are integrated and occur fully autonomously.

The robot's software architecture consists of distinct sensorimotor and conversational subsystems. The conversational subsystem is based on the COLLAGEN$^{(TM)}$ collaboration and conversation model,[26,27] but enhanced to make use of strategies for engagement. The sensorimotor subsystem is a custom, dynamic, task-based blackboard robot architecture. It performs data fusion of sound and visual information for tracking its human interlocutors in a manner similar to other systems,[24] but its connection to the conversational subsystem is unique. The communication between these two subsystems is vital for managing engagement in collaborative interactions with a human.

Information about user manipulations and gestures must be communicated in summary form as discrete events from the sensorimotor to the conversational sub-

system so the conversational side can accurately model the collaboration and the user's engagement. The conversational subsystem uses this sensory information to determine: whether the user is continuing to engage with the robot, has responded to (indirect) requests to look at objects in the scene, has nodded at the robot (which must be interpreted in light of the current conversation state as either a backchannel, an agreement, or as superfluous), is looking elsewhere in the scene, or is no longer in the scene (a signal of possible disengagement).

High-level decisions and dialog state must be communicated from the conversational to the sensorimotor subsystem, so that the robot may gesture appropriately during robot and user utterances, and so that sensor fusion can appropriately interpret user gestures and manipulations. For example, the conversational subsystem tells the sensorimotor subsystem when the robot is speaking and when it expects the human to speak, so that the robot's head may track the user's face appropriately at these times. It also indicates the points during robot utterances when the robot should perform a given beat gesture[6] in synchrony with new information in the utterance, or when it should "gaze" at (expressed only by head position, not eye movements) or point to (with its wing) objects in the demonstration scene in coordination with spoken output. The sensorimotor subsystem knows that a GLANCEAT command from the conversational subsystem temporarily overrides any default face tracking behavior corresponding to robot speech.

In many circumstances, information about the dialog state must be communicated from the conversational to the sensorimotor subsystem for the latter to properly inform the former about significant human actions/gestures and environment state. For example, the sensorimotor subsystem only tries to detect the presence of human speech when the conversational subsystem expects human speech (e.g., when the robot has a conversational partner and is itself not speaking). Similarly, the conversational subsystem tells the sensorimotor subsystem when it expects, based on the known current purpose for the conversation as specified in its dialog model, that the human will look at a given object in the environment. The sensor fusion system can then generate an appropriate semantic event to the conversational subsystem when the human is observed to move his/her head to perform such a look. If the CUP and READOUT are in approximately the same place, then a user glance in that direction will be translated as LOOKAT(HUMAN,CUP) if the dialog context expects the user to look at the cup (e.g., when the robot says "here is the cup"), as LOOKAT(HUMAN,READOUT) if the dialog context expects the human to look at the readout, and as no event if no particular look is expected.

The current architecture has an important limitation. The conversational interactions between human and robot are robot controlled, that is, the robot has control of the conversation and directs what is discussed. This format is required because of the unreliability of current off-the-shelf speech recognition tools. User turns are limited to a few types of simple utterances (for example, "hello, goodbye, yes, no, okay, please repeat"). While people seem to want to say more complex

utterances,[e] such utterances cannot be interpreted with any reliability by current commercially available speech tools unless users train the speech tools. However, our robot is intended for all users without any type of pre-training, hence the limits on speech. Improvements in speech recognition systems will permit people to use complex utterances in which they can express their desires, goals, dissatisfactions and observations during collaborations with the robot. The existing conversation and collaboration model can already interpret the intentions conveyed in more complex utterances, even though none can be given to the robot at the present time.

## 4. Studies with users

A study of the effects of engagement gestures by the robot with human collaboration partners was conducted.[30] The study consisted of 2 groups of users interacting with the robot to collaboratively perform a demo of an invention, similar to that described in Figure 2. We summarize the results of that study, as well as recent results about nodding. We discuss measures used in that study as well as additional measures that should be useful in gauging the naturalness of robotic interactions during conversations with human users.

Thirty-seven participants were tested across two different conditions. In the first, the *mover* condition, with 20 participants, the fully functional robot conducted the demonstration of an invention known as the IGlassware table (involving a special cup, a pitcher filled with water and a readout display on the table). In the second, the *talker* condition, with 17 participants, the robot gave the same demonstration in terms of verbal utterances, but was constrained to talk by moving only its beak in synchrony with the words it spoke (no wing or head movements occurred). It also initially found the participant with its vision system, but thereafter, its head remained pointed in the direction in which it first found with the participant. It performed no other gestures.

All participants completed the demo with the robot. Their sessions were videotaped and followed by a questionnaire and informal debriefing. The videotape sessions were analyzed to determine what types of behaviors occurred in the two conditions and what behaviors provided evidence that the robot's engagement behavior approached human-human interaction. Details of the questionnaire format and protocol can be found in Sidner et al.[30]

While our work is highly exploratory, we predicted that people would prefer interactions with a robot with gestures. We also expected that participants in the mover condition would exhibit more interest in the robot during the interaction. However, we did not know exactly what form the differences take. As our results show, our predictions are partially correct.

Questionnaire data focused on the robot's likability, understanding of the demo,

[e]In our experimental study, despite being told to limit their utterances to ones similar to those above, users spoke more complex utterances during their conversations with the robot.

| *Tested factor* | *Significant effects* |
|---|---|
| Liking of Robot: | No effects |
| Knowledge of the demo: | No effects |
| Confidence of knowledge of the demo: | No effects |
| Engagement in the interaction: | *Effect for female gender:*<br>Female average: 4.84<br>Male average: 4.48<br>$F[1, 30] = 3.94$<br>$p = 0.0574$ (*Borderline significance*) |
| Reliability of robot: | *Effect for talker condition:*<br>Mover average: 3.84<br>Talker average: 5.19<br>$F[1, 37] = 13.77$<br>$p < 0.001$ (*High significance*) |
| Appropriateness of movements: | *Effect for mover condition:*<br>Mover average: 4.99<br>Talker average: 4.27<br>$F[1, 37] = 6.86$<br>$p = 0.013$ ($p < 0.05$: *Significance*) |

Table 2. Summary of questionnaire results

reliability/dependability, appropriateness of movement and emotional response. Results of that data are presented in Table 2. A multivariate analysis of condition, gender, and condition crossed with gender (for interaction effects) was undertaken. No difference was found between the two groups on likability, or understanding of the demo, while a gender difference for women was found on engagement response. Participants in the mover condition scored the robot more often as making appropriate gestures (significant with $F[1, 37] = 6.86$, $p = 0.013$, $p < 0.05$), while participants in the talker condition scored the robot more often as dependable/reliable ($F[1, 37] = 13.77$, $p < 0.001$, high significance).

What users say about their experience is only one means of determining interaction behavior, so the videotaped sessions were reviewed and transcribed for a number of features. With relatively little work in this area (see Nikano et al[23] for one study on related matters with a 2D agent), the choices were guided by measures that indicated interest and attention in the interaction: length of interaction time as a measure of overall interest, the amount of shared looking (i.e., the combination of time spent looking at each other and looking together at objects) and mutual gaze (looking at each other only) as measures of how coordinated the two participants were, the amount of looking at the robot during the human's turn, as a measure of attention to the robot, and the amount of looking at the robot overall, also as an attentional measure.

Table 3 summarizes the results for the two conditions. First, total interaction time in the two conditions varied significantly (row 1 in Table 3). This difference

| Measure | Mover | Talker | Test/Result | Significance |
|---|---|---|---|---|
| Interaction time | 217.7 sec | 183.1 sec | Single factor ANOVA: $F[1, 36] = 10.34$ | Significant: $p < 0.01$ |
| Shared looking | 51.1% | 36.1% | Single factor ANOVA: $F[1, 36] = 8.34$ | Significant: $p < 0.01$ |
| Mutual gaze | 40.6% | 36.1% | Single-factor ANOVA: $F[1, 36] = 0.74$ | No: $p = 0.40$ |
| Talk directed to robot | 70.4% | 73.1% | Single-factor ANOVA: $F[1, 36] = 4.13$ | No: $p = 0.71$ |
| Look backs, overall | 19.65 avg. median: 18-19 | 12.82 avg. median: 12 | Single-factor ANOVA: $F[1, 36] = 15.00$ | Highly: $p < 0.001$ |
| Table-look 1 | 12/19 (63%) | 6/16 (37.5%) | t-tests $t(33) = 1.52$ | Weak: One-tailed: $p = 0.07$ |
| Table-look 2 | 11/20 (55%) | 9/16 (56%) | t-tests $t(34) = -1.23$ | No: One-tailed: $p = 0.47$ |

Table 3. Summary of behavior test results in human-robot interaction experiment.

may help explain the subjective sense gathered during video viewing that the talker participants were less interested in the robot and more interested in doing the demo, and hence completed the interaction more quickly.

While shared looking (row 2 in Table 3) was significantly greater among mover participants, this outcome is explained by the fact that the robot in the talker condition could never look with the human at objects in the scene. However, it is noteworthy that in the mover condition, the human and robot spent 51% of their time (across all participants) coordinated on looking at each other and the demo objects. Mutual gaze (row 3 in Table 3) between the robot and human to each other was not significantly different in the two conditions.

Attention to the robot can be measured in two additional ways. The measure of talk directed to the robot during the human's turn (row 4 in Table 3) is an average across all participants as a percentage of the total number of turns per participant. There is no difference in the rates. What is surprising is that both groups of participants directed their gaze to the robot for 70% or more of their turns. This result suggests that a conversational partner, at least one that is reasonably sophisticated in conversing, is a compelling partner, even with little gesture ability.[f] However, the second measure, the number of times the human looked back at the robot, are highly significantly greater in the mover condition. Since participants

[f]We did not eliminate beak movements by the robot since informal pre-testing indicated that users found the resulting robot non-conversational.

spend a good proportion of their time looking at the table and its objects (55% for movers, 62% for talkers), the fact that they interrupt their table looks to look back to the robot is an indication of how engaged they are with it compared with the demonstration objects. This result indicates that a gesturing robot is a partner worthy of closer attention during the interaction.

We also found grounding effects in the interaction that we had not expected. Participants in both conditions nodded at the robot, even though during this study, the robot was not able to interpret nods in any way. Eleven out of twenty participants in the mover condition nodded at the robot three or more times during the interaction (55%) while in the talker condition, seven out of seventeen participants (41%) did. Nods were counted only when they were clearly evident, even though participants produced slight nods even more frequently. The vast majority of these nods accompany "okay," or "yes," while a few accompany a "goodbye." There is personal variation in nodding as well. One participant, who nodded far more frequently than all the other participants (a total of 17 times), nodded in what appeared to be an expression of agreement to many of the robot's utterances. The prevalence of nodding, even with no evidence that it is understood, indicates just how automatic this conversational behavior is. It suggests that the conversation was enough like a human-to-human conversation to produce this grounding effect even without planning for this type of behavior. The frequency of nodding in these experiments drove in part the inclusion of nod understanding in the robot's more recent behavior repertoire.[17]

We also wanted to understand the effects of utterances where the robot turned to the demo table as a deictic gesture. For the two utterances where the robot turned to the table, we coded when participants turned in terms of the words in the utterance and the robot's movements. These utterances were: "Right there <*robot gesture*> is the IGlassware cup and near it is the table readout," and "The <*robot gesture*> copper in the glass transmits to the readout display by inductance with the surface of the table." For both of these utterances, the mover robot typically (but not always) turned its head towards and down to the table as its means of pointing at the objects. The time in the utterance when pointing occurred is marked with the label <*robot gesture*>. Note the talker robot never made such gestures.

For the first instance, Table-look 1, ("Right there. . . "), 12/19 mover participants (63%) turned their heads or their eye gaze during the phrase "IGlassware cup." For these participants, this change was just after the robot has turned its head to the table. The remaining participants were either already looking at the robot (4 participants), turned before it did (2 participants) or did not turn to the table at all (1 participant); 1 participant was off-screen and hence not codeable. In contrast, among the talker participants, only 6/16 participants turned their head or gaze during "IGlassware cup" (37.5%). The remaining participants were either already looking at the table before the robot spoke (7 participants) or looked much later during the robot's utterances (3 participants); 1 participant was off camera and

hence not codeable.

For the second declarative utterance, Table-look 2, ("The copper in the glass. . . "), 11 mover participants turned during the phrases "in the glass transmits," 7 of the participants at "glass." In all cases these changes in looking followed just after the robot's change in looking. The remaining mover participants were either already looking at the table at the utterance start (3 participants), looked during the phrase "glass" but before the robot turned (1 participant), or looked during "copper" when the robot had turned much earlier in the conversation (1 participant). Four participants did not hear the utterance because they had taken a different path through the interaction. By comparison, 12 of the talker participants turned during the utterance, but their distribution is wider: 9 turned between "copper in the glass transmits" while 3 participants turned much later in the utterances of the turn. Among the remaining talker participants, 2 were already looking when the utterance began, 1 participant was distracted by an outside intervention (and not counted), and 2 participants took a different path through the interaction.

The results for these two utterances are too sparse to provide strong evidence. However, they show that participants pay attention to when the robot turns his head, and hence his attention, to the table. When the robot does not move, participants turn their attention based on other factors (which appear to include the robot's spoken utterance, and their interest in the demo table).

While the results of this experiment indicate that talking attracts people to respond back to a robot, it appears that gestures make them even more attracted. One might argue that movement alone explains why people looked more often at the robot, but the talking-only robot does have some movement—its beak moves. So it would seem that gestures are the critical matter. The gestures used in the experiment are ones appropriate to conversation. It is possible that it is the gestures themselves, and not their appropriateness in the context of the conversation, that are the source of this behavior. Our current experiment does not allow us to distinguish between appropriate gestures and non-appropriate ones. However, if the robot were to move in ways that were inappropriate to the conversation, and if human partners ignored the robot in that case, then we would have stronger evidence for engagement gestures. We have recently completed a set of experiments that were not intended to judge these effects, but have produced a number of inappropriate gestures for extended parts of an interaction. These results may tell us more about the importance of appropriate gestures during conversation.

Developing quantitative observational measures of the effects of gesture on human-robot interaction continues to be a challenging problem. The measures used in this work, interaction time, shared looking, mutual gaze, looks during human turn, looks back overall, number of times nodding occurred and in relation to what conversation events, and observations of the effects of deictic gestures, are all relevant to judging attention and connection between the human and the robot in conversation. The measures all reflect patterns of behavior that occur in human-to-human conversation. This work has assumed that it is reasonable to expect to

find these same behaviors occurring in human-robot conversation, as indeed they do. However, we wish for finer-grained measures, that would allow us to judge more about the robot's gestures as natural or relevant at a particular point in the conversation. Such measures await further research.

## 5. Related Research

While other researchers in robotics are exploring aspects of gesture (for example Breazeal[1] and Kanda et al[15]), none of them have attempted to model human-robot interaction to the degree that involves the numerous aspects of engagement and collaborative conversation that we have set out above. A robot developed at Carnegie Mellon University serves as a museum guide[3] and navigates well while avoiding humans, but interacts with users via a 2D talking head with minimal engagement abilities. Robotics researchers interested in collaboration and dialogue[11] have not based their work on extensive theoretical research on collaboration and conversation. Research on human-robot gesture similarity[25] indicates that body gestures corresponding to a joint point of view in direction-giving affect the outcome of human gestures as well as human understanding of directions.

Most similar in spirit to work reported here is the ARMAR II robot.[9,10] ARMAR II is speech enabled, has some dialogue capabilities, and has abilities to track gestures and people. However, the ARMAR II work is focused on teaching the robot new tasks (with programming by demonstration techniques), while our work has been focused on improving the interaction capabilities needed to hold conversations and undertake tasks. Recently, Breazeal et al[2] have explored teaching a robot a physical task that can be performed collaboratively once learned.

Research on infant robots with the ability to learn mutual gaze and joint attention[16,22] offers exciting possibilities for eventual use in more sophisticated conversational interactions.

Our work is also not focused on emotive interactions, in contrast to Breazeal among others.[18] For 2D conversational agents, researchers (notably Cassell et al[5] and Johnson et al[14]) have explored agents that produce gestures in conversation. However, they have not tried to incorporate recognition as well as production of these gestures, nor have they focused on the full range of these behaviors to accomplish the maintenance of engagement in conversation.

## 6. Future work

Future work will improve the robot's conversational language generation so that nodding by humans will be elicited more easily. In particular, there is evidence in the linguistic literature, *inter alia*[7] that human speech tends to short intonational phrases with pauses for backchannels rather than long full utterances that resemble sentences in written text. By producing utterances of the short variety, we expect that people will nod more naturally at the robot. We plan to test our hypothesis by

comparing encounters with our robot where participants are exposed to different kinds of utterances to test how they nod in response.

The initiation of an interaction is an important engagement function. Explorations are needed to determine the combinations of verbal and non-verbal signals that are used to initially engage a human user in an interaction.[20] Our efforts will include providing mobility to our robot as well as extending the use of current vision algorithms to "catch the eye" of the human user and present verbal feedback in the initiation of engagement.

Current limits on the robot's vision make it impossible to determine the identity of the user. Thus if the user leaves and is immediately replaced by another person, the robot cannot tell that this change has happened. Identity recognition algorithms, in variable light without color features, will soon be used, so that the robot will be able to recognize the premature end of an interaction when a user leaves. This capability will also allow the robot to judge when the user might desire to disengage due to looks away from either the robot or the objects relevant to collaboration tasks.

## 7. Conclusions

In this paper we have explored the concept of engagement, the process by which individuals in an interaction start, maintain and end their perceived connection to one another. We have reported on one aspect of engagement among human interactors—the effects of tracking faces during an interaction. We have reported on a humanoid robot that participates in conversational, collaborative interactions with engagement gestures. The robot demonstrates tracking its human partner's face, participating in a collaborative demonstration of an invention, and making engagement decisions about its own behavior as well as the human's during instances where face tracking was discontinued in order to track objects for the task. We also reported on our findings of the effects on human participants who interacted when the robot performed or did not perform engagement gestures.

While this work represents a first step in understanding the engagement process, it demonstrates that engagement gestures have an effect on the behavior of human interactors with robots that converse and collaborate. Simply said, people direct their attention to the robot more often in interactions where gestures are present, and they find these gestures more appropriate than when they are not present. We believe that as the engagement gestural abilities of robots become more sophisticated, human-robot interaction will become smoother, be perceived as more reliable, and will make it possible to include robots into the everyday lives of people.

## References

1. C. Breazeal, *Affective interaction between humans and robots*, Proceedings of the 2001 European Conference on Artificial Life (ECAL2001) (Prague, Czech Republic), 2001.
2. C. Breazeal, G. Hoffman, and A. Lockerd, *Teaching and working with robots as a*

*collaboration*, The Third International Conference on Autonomous Agents and Multi-Agent Systems AAMAS 2004, ACM Press, July 2004, pp. 1028–1035.
3. W. Burgard, A. B. Cremes, D. Fox, D. Haehnel, G. Lakemeyer, D. Schulz, W. Steiner, and S. Thrun, *The interactive museum tour guide robot*, In Proceedings of AAAI-98, AAAI Press, Menlo Park, CA, 1998, pp. 11–18.
4. J. Cassell, *Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents.*, Embodied Conversational Agents (J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, eds.), MIT Press, Cambridge, MA, 2000.
5. J. Cassell, J. Sullivan, S. Prevost, and E. Churchill (eds.), *Embodied conversational agents*, MIT Press, Cambridge, MA, 2000.
6. J. Cassell, H. Högni Vilhjálmsson, and T.W. Bickmore, *BEAT: The behavior expression animation toolkit*, SIGGRAPH 2001, Computer Graphics Proceedings (Eugene Fiume, ed.), ACM Press / ACM SIGGRAPH, 2001, pp. 477–486.
7. H. H. Clark, *Using language*, Cambridge University Press, Cambridge, 1996.
8. P. H. Dietz, D. L. Leigh, and W. S. Yerazunis, *Wireless liquid level sensing for restaurant applications*, IEEE Sensors **1** (2002), 715–720.
9. R. Dillman, R. Becher, and P. Steinhaus, *ARMAR II – a learning and cooperative multimodal humanoid robot system*, International Journal of Humanoid Robotics **1** (2004), no. 1, 143–155.
10. R. Dillman, M. Ehrenmann, P. Steinhaus, O. Rogalla, and R. Zoellner, *Human friendly programming of humanoid robots–the German Collaborative Research Center*, The Third IARP Intenational Workshop on Humanoid and Human-Friendly Robotics (Tsukuba Research Centre, Japan), December 2002.
11. T. Fong, C. Thorpe, and C. Baur, *Collaboration, dialogue and human-robot interaction*, 10th International Sumposium of Robotics Research (Lorne, Victoria, Australia), November 2001.
12. B. J. Grosz and S. Kraus, *Collaborative plans for complex group action*, Artificial Intelligence **86** (1996), no. 2, 269–357.
13. B. J. Grosz and C. L. Sidner, *Attention, intentions, and the structure of discourse*, Computational Linguistics **12** (1986), no. 3, 175–204.
14. W. L. Johnson, J. W. Rickel, and J. C. Lester, *Animated pedagogical agents: Face-to-face interaction in interactive learning environments*, International Journal of Artificial Intelligence in Education **11** (2000), 47–78.
15. Ishiguro H. Imai. M. Ono T. Kanda, T. and K. Mase, *A constructive approach for developing interactive humanoid robots*, Proceedings of IROS 2002, IEEE Press, New York, 2002.
16. H. Kozima, C. Nakagawa, and H. Yano, *Attention coupling as a prerequisite for social interaction*, Proceedings of the 2003 IEEE International Workshop on Robot and Human Interactive Communication, IEEE Press, New York, 2003, pp. 109–114.
17. C. Lee, N. Lesh, C. Sidner, L.-P. Morency, A. Kapoor, and T. Darrell, *Nodding in conversations with a robot*, Proceedings of the ACM International Conference on Human Factors in Computing Systems, April 2004.
18. C.L. Lisetti, S.M. Brown, K. Alvarex, and A.H. Marpaung, *A social informatics approach to human-robot interaction with a service social robot*, IEEE Transactions on Systems, Man and Cybernetics **34** (2004), no. 2, 195–209.
19. K. E. Lochbaum, *A collaborative planning model of intentional structure*, Computational Linguistics **24** (1998), no. 4, 525–572.
20. D. Miyauchi, A. Sakurai, A. Makamura, and Y. Kuno, *Active eye contact for human-robot communication*, Proceedings of CHI 2004–Late Breaking Results, no. CD Disc 2, ACM Press, 2004, pp. 1099–1104.

21. L.-P. Morency, P. Sundberg, and T. Darrell, *Pose estimation using 3D view-based eigenspaces*, ICCV workshop on Analysis and Modeling of Face and Gesture (Nice, France), October 2003.
22. Y. Nagai, K. Hosoda, A. Morita, and M. Asada, *A constructive model for the development of joint attention*, Connection Science **15** (2003), no. 4, 211–229.
23. Y. Nikano, G. Reinstein, T. Stocky, and J. Cassell, *Towards a model of face-to-face grounding*, Proceedings of the 41st meeting of the Association for Computational Linguistics (Sapporo, Japan), 2003, pp. 553–561.
24. H. G. Okuno, K. Nakadai, K-I. Hidai, H. Mizoguchi, and H. Kitano, *Human robot non-verbal interaction empowered by real-time auditory and visual multiple-talker tracking*, Advanced Robotics **17** (2003), no. 2, 115–130(16).
25. T. Ono, M. Imain, and H. Ishiguro, *A model of embodied communications with gestures between humans and robots*, Proceedings of the 23rd meeting of the Cognitive Science Society (J.D. Moore and K. Stenning, eds.), Lawrence Erlbaum Associates, Mahwah, NJ, 2001, pp. 760–765.
26. C. Rich and C. L. Sidner, *COLLAGEN: A collaboration manager for software interface agents*, User Modeling and User-Adapted Interaction **8** (1998), no. 3/4, 315–350.
27. C. Rich, C. L. Sidner, and N. Lesh, *COLLAGEN: Applying collaborative discourse theory to human-computer interaction*, AI Magazine **22** (2001), no. 4, 15–25, Special Issue on Intelligent User Interfaces.
28. H. Sacks, E.A. Schegeloff, and F. Jefferson, *A simplest systematics for the organization of turn taking on conversation*, Language **50** (1974), no. 4, 696–735.
29. E. Schegeloff and H. Sacks, *Opening up closings*, Semiotica **7** (1973), no. 4, 289–327.
30. C. L. Sidner, C. D. Kidd, C. H. Lee, and N. Lesh, *Where to look: A study of human-robot engagement*, ACM International Conference on Intelligent User Interfaces (IUI), ACM, January 2004, pp. 78–84.
31. C. L. Sidner, C. H. Lee, and N. Lesh, *Engagement when looking: behaviors for robots when collaborating with people*, Diabruck: Proceedings of the 7th workshop on the Semantic and Pragmatics of Dialogue (I. Kruiff-Korbayova and C.Kosny, eds.), University of Saarland, 2003, pp. 123–130.
32. D. Traum and J. Rickel, *Embodied agents for multi-party dialogue in immersive virtual world*, Proceedings of the International Joint Conference on Autonomous Agents and Multi-agent Systems (AAMAS 2002), July 2002, pp. 766–773.
33. P. Viola and M. Jones, *Rapid object detection using a boosted cascade of simple features*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (Hawaii), 2001, pp. 905–910.