MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Where to Look: A Study of Human-Robot Engagement

Candace L. Sidner, Cory D. Kidd, Christopher Lee and Neal Lesh

TR2003-123 November 2003

Abstract

This paper reports on a study of human subjects with a robot designed to mimic human conversational gaze behavior in collaborative conversation. The robot and the human subject together performed a demonstration of an invention reated at our laboratory; the demonstration lasted 3 to 3.5 minutes. We briefly discuss the robot architecture and then focus the paper on a study of the effects of the robot operating in two different conditions. We offer some conclusions based on the study about the importance of engagement for 3D IUIs. We will present video clips of the subject interactions with the robot at the conference.

IUI 2004

This paper was accepted for publication in the Proceedings of the Intelligent User Interfaces Conference, 2004.

Where to Look: A Study of Human-Robot Engagement

Candace L. Sidner* , Cory D. Kidd, Christopher Lee* and Neal Lesh***
Mitsubishi Electric Research Labs * and MIT Media Lab**
Cambridge MA 02139
{sidner, lee, lesh}@merl.com, coryk@media.mit.edu

ABSTRACT

This paper reports on a study of human subjects with a robot designed to mimic human conversational gaze behavior in collaborative conversation. The robot and the human subject together performed a demonstration of an invention created at our laboratory; the demonstration lasted 3 to 3.5 minutes. We briefly discuss the robot architecture and then focus the paper on a study of the effects of the robot operating in two different conditions. We offer some conclusions based on the study about the importance of engagement for 3D IUIs. We will present video clips of the subject interactions with the robot at the conference.

Categories and Subject Descriptors

H.5.2 Information systems: User Interfaces.

Keywords

Human-robot interaction, engagement, intelligent user interfaces, collaborative conversation.

1. INTRODUCTION

The creation of two and three-dimensional collaborative partners raises important challenges in the behavior of these computational entities. This paper reports on results of creating a 3D robot with engagement capabilities [17]. By engagement, we mean the process by which two (or more) participants establish, maintain and end their perceived connection. This process includes: initial contact, negotiating a collaboration, checking that other is still taking part in the interaction, evaluating whether to stay involved, and deciding when to end the connection. The robot we have developed interacts with a single user in a collaboration that involves: spoken language (both understanding and generation), beat gestures with its arm, and head gestures to track the user and to turn to look at objects of interest in the interaction. The robot also initiates interactions with users, and performs typical preclosings and goodbyes to end the conversation. All these capabilities increase the means by which the robot can engage the user in an interaction.

These capabilities make it possible for a robot to have a face-to-face conversation with a person. But such conversations presumably require more than just talking. The robot must use its
face and use it well. It must also use its vision capabilities to assess the activities of its human conversational partner.

IUI'04, January 13-16, 2004, Madeira, Funchal, Portugal.

Effective use of these capabilities requires careful study and evaluation of multiple users interacting with robots. In this paper we explore the impact of where a robot looks during conversation, in particular with regard to objects of interest in the conversation. The paper reports on a user study with 37 subjects who interacted with our robot on the task of collaboratively performing a demonstration of an invention created in our laboratory.

The paper first describes how this robot was created and provides an example interaction with a user. Video clips will be available of users interacting with the robot for presentation at the conference. The main body of the paper discusses the user study.

2. CREATING AN ENGAGING ROBOT

Our robotic agent is a homegrown stationary robot created at Mitsubishi Electric Research Labs (MERL). It uses 5 servomotors to control the movement of the robot's head, mouth and two wings. The robot takes the appearance of a penguin (called Mel). Mel can open and close his beak, nod and turn his head, and flap his "wings" up and down. A speaker provides audio output. Two cameras near Mel provide vision capabilities, and three microphones provide speech recognition (1 far distance microphone) and sound location (two microphones in the same focal plane as one of the vision cameras). Figure 1 shows Mel and his associated hardware.

Figure 1. Mel the robotic penguin

Our architecture for collaborative interactions uses several different systems and algorithms, largely developed at MERL. The architecture is illustrated in Figure 2. The conversational and collaborative capabilities of our robot are provided by the Collagen™ middleware for collaborative agents [15, 16], and commercially available speech recognition software (IBM ViaVoice). We use a face detection algorithm [20], a sound location algorithm, a speech detection algorithm, and an object recognition algorithm [1] and fuse the sensory data before

passing results to the Collagen™ system. The agent control makes decisions about how to proceed in the interaction based on rules about engagement (how to proceed at the beginning, middle and ends of an interaction) and the state of the dialogue (provided by the Collagen™ system). Agent actions from the agent control are passed to a speech synthesizer and to the robot control algorithms to produce gestures. All these operations occur in real-time. Further details about the architecture and current implementation can be found in [18].

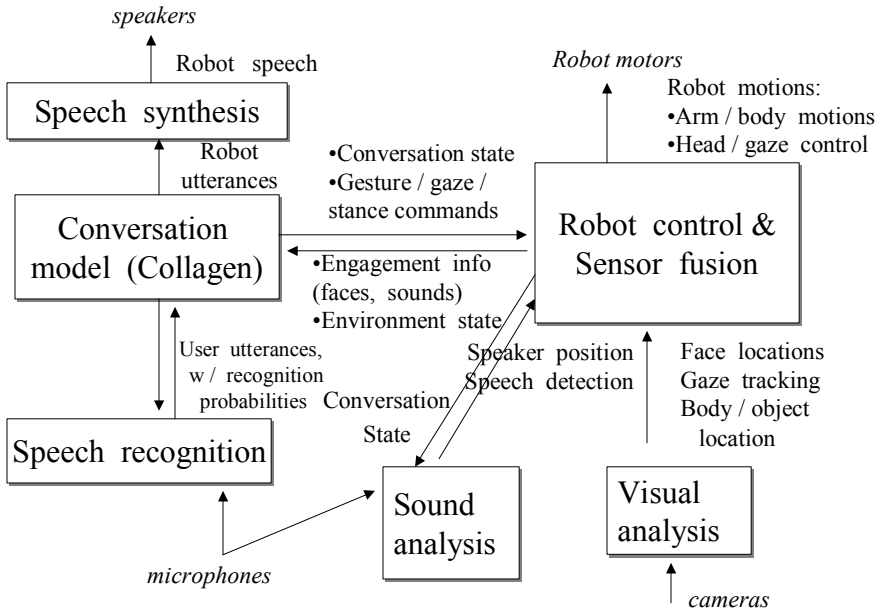

Figure 1: An Architecture for Human-Robot Interaction

The engagement rules for Mel are drawn from analysis of human-human interactions based on videotapes of a pair of people demonstrating and observing equipment at MERL [19]. These rules determine how the robot should gesture during the user's turn, and during its turn both how to gesture and what to say. In particular, only during Mel's turn will he look towards objects relevant to the conversation. At other times, he looks at the user as he speaks. He also expects the user to look at him, except when he points out equipment, in which case the user is expected to view the equipment. Failure to do so will cause Mel to choose a response to further guide the user's attention.

While other researchers in robotics are exploring aspects of gesture (for example, [2], [8]), the current work is unique in modeling human-robot interaction to a degree that involves the numerous aspects of engagement and collaborative conversation that are set out above. Robotics researchers interested in collaboration and dialogue [6] have not based their work on extensive theoretical research on collaboration and conversation, as has been accomplished for Mel. Our work is also not focused on emotive interactions, in contrast to [2] among others. For 2D conversational agents, researchers (notably, [5],[7]) 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. A robot developed at Carnegie Mellon University serves as a museum guide [4] and navigates well while avoiding humans, but interacts with users via a 2D talking head with minimal engagement abilities.

Because Mel is a stationary robot with no hands to manipulate objects, the typical robot task of navigating a space, picking up objects and delivering them was impossible. The challenge became choosing a useful task that it could do with people. Given a focus in our group on hosting activities for robots (that is, activities where an agent in an environment provides services, particularly information and entertainment services), we concluded that it would be challenging and unique for the robot to give a demo similar to the demo in videotapes of the human-human interactions.

An invention from MERL, called IGlassware, lent itself to human-robot interaction because the human could do all the physical actions, with directions and explanations coming from the robot. IGlassware is a specially designed cup that senses fluid level using capacitance and communicates the level to a special table by inductance. Mel was programmed to demonstrate this table to a human visitor. A typical interaction is shown in Figure 3; the robot's gestures with its head towards the table are noted in the text. However, many different conversations are possible, including one that is not about the demo at all (if the human participant refuses to be shown a demo).

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 I Glassware 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: I Glassware stands for Instrumented glassware! Sam, right there <robot turns head to table> is the I Glassware 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 glass to the pitcher to fill it up. Then make sure the glass is on the IGlassware table.
U: Okay.
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.
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 I Glassware. So long!
U: Good-bye.

Figure 3. A typical interaction with Mel

In an interaction, Mel uses his vision system to find a conversational partner (from a set of people in a room) by zeroing in on the person who responds to his hello using the face detection algorithms (which can detect up to 25 faces) and the sound location algorithms. Thereafter Mel tracks the conversational partner's face and adjusts his "gaze" towards the partner (even when the partner moves about). Mel has eyes in his head, but they do not see, so his gaze merely communicates his focus of attention to the partner. Mel does not look at the human partner at all times, because during the demo, he turns to look at the table and its contents as he speaks about them. Mel also prompts a partner who fails to look at the table to notice the objects there. After the demo and explanation conclude, Mel wishes the partner goodbye, waves and drops his head to his chest to indicate that he is no longer available.

Note that interactions with Mel are greatly affected by the uncertainty of sensory information. The Mel interactions are designed for any speaker of English without training. There are speech recognition errors (sometimes brief, sometimes of several exchanges). In addition, early on, we discovered that given the opportunity to say something, users say an unpredictable set of responses to Mel. Hence we designed the demo interaction with Mel as a "robot controlled" conversation, that is, the robot directs most of the conversation and elicits limited types of responses from users. This design reduced the unpredictability of user exchanges, but did not eliminate them entirely. In our user study, users asked questions, offered explanations as part of their refusals, and made statements about the demonstration. Likewise, interpretation to vision input relies on uncertain information, and Mel sometimes looses faces of his users. Often he is able to regain them, but occasionally the user moves so that our camera cannot detect the face. In such cases, Mel either finds another user to look at, or if none are present, he looks to the last place he saw a user.

3. USER STUDY

When we began our study, our intended goal was to determine how effective Mel was at mimicking human conversational behavior. We wanted to know if Mel's gestures were appropriate ones, and ones that would cause users to behave as intended and to feel more natural in interacting with the robot.

What we learned from the evaluation went beyond our intended goal. Our results do provide some information about the appropriateness of the robot's gestures and how to improve those gestures. However, one of our data sources, videotapes of the subjects with Mel, provided a great deal of material about how each subject proceeded in the conversation. To make sense of our observations, we devised categories for the conversational behaviors of subjects along with measures for each. These measures revealed more about what happens when subjects talk to a robot that has just a talking head compared with one that has an active head and body.

Study circumstances: Thirty-seven subjects were tested in two different conditions. In the first, the *mover condition*, the fully functional robot conducted the demonstration of the IGlassware table. In the second, the *talker* condition, 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. It also initially found the subject with its vision system, but thereafter, its head remained looking in the direction in which it first found with the subject. This constraint meant in many cases that the robot did not look at the subject during most of the demo. The entire interaction was videotaped as well as audiotaped (see Figure 4). The study used a between-subjects design, and hence no subject interacted with the robot in both conditions.

Protocol for the study: Each participant was randomly pre-assigned into one of the two conditions. 20 subjects participated in mover condition and 17 in talker condition. A video camera was turned on after the subject arrived. The subject was introduced to the robot (as Mel) and told the stated purpose of the interaction (i.e. to see a demo from Mel). Subjects were told that they would be asked a series of questions at the completion of the interaction.

When the robot was turned on, the subject was instructed to approach Mel. The interaction began, and the experimenter left the room. After the demo, subjects were given a short questionnaire that contains the scales described in the Results section below. Lastly they also reviewed the videotape with the experimenter to discuss any thoughts they had about the interaction.

Figure 4. Subject interacting with Mel

Our results come from two different sources, questionnaires meant to elicit from subjects their response to the interaction as they perceived it, and behavioral assessments taken from observations of the video data.

4. RESULTS

4.2 Questionnaires

Subjects were provided with post-interaction questionnaires. Questionnaires were devoted to five different factors concerning the robot:

- General liking of Mel (devised for experiment; 3 items) - This measure gives the participants' overall impressions of the robot and their interactions with it.
- Knowledge and confidence of knowledge of demo (devised for experiment; 6 items) - The former concerns task differences. A difference among subjects was not expected, but such a difference would be very telling about the two conditions of interaction. Confidence in the knowledge of the demo is a finer-grained measure of task differences.
- Engagement in the interaction (adapted from [10,11]; 5 items) - Lombard and Ditton's notion of engagement (different from ours) is a good measure of how natural and interactive the experience seemed to the person interacting with the robot.
- Reliability of the robot (adapted from [9], 4 items) - While not directly related to the outcome of this interaction, the perceived reliability of the robot is a good indicator of how much the participants would be likely to depend on the robot for information on an ongoing basis. A higher rating of reliability means that the robot will be perceived more positively in future interactions.
- Effectiveness of movements (devised for this experiment; 5 items) - This measure is used to determine the quality of the gestures and looking.

A multivariate analysis of condition, gender, and condition crossed with gender (for interaction effects) provided the following results by category (summarized) in the table below:

For factors where there is no difference in effects, it is evident that all subjects understood the demo and were confident of their response. Knowledge was a right/wrong encoding of the answers to the questions. In general, most subjects got the answers correct (overall average = 0.94; movers = 0.90; talkers = 0.98). Confidence was scored on a 7-point Likert scale. Both conditions rated highly (overall average = 6.14; movers = 6.17; talkers = 6.10). All subjects also liked Mel more than they disliked him. On a 7-point Likert scale, the overall average was 4.86. The average for the mover condition was 4.78, while the talker condition was actually higher, at 4.96. If one subject who had difficulty with the interaction is removed, the mover average comes up to 4.88. None of these differences between conditions is significant.

Table 1. Summary of Questionnaire Results

Liking of Mel: no effects
Knowledge of the demo: no effects
Confidence of knowledge of the demo: no effects
Engagement in the interaction -- effect for female gender:

Female average: 4.84
Male average: 4.48
$F[1,30] = 3.94$ $p = 0.0574$ (Borderline significance)

Reliability of Mel -- effect for talker condition:

Mover average = 3.84
Talker average = 5.19
$F[1,37] = 13.77$ $p < 0.001$ (High significance)

Appropriateness of movements -- effect for mover condition:

Mover average = 4.99
Talker average = 4.27
$F[1,37] = 6.86$ $p=0.013$ ($p < 0.05$) (Significance.)

The three factors with effects on subjects provide some insight into the interaction with Mel. First, consider the effects of gender on engagement. The sense of engagement in [10,11] concerns being "captured" by the experience. Questions for this factor included:

How engaging was the interaction?
How relaxing or exciting was the experience?
How completely were your senses engaged?
The experience caused real feelings and emotions for me.
I was so involved in the interaction that I lost track of time.

While these results are certainly interesting, we conclude only that male and female users may interact in different ways with fully functional robots. This result mirrors work by [9,14] who found differences in gender, not for engagement, but for likeability and credibility.

Concerning appropriateness of movements, mover subjects perceived the robot as moving appropriately. In contrast, talker subjects felt Mel did not move appropriately. However, the talker subjects did indicate that they thought he moved. This effect confirms our sense that a talking head is not doing everything that a robot should be doing in an interaction, when people and objects are present. Mover subjects' responses

indicated that they thought that "The interaction with Mel was just like interacting with a real person; Mel always looked at me at the appropriate times," and "Mel did not confuse me with where and when he moved his head and wings."

However, it is striking that subjects in the talker condition found the robot more reliable. Subjects responded to statements "I could depend on Mel to work correctly every time, Mel seems reliable, If I did the same task with Mel again, he would do it the same way," and "I could trust Mel to work whenever I need him to." There are two possible conclusions to be drawn about reliability given the response to appropriateness: (1) some of the robot's behaviors were either not correct or not consistently produced, or (2) devices such as robots with moving parts are seen as more complicated, more likely to break and hence less reliable. Clearly much more remains to be done before users are perfectly comfortable with a robot.

4.2 Behavioral Observations

In this section the behavior of subjects that were observed from videos taken of their interactions with the robot is reviewed. The videos showed a number of ways to improve the robot: changing individual gestures, improving recovery from speech recognition errors, recovery from loss of the subject's face and the like. However, we also wanted to know if there were any differences in the subjects' conversational behavior with the robot acting in the two conditions, and if so, what these were.

We are unaware of studies that have looked at human-robot conversational behavior in any detail (although some preliminary results are reported in [13]). Therefore we had to decide what behaviors to consider. We choose to consider length of interaction time, the amount of shared looking (i.e. looking at each other and looking together at objects) as a measure of how coordinated the two participants were, the amount of looking at the robot during the subject's turn, as a measure of attention to the robot, and the amount of looking at the robot overall, also an attentional measure. We also wanted to understand the effects of utterances where the robot turned to the demo table. For the two utterances where the robot turned to the table, we coded when subjects turned in terms of the words in the utterance and the robot's movements. We summarize our results for each of these measures in Table 2. We then explain each measure and the results in more detail.

First, total **interaction time** by the two conditions varied by a significant amount (row 1 in Table 2). This difference coincides with our subjective sense that the talkers were less interested in the robot and more interested in doing the demo.

The nature of the two subject pools with respect to **shared looking** was coded. Shared looking occurred when subject and robot looked at each other (so called mutual gaze) and when they looked at the same object (the IGlassware table and its contents). Shared looking is an indication of how coordinated two participants are in their interaction. The more shared looking the more the participants share an interest and hence engagement in the interaction. Shared looking is more relevant than simply mutual gaze, because participants in a collaboration where other objects are discussed or used must pay attention to these as well as their partner in coordination with the content of the conversation.

In the study, the robot, when it looked at the table, turned its head to the table in two directions (left and down), with its beak serving as a well-defined pointer. While the robot does not have seeing eyes in its head, its turns to the table provided clear information that it was "looking" at the table and not at other devices nearby in the room (such as computer monitors and laptops). Only the general view of the table was considered because we did not have a means of telling exactly which objects the subject or the robot were viewing.

Table 2. Summary of behavioral test results

Measure	Mover	Talker	Test/Result	Significance
Interac-tion Time	217.7 seconds	183.1 seconds	Single-factor, ANOVA: $F(1,36)= 10.34$	Significant difference: $p < 0.01$
Shared Looking	51.1%	35.9%	Single factor ANOVA: $F(1,36)= 8.34$	Significant difference: $p < 0.01$
Mutual Gaze	40.6%	36.1%	Single-factor, ANOVA: $F(1,36) = 0.74$	No significant difference: $p = 0.40$
Talk directed to Mel	70.4%	73.1%	Single-factor, ANOVA, $F[1,36]= 4.13$	No significant difference: $p=0.71$
Look backs overall	19.65 looks; median 18-19	12.82 looks; median 12	Single-factor, ANOVA: $F[1,36]= 15.00$	Highly significant difference: $p < 0.001$
Table Look 1	12/19, 63%	6/16, 37.5%	t-tests, $t(33)= 1.52$	Weak significance: One-tailed: $p=0.07$
Table Look 2	11/20, 55%	9/16, 56%	t-tests, $t(34)= -1.23$	No significance: One-tailed: $p = 0.47$

We measured the percentage of the entire interaction during which the participants were engaged in shared looking. The mover subjects engaged in shared looking with the robot significantly more than the talker subjects (row 2 in Table 2).

However, to understand this effect, it is necessary to look at how much of it is determined by **mutual gaze**. We reasoned that while shared looking differences indicate that something was happening as a result of the robot being able to look around at the subject and the table, the components of that effect were unclear. Mover and talker subjects have only slightly different rates of mutual gaze (which are not statistically significant), measured as a percentage of total interaction time (row 3 of Table 2).

Clearly, mutual gaze does not account for the differences in shared looking. The differences in shared looking then have to do with when the robot and the subject are looking at the table together. Since the talking only version of the robot never looks at the table, it is the fully functional robot that makes the difference in shared looking.

However, additional analyses offer more insight into engaging robots. We discovered that both mover and talker subjects offer their **talk directly** to Mel when they take a turn in the interaction at similar rates. The measure considers averages across all subjects as a percentage of the total interaction time

per subject (row 4 in Table 2). This result greatly surprised us. We did not expect either group to be so conversationally involved with the robot. It seems that a talking head, whether moving around or not, is a compelling conversation partner. However, the features of interaction presented so far do not indicate if the subjects in one or the other condition were affected at all by the gestural abilities of the robot. To consider these several additional aspects of interaction were considered.

One significant difference in behavior is the number of times the subjects **looked back** at the robot when they were looking at the table. Since subjects spend a good proportion of their time looking at the table (55% for movers, 62% for talkers[1]), the fact that they interrupt their table looks to look back to Mel is an indication of how engaged they are with Mel compared with the demonstration objects. All subjects turned their bodies to the demo table when they began interacting with it, so their primary focus, based on body stance, was the table. Mover subjects looked back to the robot far more often than talker subjects did (average number of looks per interaction across subjects, row 5 in Table 2).

Finally, subject behavior was considered during utterances that are not direct commands,[2] but where the robot generally changed its looking. Two declaratives, one with a deictic ("right there") occur as beginnings of the robot's turns: "Right there is the IGlassware cup and near it is the table readout," and "The copper in the glass transmits to the readout display by inductance with the surface of the table." For both of these, the mover robot typically (but not always) turned its head towards and down to the table, while the talker robot never did so.

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

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

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

A talking robot engages people, even if just the head is talking, and no other movement occur. Engagement is compelled because speech and conversation are powerful devices for engaging people in interactions. However, looking gestures provide additional power. They cause people to pay more attention to the robot, and they may also cause people to adjust their looking based on the robot's looking.

5. CONCLUSIONS

The results of this study suggest that there are interactional differences between a robot that uses its body and head to gesture, look at the user and at objects. Gesturing, talking robots capture the user's attention more often, and users seem to respond to changes in head direction and gaze by changing their own gaze or head direction. Users engage in mutual gaze with these robots, direct their gaze to them during turns in the conversation, and follow their commands when asked to perform tasks. Even robots that are just "talking heads" are influential conversational partners. Users mutually gaze at them, talk directly to them when they take a turn in the conversation, and follow their commands.

Users also appear to be sensitive to the appropriateness of gestures and are aware that just a talking head is not what they expect from a 3D conversational participant. The robot must use its body to indicate its attention to the human and to objects of relevance to the interaction. In the coming years, as robot partners in interactions become more commonplace, engagement in interaction, including capturing head gestures, arm gestures, gaze and conversation management in ways that people expect will be the continuing challenge for 3D intelligent user interfaces.

6. FUTURE RESEARCH

A careful reading of the conversation in Figure 3 will reveal that the robot's turns in the conversation are much too long. Human conversation contains much smaller chunks, punctuated by backchannels from the conversation participant. Many backchannels are not spoken but rather gestural; they come in the form of nods. In fact, many of our subjects nodded to Mel, especially during positive response turns. Their behavior suggests that utterance chunks and use of backchannels would produce a more typical conversation style. To recognize nods from users, we are now outfitting Mel with a stereoptic camera, and will make use of head position tracking [12] and an algorithm for recognizing head nods. We will be experimenting with the effects of recognition of nods as well as production of nods by Mel in conversation.

---

[1] The rest of the time subjects either looked elsewhere in the room or looked at the robot when it was looking at the table.

[2] All subjects in both conditions performed the actions expressed in imperative utterances.

Our current gestural rules are still very primitive. First, while we have experimented with how to proceed when the user looks away and does not take a turn, Mel does not change his behavior if a user looks away for a long time (as long as they take their turn in the conversation). Clearly this behavior is faulty. Secondly, from human-human observations [19], we know that people do not track each other at all times. They look away to see what else is going on and to time-share with other tasks they must do. So natural looking is still more complex than Mel currently undertakes. Third, when Mel points, he currently does so with his beak. Recently we outfitted Mel with 2 degrees of freedom in each wing, so that he can point with his wings. However, now we must produce natural gestures for the head and the wing together in pointing (in humans, people look first and bring their arms/hands to point after, but with very close timing between the two).

A mobile robot can engage users to begin conversations as well as to indicate focus of attention during them. We plan to mobilize Mel so that he can attend to users and greet them [3], by not only finding their faces and offering greetings, but by approaching them. In addition, once mobile, Mel will be able to turn to face objects of interest in the conversation. This change will allow us to understand the role of body stance as an indicator of focus of attention.

7. ACKNOWLEDGEMENTS

Thanks to Chuck Rich for assistance on Collagen for Mel and to the anonymous reviewers for ideas on improvements in the paper.

8. REFERENCES

[1] Beardsley, P.A. *Piecode Detection*. Mitsubishi Electric Research Labs TR2003-11, Cambridge, MA, February, 2003.

[2] Breazeal, C. Affective interaction between humans and robots. In *Proceedings of the 2001 European Conference on Artificial Life (ECAL2001*). Prague, Czech Republic, 2001.

[3] A. Bruce, I. Nourbakhsh, R. Simmons. The Role of Expressiveness and Attention in Human Robot Interaction. In *Proceedings of the IEEE International Conference on Robotics and Automation,* Washington DC, May 2002.

[4] Burgard, W., Cremes, A. B., Fox, D., Haehnel, D., Lakemeyer, G., Schulz, D., Steiner, W. & Thrun, S. The Interactive Museum Tour Guide Robot. In *Proceedings of AAAI-98*, pp. 11-18, AAAI Press, Menlo Park, CA, 1998.

[5] Cassell, J., Sullivan, J., Prevost, S. and Churchill, E. *Embodied Conversational Agents*. MIT Press, Cambridge, MA, 2000.

[6] Fong, T., Thorpe, C., and Baur, C. Collaboration, Dialogue and Human-Robot Interaction. In *10th International Symposium of Robotics Research,* Lorne, Victoria, Australia, November, 2001.

[7] Johnson, W.L., Rickel, J.F. and Lester, J.C. Animated Pedagogical Agents: Face-to-Face Interaction in Interactive Learning Environments. *International Journal of Artificial Intelligence in Education*, 11: 47-78, 2000.

[8] Kanda, T., Ishiguro, H., Imai, M., Ono, T., and Mase, K. A constructive approach for developing interactive humanoid robots. In *Proceedings of IROS 2002*, IEEE Press, NY, 2002.

[9] Kidd, C. *Sociable Robots, The Role of Presence and Task in Human-Robot Interaction*. M.S. thesis, MIT Media Laboratory, June 2003.

[10] Lombard, M. and Ditton, T.B. At the heart of it all: the concept of presence. *Journal of Computer-Mediated Communication*, 3(2). University of S. Ca Annenberg School of Communiation, 1997.

[11] Lombard, M., Ditton, T.B., Crane, D., Davis, B., Gil-Egul, G. Horvath, K. and Rossman, J. Measuring presence: a literature-based approach to the development of a standardized paper and pencil instrument. In *Presence 2000: The Third International Workshop on Presence*, Delft, The Netherlands, 2000.

[12] Morency, L.-P.; Rahimi, A.; Darrell, T.; Adaptive view-based appearance models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 1:8-20, June, 2003.

[13] Nikano,Y., Reinstein, G., Stocky, T. Cassell, J. Towards a Model of Face-to-Face Grounding. *In Proceedings of the 41st ACL meeting*, Sapporo, Japan, pp. 553-561, 2003.

[14] Reeves, B. Wise, K,., Maldonado, H., Kogure, K., Sinozawa, K. and Naya, F., Robots Versus On-Screen Agents: Effects on Social and Emotional Responses. In *Proceedings of CHI 2003*, ACM press, 2003.

[15] Rich, C. and Sidner, C.L "COLLAGEN: A Collaboration Manager for Software Interface Agents," *User Modeling and User-Adapted Interaction*, Vol. 8, No. 3/4, 1998, pp. 315-350, 1998.

[16] Rich, C., Sidner, C.L. and Lesh, N. COLLAGEN: Applying Collaborative Discourse Theory to Human-Computer Interaction. *AI Magazine*, *Special Issue on Intelligent User Interfaces*, AAAI Press, Menlo Park, CA, Vol. 22: 4: 15-25, 2001.

[17] Sidner, C.L., Dzikovska, M. Human-Robot Interaction: Engagement Between Humans and Robots for Hosting Activities. In the Proceedings of the *IEEE International Conference on Multimodal Interfaces*, pp. 123-128, 2002.

[18] Sidner, C.L. and Lee, C. *An Architecture for Engagement in Collaborative Conversations between a Robot and Humans*. MERL Technical Report, TR2003-12, June 2003.

[19] Sidner, C.L., Lee, C. and Lesh, N. Engagement when looking: behaviors for robots when collaborating with people. In *Diabruck: Proceedings of the 7th workshop on the Semantic and Pragmatics of Dialogue*, I. Kruiff-Korbayova and C.Kosny (eds.), University of Saarland, pp. 123-130, 2003.

[20] Viola, P. and Jones, M. Rapid Object Detection Using a Boosted Cascade of Simple Features. In the Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition*, Hawaii, pp. 905-910, 2001.