A First Experiment in Engagement for Human-Robot Interaction in Hosting Activities

Candace L. Sidner and Myroslava Dzikovska

TR2003-134   December 2003

Abstract

To participate in conversations with people, robots must not only see and talk with people but make use of the convetnions of conversation and of the means to be conntected to their human counterparts. This paper reports on initial research on engagement in human-human interaction and applications to stationary robots interacting with humans in hosting activities.

Advances in Natural Multimodal Dialogue Systems, Dec 2005

CANDACE L. SIDNER AND MYROSLAVA DZIKOVSKA

A FIRST EXPERIMENT IN ENGAGEMENT FOR HUMAN-ROBOT INTERACTION IN HOSTING ACTIVITIES[1]

Abstract. To participate in conversations with people, robots must not only see and talk with people but also make use of the conventions of conversation and of the means to be connected to their human counterparts. This paper reports on initial research on engagement in human-human interaction and applications to stationary robots interacting with humans in hosting activities.

1. INTRODUCTION

As part of our ongoing research on collaborative interface agents, we have begun to explore engagement in human interaction. Engagement is the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake. This process includes: establishment of the initial contact with another participant, negotiation of a collaboration to undertake activities of mutual interest, determination of the ongoing intent of the other participant to continue in the interaction, evaluation of one's own intentions in staying involved, and determination of when to end the interaction.

To understand the engagement process we are studying human-to-human engagement in interactions. Study of human-to-human engagement provides an understanding of the capabilities required for human-robot interaction. At the same time, experimentation with human-robot interaction provides a valid means to test theories about engagement as well as to produce useful technology results. In this paper we report on our initial experiments in programming a stationary robot to have initial engagement abilities.

2. HOSTING ACTIVITIES

[1] Portions of this paper are reprinted with permission from C. Sidner and M. Dzikovska, "Human-Robot Interaction: Engagement between Humans and Robots for Hosting Activities," The Fourth IEEE International Conference on Multi-modal Interfaces, October, 2002, pages 123-128. @ 2002 IEEE.

T.H.E. Editor(s) (ed.), Book title, 1—6.

The domain of activity in which this research on engagement is framed concerns the activity of hosting. 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 (which may be an artificial or the natural world) and may also request that the user undertake actions to support the fulfilment of those services. Hosting activities are situated or embedded activities, because they depend on the surrounding environment as well as the participants involved. They are social activities because, when undertaken by humans, they depend upon the social roles that people play to determine the choice of the next actions, timing of those actions, and negotiation about the choice of actions. In this research, agents, 2D animated ones or physical robots, who serve as guides, are the hosts of the environment. Tutoring applications require hosting activities; this paper reports on experience with a robot host who is acting as a tutor. Some portions of tutoring, such as testing a student's knowledge, trouble shooting concepts that a student fails to grasp, and keeping track of what a student knows go beyond the informational services of hosting activities. Thus hosting is part of tutoring but not vice versa.

Another common hosting activity is hosting a user in a room with a collection of artifacts. In such an environment, the ability of the host to interact with the physical world and visitors becomes essential, and justifies the creation of physical agents. Room hosting is a core activity in tour guiding in museums, and other indoor and outdoor spaces (see Burgard et al, 1998, for a robot that can guide a museum tour). Sales activities include hosting as part of their mission in order to make customers aware of types of products and features, locations, personnel, and the like. In many activities, hosting may be intermingled with other tasks, e.g. with selling items in retail sales or evaluation tasks in tutoring.

Hosting activities are collaborative because neither party completely determines the goals to be undertaken nor the means of reaching the goal; these must be shared between the parties. While the visitor's interests in the room may seem paramount in determining shared goals, the host's (private) knowledge of the environment also constrains the goals that can be achieved. Typically the goals undertaken will need to be negotiated between visitor and host. Even in tutoring, where the tutor-host's plans for how to tutor the student may seem to drive the interaction, the tutor and student negotiate on the problems they will undertake in their encounter.

This work hypothesizes that by creating computer agents which function more like human hosts, the human participants will focus on the hosting activity and be less distracted by the agent interface. For example, the agent will gaze at the human partner and at domain objects in ways that appropriately indicate the agent's attention to each. When the agent gazes at a partner instead of gazing at an object, that gesture conveys information about the agent's interest in its partner. When a human partner gazes away from the robot or objects of discussion, the robot must be able to assess whether the human has lost interest, and if so, determine how to re-establish or end the engagement between the two participants. Assuring that a robot behaves in ways with which people are familiar in human-to-human interactions increases the likelihood that the interaction will not break down due to the robot's misusing or misunderstanding the cues of engagement.

3. WHAT IS ENGAGEMENT?

Engagement is fundamentally a collaborative process (see Grosz & Sidner, 1990, Grosz & Kraus, 1996), although it also requires significant private planning on the part of each participant in the engagement. Engagement is collaborative principally because the interactors intend to connect together. However, they may be less aware of the actions involved in accomplishing the joint engagement goals, e.g. gaze, head and hand gestures, unlike the conscious actions in other types of collaboration. Engagement, like other types of collaborations, consists of establishing the collaborative goal (the goal to be connected), maintaining the connection, and then ending the engagement. The collaboration process may include negotiation of the goal because a potential collaborator might not decide to become engaged right away or at all. In addition, participants might have to negotiate the means to achieve their goals (Sidner, 1994a,b). For engagement in an interaction, participants negotiate the means for achieving engagement through the various ways they maintain engagement, and repair engagement when it appears to be failing. Described this way, engagement is similar to other collaborative activities.

Engagement is an activity that contributes centrally to collaboration on other activities in the world and the conversations that support them. In fact, conversation is impossible without engagement. This claim does not imply that engagement is just a part of conversation. Rather engagement is a collaborative process that occurs in its own right, simply to establish connection between people, a natural social phenomenon of human existence. It is entirely possible to engage another without a single word being said and to maintain the engagement process with no conversation. That is not to say that engagement is possible without any communication; it is not. A person who engages another without language must rely effectively on some form of gestural communication to establish the engagement joint goal and to maintain the engagement. Gesture is also a significant feature of face-to-face interaction where conversations are present (McNeill, 1992).

Being engaged with another can also be the sole purpose of an interaction. The use of just a few words and gestures can establish and maintain connection with another when no other intended goals are relevant. For example, an exchange of hellos, a brief exchange of eye contact and a set of good-byes can accomplish an interaction just to be engaged. In such interactions, one can reasonably claim that the only purpose is to be connected. The current work focuses on interactions, ones that include conversations, where the participants wish to accomplish action in the world rather than just the relational connection that engagement can provide.

Much of the engagement process can be accomplished by linguistic means only. Evidence for this statement derives from telephone conversations where participants are engaged with each other and have only the words they say, and prosodic effects (pitch, timing, duration, voice quality and the like) to indicate their desire for establishing, continuing and ending their connection to each other. However, in face-to-face interaction, people look at one another as they talk, and make use of gestures to indicate their interest in what the other has to say, to indicate that they wish to continue while at the same time using gestures to access other information in the environment. The engagement process must always balance the need to convey ongoing engagement with the other (or signal its demise) with the need to look at objects in the environment, perform actions called for by the collaboration (in addition to ones that are independent of it), as well as interpreting those gestures from the other participant where the same requirement to balance these needs is in effect.

4. FIRST EXPERIMENT IN HOSTING: A POINTING ROBOT

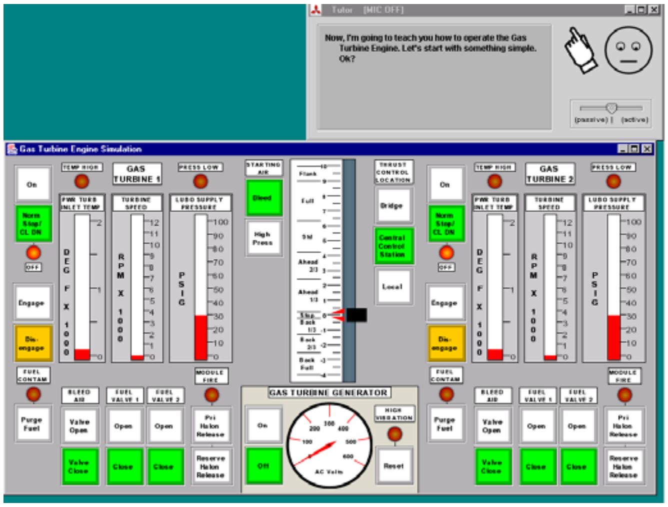
In order to experiment with engagement in hosting activities, this effort began with a well-delimited problem: appropriate pointing and beat gestures for a stationary robot, called Mel, while conducting a conversation. Mel's behavior is a direct product of extensive research on animated pedagogical agents (Johnson et al, 2000). It shares with those agents concerns about conversational signals and pointing. Unlike these efforts, Mel has greater dialogue capability, and its conversational signaling, including deixis, comes from combining the Collagen^{TM} and Rea architectures (Cassell et al, 2001b). Furthermore, while 2D embodied agents (Cassell et al, 2000c) can point to things in a 2D environment, 2D agents cannot effectively point in a 3D space. So it seemed appropriate to explore the effects of deictic behavior with a robot.

To build a robot host, the effort relied significantly on the PACO agent (Rickel et al, 2002) built using Collagen^{TM} (Rich et al, 2001, Rich & Sidner, 1998) for tutoring a user on the operation of a gas turbine engine. The PACO agent tutors a student on the procedures needed to control two engines by their various buttons and dials. Mel served as the tutor in this application and took on the task of speaking all the output and pointing to the portions of the display, tasks normally done by a 2D on-screen agent in the PACO system. The student's operation of the display, through a combination of speech input and mouse clicks, remained unchanged. Understanding of the student's speech was accomplished with the IBM ViaVoice^{TM} speech recognizer, the IBM JSAPI[2] to parse and interpret utterances, and the Collagen^{TM} middleware to provide dialogue

[2] See the ViaVoice SDK, at www4.ibm.com/software/ speech/dev/sdk_java.html.

interpretation and next moves in the conversation, to manage the tutoring goals and to provide a student model for tutoring.

The PACO 2D screen for gas turbine engine tutoring is shown in Figure 1. The agent is represented in a small window, where text, a cursor hand and an iconic face appear. The face changes to indicate six states: the agent is speaking, is listening to the user, is waiting for the user to reply, is thinking, is acting on the interface, and has failed due to a system crash. The cursor hand is used to point out objects in the display.



Now, I'm going to teach you how to operate the Gas Turbine Engine. Let's start with something simple. Ok?

Figure 1. The PACO agent for gas turbine engine tutoring

The robotic agent, Mel, is a stationary robot created at Mitsubishi Electric Research Labs, and consists of 5 servomotors to control the movement of the robot's head, mouth and two appendages. The robot takes the appearance of a penguin. Mel can open and close his beak, move his head in up-down, and left-right combinations, and flap his "wings" up and down. He also has a laser light on his beak, and a speaker provides audio output for him. See Figure 2 for Mel pointing to a button on the gas turbine control panel.

For gas turbine tutoring, Mel sits in front of a large (2 feet x 3 feet) horizontal flat-screen display on which the gas turbine display panel is projected. To conduct a conversation with the student, Mel addresses the student face-on, and beats with his wings at appropriate points in his turn in the conversation. He uses the PACO system to teach the student procedures on the display. When he wishes to point to a button or dial on the display panel, he points with his beak. When he finishes pointing, he addresses the student face-on again. While Mel's motor operations are extremely limited, they offer enough movement to undertake beat gestures, which indicate new and old information in utterances (Cassell et al, 2001a). The head movement is also sufficient to point effectively at objects, so that students can readily see the objects on the panel.

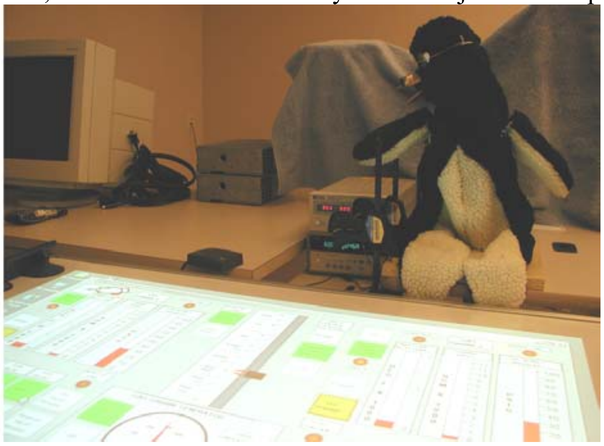


Figure 2. Mel pointing to the gas turbine control panel

The architecture of a Collagen^{TM} agent and an application using Mel is shown in Figure 3. Specifics of the Collagen^{TM} internal organization and the means by which Collagen^{TM} is connected to applications are beyond the scope of this paper; see (Rich & Sidner, 1998; Rich et al, 2001) for more information. Basically, the application is connected to the Collagen^{TM} system through the application adapter. The adapter translates between the semantic events Collagen^{TM} understands and the events/function calls understood by the application. The agent controls the application by sending events to perform to the application, and the adapter sends performed events to Collagen^{TM} when a user performs actions on the application. Collagen^{TM} is notified of the propositions uttered by the agent via uttered events. They also go to the AgentHome window, which is a graphical component responsible in Collagen^{TM} for showing the agent's words on screen as well as generating speech in a speech-enabled system. The shaded area highlights the components that were added to the standard Collagen^{TM} middleware. With these additions, utterance events go through the Mel annotator and the BEAT system (Cassell et al, 2001a) in order to generate gestures as well as the utterances that Collagen already produces. More details on the architecture and Mel's function with it can be found in (Sidner & Dzikovska, 2002).

In tutoring, the Collagen^{TM} architecture is instantiated by means of a detailed set of recipes for the tutoring domain that must be specified in the Planning and Discourse module. Recipes are the means by which the hosting environment is specified for the robot. The recipes do not specify dialogue actions, but instead must detail the actions needed to operate the gas turbine panel and the actions needed to tutor students to use the panel. Additional rules in the Agent help the robot tutor decide what to say or do next, by choosing from a list of next moves created by the Planning and Discourse module. The recipes and rules apply to the domain that is being tutored only, and do not affect the engagement mechanisms that determine the robot's wing and head gestures. The engagement mechanisms are usable in any tutoring activity. However, the tutoring domain recipes do provide a piece of information for the engagement mechanisms, namely what items in the display need to be pointed out. In this way, the current architecture separates linguistic and gesture functions. Thus, like a person, the robot could convey engagement with its linguistic behavior but also convey the desire to disengage with its gestures.

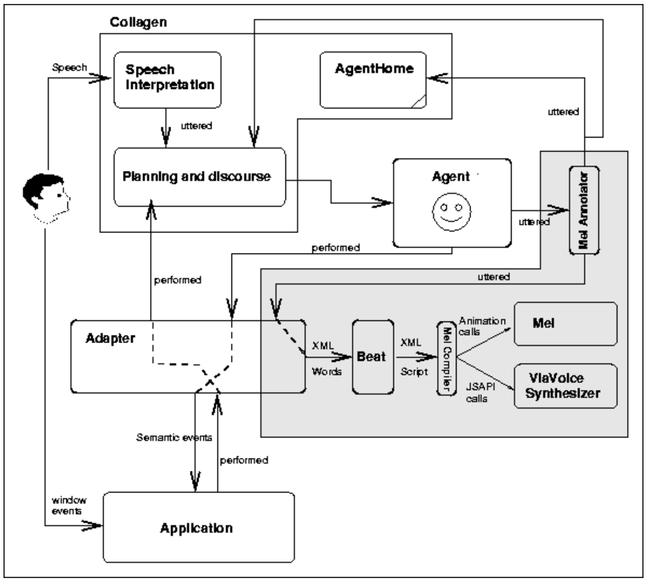


Figure 3: Architecture of Mel

5. MAKING PROGRESS ON HOSTING BEHAVIORS

Mel is quite effective at pointing in a display and producing a gesture that can be readily followed by humans. Mel's beak is a large enough pointer to operate in the way that a finger does. Pointing within a very small margin of error (which is assured by careful calibration before Mel begins talking) makes possible the location of the appropriate buttons and dials on the screen. Mel returns his "gaze" [3] to the student after pointing, which is a signal that he is staying engaged with the user. These two behaviors are a first step in creating engagement. They make it evident to the student that Mel is interacting with the student and that his looks away are not intended to disengage but rather to accomplish a part of the task at hand. However, human engagement is far richer than what this robot can currently do (see Sidner, 2003). Most significantly engagement is a two-part activity: engaging behaviors must be produced, and at the same time, the engaging agent must interpret engaging behaviors from its conversational collaborator.

Two of the most basic aspects of engagement are beginning and ending it. The means by which one begins and ends a conversation with Mel are unsatisfactory. While Mel responds to a student greeting to start the conversation, it does not have any means or goals to decide when and how to begin the conversation itself. Mel also does not know how to end a conversation or when it is appropriate to do so. Furthermore, Mel has only two weak ways of checking on his partner's signals of engagement during their interaction: to ask "okay?" and await a response from the user after every explanation he offers, and to await (including indefinitely) a user response (utterance or action) after each time he instructs the user to act. In human-to-human interactions, engagement activities range over far more linguistic and gestural behaviors. In more recent system building efforts using Mel (see Sidner et al, 2004), Mel produces a much wider range of interactions to gaze and interpret some gazing acts from the human participant, gesture at objects, and begin and end conversations. All these capabilities result from more sophisticated subsystems for Mel (such as vision algorithms for detect human faces and sound location algorithms) as well as careful study of human-human scenarios and video data [Sidner et al, 2003] for determining the types of engagement strategies that humans use effectively in hosting situations.

To understand more about the variety of behaviors to signal engagement, consider the interaction (conversation and gestures) presented in figure 4. It illustrates a constructed engagement scenario with a number of features of the engagement process for room hosting. This scenario was chosen because the visitor is less than enthusiastically engaged in the interaction, and so both the means to stay engaged and to convey less engagement are illustrated. These features of engagement include: failed negotiations of engagement goals, successful rounds of collaboration, conversational capabilities such as turn taking, change of initiative, and negotiation of differences in engagement goals, individual planning and decision making, and execution of end-of-engagement activities. There are also collaborative behaviors that support the action in the world activities (i.e. the domain task) of the participants, in this case touring a room. In a more detailed discussion of this example below, these different collaborations will be distinguished.

[3] Mel's eyes do not move, so to look at the person, the whole head must turn.

Significant to the interaction are the use of intentionally communicative gestures such as pointing and movement, as well as use of eye gaze and recognition of eye gaze to convey engagement or disengagement in the interaction.

In part 1 of this scenario, the visitor in the room hosting does not immediately engage with the host, who uses a greeting and an offer to provide a tour as means of (1) engaging the visitor and (2) proposing a joint activity in the hosting world. Neither the engagement nor the joint activity are accepted by the visitor. The visitor accomplishes this non-acceptance by ignoring the uptake of the engagement activity, which also quashes the tour offer.

However, at part 2, the visitor at her next speaking turn chooses to engage the host in several rounds of questioning, a simple form of collaboration for touring. Questioning maintains the engagement by its very nature. In addition, the visitor’s gaze at the host, as well as her response to requests to follow the host and to look at objects that the host points out are also evidence of engagement. While the scenario does not stipulate much about gaze, in real interactions, much of parts 2 through 6 would include various uses of hands, head turns and eye gaze to maintain engagement as well as to indicate that each participant understood what the other said.

In part 4, the host takes on a new task expressed through the conversation to offer to demonstrate a device in the room; this offer to demonstrate is also an offer to collaborate. The visitor's response is not linguistically complex, but its intent is more challenging to interpret because it conveys that the visitor has not accepted the host's offer and is beginning to negotiate non-performance of the picture taking. The host, a sophisticated negotiator, provides a solution to the visitor's objection, and the demonstration is undertaken. Gestures that typically would accompany the visitor’s utterances would include glances away from the robot with a return only when the host began to speak. Here, negotiation of collaboration on the domain task keeps the engagement happening. A failure to successfully negotiate at this stage would also make it possible to signal the desire to end the interaction, most easily by failing to take a turn or by looking away and then wandering away.

In part 6, the host's next offer is not accepted, not by conversational means, but by lack of response, an indication of disengagement. The host, who could have chosen to re-state his offer (with some persuasive comments), instead takes a simpler negotiation tack and asks what the visitor would like to see. This aspect of the interaction illustrates the private assessment and planning which individual participants undertake in engagement. Essentially, it addresses the private question: what will keep us engaged? With the question directed to the visitor, the host also intends to re-engage the visitor in the interaction, which is minimally successful. The visitor responds but uses the response to indicate that the interaction is drawing to a close. The closing ritual (Schegeloff & Sacks, 1973), a disengagement behavior, is, in fact, odd here, given the overall interaction that has preceded it because the visitor does not follow the American cultural convention of expressing appreciation or offering a simple thanks for the activities performed by the host.

Part 1

<Visitor enters and is looking around the room when host notices visitor. >

Host: Hello, I'm the room host. Would you like me to show you around?

Visitor: <Visitor ignores host and continues to look around.>

Part 2

Visitor: What is this? <Visitor looks at and points to an object.>

Host: That's a camera that allows a computer to see as well as a person to track people as they move around a room.

Visitor: <looks at host> What does it see?

Host: Come over here <Host moves to the direction of the object of interest.> and look at this monitor <points>. It will show you what the camera is seeing and what it identifies at each moment.

Part 3

Visitor: <follows host and then looks at monitor> Uh-huh. What are the boxes around the heads?

Host: The program identifies the most interesting things in the room--faces. That shows it is finding a face.

Visitor: Oh, I see. Well, what else is there?

Part 4

Host: I can show you how to record a photo of yourself as the machine sees you.

Visitor: Well, I don't know. Photos usually look bad.

Host: You can try it and throw away the results.

Part 5

Visitor: Ok. What do I do?

Host: Stand before the camera.

Visitor: Ok.

Host: When you are ready, say "photo now."

Visitor: Ok. Photo now.

Host: Your picture has been taken. It will print on the printer outside this room.

Visitor: Ok.

Part 6

Host: Let's take a look at the multi-level screen over there <points> <then moves toward the screen>.

Visitor: <The visitor does not follow pointing and instead looks in a different direction for an extended period of time.>

Host: <Host notices and decides to see what the visitor is looking at. > Is there something else you want to see?

Visitor: No, I think I’ve seen enough. Bye.

Host: Ok. Bye.

Figure 4. Scenario for Room Hosting

While informal constructed scenarios illustrate some features of engagement, they do not provide a detailed way to understand how people collaborate to stay in connection to one another as they interact. A more solid basis of study of human hosting is needed. To that end, ongoing analysis of several videotaped interactions between human hosts and visitors in a natural hosting situation provides details about the use of gaze in engagement in hosting [Sidner et al, 2003]. In each session, the host was a lab researcher, while the visitor was a guest invited by the first author to visit and see the work going on in the lab. The host demonstrated new technology in a research lab to the visitor for between 28 and 50 minutes, with the variation determined by the host and the equipment available.

6. ENGAGEMENT AMONG HUMAN HOSTS AND VISITORS

The nature of engagement between human hosts and their human visitors provides an informative picture of hosting for humans and robots. First, human-to-human hosting is a common enough activity that many people have participated in such an activity in museums, outdoor tours, and retail settings. Gathering data using videotaping is somewhat intrusive on the typical hosting encounter, but not so much so that people are aware of the taping at all times. So their behavior is a reliable indicator of typical hosting interactions. Second, because hosting involves more than just engagement, that is, the host and visitor have a joint task to perform, namely, to see that the visitor is hosted, it allows researchers to view engagement in a task oriented setting where collaboration on a domain task is ongoing. While this view produces the problem of distinguishing engagement from the hosting collaboration, it also makes it possible to understand how engagement goes on in the context of everyday activities.

Engagement is a collaboration that generally happens together with collaboration on a domain task. In effect, at every moment in the hosting interactions, there are two collaborations happening, one for the participants to accomplish hosting (for example, to tour a lab, which is the domain task), and the other for the participants to stay engaged with each other. While the first collaboration provides evidence for the ongoing process of the second, it is not enough in and of itself. Engagement depends on many gestural actions as well as conversational comments. Furthermore, the initiation of engagement generally takes place before the domain task is explored, and engagement happens when there are no domain tasks being undertaken. Filling out this story is one of our ongoing research tasks.

In the hosting situations observed from videotaped data, engagement begins with two groups of actions. The first is the approach of the two participants accompanied by gazing. Each notices the other. Then, a second group of actions takes place, namely those for opening ritual greetings (Luger, 1983), name introductions and handshakes. Introductions and handshakes are customary American rituals that follow greetings between strangers. For people, who are familiar with one another, engagement can begin with an approach, gaze at the potential partner and optionally a mere "hi." These brief descriptions of approach and opening rituals only begin to describe some of the variety in these activities. The salient point is that approach is a collaboration because the two participants must achieve mutual notice. The critical point about openings is that an opening ritual is necessary to establish connection, and hence is part of the engagement process.

All collaboration initiations can be thwarted, and the same is true of the collaboration for engagement, as is illustrated in the constructed scenario in Figure 4 in part 1. However, in the videotaped sessions, no such failures occur, in large part due to the participants having pre-agreed to the videotaped encounter.

Once connected, collaborators must find ways to stay connected. In relational only encounters, eye gaze, smiles and other gestures may suffice. However, for domain tasks, the collaborators begin the collaboration on the domain task. Collaborations always have a beginning phase where the goal is established, and proposing the domain task goal is a typical way to begin the domain collaboration. In the videotaped hosting activities, the participants have been set up in advance (as part of the arrangement to videotape them) to participate in hosting, so they do not need to establish this goal. They instead check that the hosting is still their goal and then proceed. The host performs his part by showing several demos of prototype systems. In three of the videotaped sessions, the host (who is the same person in all the sessions) utters some variant of "Let's go see some demos." This check on whether hosting has started is accompanied by looking at the visitor, smiles and in some cases, a sweep of the hand and arm, which appears to indicate either conveying a direction to go in or offering a presentation.

How do participants in a domain task collaboration know that the engagement process is succeeding, that is, that the participants are continuing to engage each other? When participants perform the actions to accomplish a domain task collaboration, they have evidence that the engagement is ongoing by virtue of what is said and done in the domain task collaboration. In addition, other behaviors provide signals between the participants that they are still engaged. These signals are not necessary, as is evidenced by the fact that participants have done tasks in laboratories with only computer terminal contact between participants. However, without these signals, the collaboration is a slow, inefficient enterprise and likely to breakdown because at least some actions can be interpreted as not continuing engagement or participation in the domain task. Some of these signals are also essential in conversation for the same reason. Furthermore, when misused, these behaviors indicate that engagement is somehow off the track. The signals include:

- talking about and performing the task,
- turn taking (Clark, 1996)
- timing (i.e. the pace of uptake of a turn),
- use of gaze at the speaker, gaze away for taking turns (Duncan, 1974; Cassell, 2000b),
- use of gaze at speaker to track speaker gestures with objects,
- use of gaze by speaker or non-speaker to check on the attention of other,
- hand gestures for pointing, iconic description, beat gestures, etc (see Cassell, 2000a; Johnson et al, 2000), and in the hosting setting, gestures associated with domain objects,
- head gestures (e.g. nods, shakes, sideways turns)
- body stance (i.e. facing towards the other, turning away, standing up when previously sitting and sitting down),
- facial gestures (not explored in this work but see Pelachaud et al, 1996),
- non-linguistic auditory responses (e.g. snorts, laughs),
- social relational activities (e.g. telling jokes, role playing, supportive rejoinders).

Several of these signals have been investigated by other researchers, and hence only a few are discussed here. The pace in the uptake of a turn concerns the delay between the end of one participant's utterances and the next participant's start at speaking. It appears that participants have expectations about next speech occurring at an expected interval. They take variations to mean something. In particular, delays in uptake can be signals of disengagement or of conversational difficulties. Due to this ambiguity, uptake delay clearly signals disengagement only when other cues also indicate the possibility of disengagement: looking away, walking away, or turning one’s body away from the other participant.

In some hosting situations, domain activities can require the use of hands (and other parts of the body) to operate equipment or display objects. In the videotaped sessions, the host often turns to a piece of equipment to operate it as part of a demonstration. The visitors interpret these extended periods of attention to something other than the visitor as part of the domain task collaboration, and hence do not take their existence as evidence that the performer is distracted from the task and the engagement. The important point here is that when relevant to the domain task, gestures related to operating equipment and object display indicate that the collaboration is continuing, and no disengagement is occurring. When they are not relevant to the domain task, they could be indicators that the performer is no longer engaged, but further study is needed to gauge this circumstance. This observation can be taken as indicative of a principle of engagement: Activities relevant to the domain collaboration provide evidence for the continuance of engagement.

Hosting activities seem to bring out what will be called *social relational activities*, that is, activities that are not essential for the domain task, but seem social in nature, and yet occur during it with some thread of relevance to the task. (Bickmore, 2003) notes that social dialogue (even without the performance of accompanying physical actions) increases the trust between dialogue participants. The hosts and visitors in the videotaped sessions tell humorous stories, offer rejoinders or replies that go beyond conveying that the information just offered was understood, and even take on role playing with the host and the objects being exhibited. Figure 5 contains a portion of a transcript of one hosting session. In that session, the visitor and the host spontaneously play the part of two children using the special restaurant table that the host was demonstrating. The reader should note that their play is coordinated and interactive and is not discussed before it occurs. The role-playing begins at 10 in the figure and ends at 17. This segment of the transcript is preceded by the host P having shown the visitor C how restaurant customers order food in an imaginary restaurant using an actual electronic table, and having explained how waitstaff might use the new electronic table to assist customers. Note that utterances by P and C are labeled with their letter, a colon, and italics, while other material describes their body actions.

54: P turns head/eyes to C, raises hands up.
C's head down, eyes on table.
55: P moves away from C and table, raises hands and shakes them;
moves totally away, fully upright .
56: P: *Uh and show you how the system all works*
C looks at P and nods.
58: P sits down.
P: *ah*
00: P: *ah another aspect that we're*
P rotates each hand in coordination.
C looks at P.
01: P: *worried about*
P shakes hands.
02: P: *you know*

C nods.
04: P: *sort of a you know this would fit very nicely in a sort of theme restaurant*
P looks at C; looks down.
05: C: *MM-hm*
C looks at P, nods at "MM-hm."
P: *where you have lots of*
06: P draws hands back to chest while looking at C.
C: *MM-hm*
P: *kids*
C nods, looking at P.
07: P: *I have kids. If you brought them to a*
P has hands out and open, looks down then at C.
C still nods, looking at P.
09: P: *restaurant like this*
P brings hands back to chest.
C smiles and looks at P.
10: P looks down; at "oh oh" lunges out with arm and together points to table and looks at table.
P: *they would go oh oh*
11: C: *one of these, one of these, one of these*
C points at each phrase above and looks at table.
P laughs.
13: P: *I want ice cream* <point>*, I want cake* <point>
C: *yes yes* <simultaneous with "cake">
C points at "cake" looks at P, then brushes hair back.
P looking at table.
15: P: *pizza* <points>
P looking at table.
C: *Yes yes French fries* <point>
C looks at table as starts to point.
16: P: *one of everything*
P pulls hands back ,looks at C.
C: *yes*
C looks at P.
17: P: a*nd if the system just ordered {stuff} right then and there*
P looks at C, hands out and {shakes}, shakes again after "there."
C looking at P; brushes hair.
C: *Right right* (said after "there")
20: P: *you'd be in big trouble* || <laughs>
P looking at C and shakes hands again in same way as before.
C looking at P, nods at ||.

23: C: *But your kids would be ecstatic*
C looking at P.
P looks at C, puts hands in lap

Figure 5: A Playtime Example

One might argue that social relational activities occur to support other relational goals between participants in the engagement and domain task. In particular, many researchers claim that participants are managing their social encounters, their "social face," or their trust (Bickmore & Cassell, 2001; Katagiri et al, 2001) in each other, in addition to achieving some task domain goals. Social relational activities may occur to support these concerns. However, one need not analyze the details of the social model for face management, or other interpersonal issues such as trust, in order to note that either indirectly as part of social management, or directly for engagement, the social relational activities observed in the videotaped sessions contribute to maintaining the connection between the participants. Social relational activities such as the role playing in Figure 5 allow participants to demonstrate that they are socially connected to one another in a demonstrable way. They are more than just looking at each other and nodding to one another, especially to accomplish their domain goals. They actively seek ways to indicate to the other that they have some social relation to each other. Telling jokes to amuse and entertain, conveying empathy in rejoinders or replies to stories, and playing roles are all means to indicate social relational connection. In sum, relational connection is evidence that engagement is ongoing.

The challenge for participants in collaborations on domain tasks is to weave the relational connection into the domain collaboration. Alternatively participants can mark a break in the collaboration to tell stories or jokes. In the hosting events studied here, the subjects are very facile at accomplishing the integration of social relational connection and the domain task collaboration.

All collaborations have an end, either because the participants give up on the goal (c.f. Cohen & Levesque, 1990), or because the collaboration succeeds in achieving the desired goals. When collaboration on a domain task ends, participants can elect to negotiate an additional task collaboration or refrain from doing so. When they refrain, they then undertake to close their interaction and end the engagement. Their means to do so are presumably as varied as the rituals to begin engagement, but the common patterns prevail for pre-closing, expressing appreciation, saying goodbye, with an optional handshake, and then moving away from one another. Preclosings (Schegeloff & Sacks, 1973) convey that the end is coming. Expressing appreciation is part of a socially determined custom in the US (and many other cultures) when someone has performed a service for an individual. In the hosting data, the visitor expresses appreciation, with acknowledgement of the host. Where the host has had some role in persuading the visitor to participate, the host may express appreciation as part of the preclosing. Moving away is a strong cue that the disengagement has taken place.

Collaboration on engagement transpires before, during and after collaboration on a domain task. So what theoretical models will help explain this multi-collaboration process? One might want to argue that more complex machinery is needed than that so far suggested in conversational models of collaboration (cf. Grosz & Sidner, 1990, Grosz & Kraus 1996, Lochbaum, 1998). However, we believe it is possible to account for engagement within this framework and to use the framework to develop a working computer agent, in particular a robot. In the conversational models of collaboration, collaboration on a domain task or tasks proceeds from group activity and is accompanied by conversation. The conversation reflects the tasks being undertaken through the structure of the segments of the conversation and the intentions conveyed by participants in those segments. Tasks, or as the theory dubs them, *goals* are modeled by a set of recipes that specify how actions are performed in the domain to achieve the goal. Actions in the recipe can be performed by either participant, and the participants are presumed to mutually believe or come to mutually believe the recipes of the collaboration. Participants also come to believe individual intentions to perform actions in the recipe. The theory assumes that each participant uses the actions and recipes (1) to recognize how actions by the other participant contribute to the goal and (2) to plan his or her own acts. Conversational collaboration theory does not specifically consider the nature of collaboration for engagement as part of conversation, but the theory and model are specified in a generic way that should also apply to engagement as a collaboration.

To apply this theory to engagement, our challenge is to specify the set of rules and recipes that participants in hosting believe will achieve the goals of starting, maintaining and ending engagement. Furthermore, to express this theory computationally, a computational participant (such as a robot) must be able to recognize actions that use those recipes. Clearly recipes for the opening and closing of a conversation as a means of starting and ending engagement can be expressed in terms of actions on the part of the participants in a domain task collaboration. What remains to be discovered is the sets of actions and action groups that form the process of maintaining engagement during a domain task collaboration. In particular, turns in the conversation about the domain task as well as certain gaze, body stance and pointing gestures form the class of engagement actions. The exact composition of that class is as yet unclear. What is clear is that the robot has a two-part task: to engage with the visitor and to track engagement behaviors from the visitor.

Finally, social relational behaviors play a part in both the domain collaboration and the engagement process. How does one account for the social relational behaviors discussed above in collaboration theory? While social relational behaviors also tell participants that their counterparts are engaged, they are enacted in the context of the domain task collaboration, and hence must function with the mechanisms for that purpose. Intermixing relational connection, a social goal, and domain collaboration are feasible in collaboration theory models. In particular, the goal of making a relational connection can be accomplished via actions that *contribute* to the goal of the domain task collaboration. However, each collaborator must ascertain through presumably complex reasoning that the actions (and associated recipes) will serve their social goals as well as contribute to the domain goals. Hence they must choose actions that contribute to social goals as well as domain goals. Then they must also ascertain that the social goals are compatible with ongoing engagement collaboration. Furthermore, they must undertake these goals jointly.

The remarkable aspect of the playtime example is that the participants do not explicitly agree to demonstrate how kids will act in the restaurant. Rather the host, who has previously demonstrated other aspects of eating in the electronic restaurant, relates the problem of children in a restaurant and begins to demonstrate the matter when the visitor jumps in and participates jointly. The host accepts this participation by simply continuing his part in it. It appears that they are jointly participating in the hosting goal, but at the same time they are also participating jointly in a social interaction. The details that describe and explain how hosting agents and visitors accomplish this second collaboration are an important goal of ongoing research.

Presumably not all social behaviors can be interpreted in the context of the domain task. Sometimes participants interrupt their collaborations to tell a story that is either not pertinent to the collaboration or while pertinent, is somehow out of order. These stories are interruptions of the current domain task collaboration and are understood as having some other conversational purpose. As interruptions, they signal that engagement is happening as expected as long as the conversational details of the interruption operate to signal engagement. It is not interruptions in general that signal disengagement or a desire to move to disengage; it is failure to take up the interruption that signals disengagement possibilities.

6.1 OPEN QUESTIONS

The discussion above raises a number of questions that must be addressed in ongoing work. First, in the video data, the host and visitor often look away from each other at non-turn taking times, especially when they are displaying or using demo objects. They also look up or towards the other's face in the midst of demo activities. The conversational collaboration model does not account for the kind of fine detail required to explain gaze changes, nor do the standard models of turn taking. How are we to account for these gaze changes as part of engagement? What drives collaborators to gaze away and back when undertaking actions with objects so that they and their collaborators remain engaged?

Second, in the data, participants do not always explicitly acknowledge or accept what another participant has uttered. Sometimes they use laughs, snorts or expressions of surprise (such as "wow") to indicate that they have heard and understood and even confirm what another has said. These verbal expressions are appropriate because they express appreciation of a joke, a humorous story or outcome of a demo. We are interested in the range and character of these phenomena as well as how they are generated and interpreted.

Third, this paper argues that much of engagement can be modeled using the computational collaboration theory model of (Grosz & Sidner, 1990, Grosz & Kraus 1996, Lochbaum, 1998). However, a fuller computational picture is needed to explain how participants decide to signal engagement as continuing and how to recognize these signals.

7. A NEXT GENERATION MEL

While pursuing theory of human-human engagement, we continue to add new capabilities for Mel that are founded on human communication. To accomplish this, the next generation Mel combines hosting conversations with other research at MERL on face tracking and sound location (Sidner et al, 2004). This combination makes it possible to locate visitors and then greet them in ways similar to human experience. These vision and algorithms as well as others permit Mel make use of nodding and gaze change (though not what a human gazes at), which are important indicators of conversation for turn taking as well as expressions of attention. Mel's architecture continues to evolve to the point that Mel has both a "brain," performing Collagen^{TM} related functions and a "body," fusing sensory data to feed to the brain and controlling Mel's motions. Building a robot that can detect faces, track them and notice when the face disengages for a brief or extended period of time demonstrates more engagement behavior than has been possible before.

One challenge for a robot host is to experiment with techniques for dealing with unexpected speech input. People, it is said, say that darndest things. While the Collagen^{TM} middleware continues to be used for modeling conversation, the struggle goes on to find reasonable behaviors for unexpected visitor utterances. For example, when demonstrating a device that requires filling a cup with water, a visitor may make a mistake and spill water on the floor or table and say "Oops, I spilled water on the floor." To understand this utterance, the speech recognizer must correctly process the words, and the sentence semantics must give it a meaningful description, after which the conversation engine must determine the purpose of the meaning description. Finally Mel must respond to it. If this sort of utterance was not predicted to occur (and there will be many such utterances), the best response that Mel can currently produce is "I do not understand the purpose of your utterance. Please find a human to help me." Even that response is only possible if Mel has understood the visitor utterance up to its purposive intent. Failures at speech recognition or sentence interpretation will produce even less informative error messages. These difficulties result from the limits of current speech understanding technology and continue to limit the naturalness of interaction with Mel.

8. SUMMARY

Hosting activities are a natural and common activity among humans, and one that can be accommodated by human-robot interaction. Making the human-machine experience natural requires understanding the nature of engagement and applying the same types of human engagement behavior to robot participation in hosting activities. Engagement is a collaborative activity that is accomplished through both linguistic and gestural means. The experiments described in this paper with a stationary robot that can converse and point provide an initial example of an engaged conversationalist. Through study of human-human hosting activities, new models of engagement for human-robot hosting interaction will provide us with a more detailed means of interacting between humans and robots.

9. ACKNOWLEDGMENTS

The authors wish to acknowledge the work of Paul Dietz on Mel, Neal Lesh, Charles Rich, and Jeff Rickel on Collagen and PACO.

10. REFERENCES

Bickmore, T. (2003). *Relational Agents: Effecting Change through Human-Computer Relationships*. PhD. Thesis, Media Arts & Sciences, Massachusetts Institute of Technology.

Bickmore, T. and Cassell, J. (2001). Relational Agents: A model and implementation of building user trust. *Proceedings of CHI-2001*. New York: ACM Press pp. 396-403.

Burgard, W., Cremes, A. B. , Fox, D., Haehnel, D., Lakemeyer, G., Schulz, D., Steiner, W. & Thrun, S. (1998). The Interactive museum tour guide robot. *Proceedings of AAAI-98*. Menlo Park, CA: AAAI Press, pp. 11-18.

Cassell, J. (2000a). Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents. in *Embodied Conversational Agents*, J. Cassell, J. Sullivan, S. Prevost, and E. Churchill (eds.), Cambridge, MA: MIT Press.

Cassell, J., Bickmore, T., Campbell, L., Vilhjálmsson, H., & Yan, H. (2000b) Human conversation as a system framework: Designing embodied conversational agents," in *Embodied Conversational Agents*, J. Cassell, J. Sullivan, S. Prevost, and E. Churchill (eds.), Cambridge, MA: MIT Press.

Cassell, J. , Sullivan, J. , Prevost, S., & Churchill, E. (2000c). *Embodied conversational agents*. Cambridge, MA: MIT Press.

Cassell, J., Vilhjálmsson, H., & Bickmore, T. W. (2001a). BEAT: the behavior expression animation toolkit. *Proceedings of SIGGRAPH 2001*. New York: ACM Press. pp. 477-486.

Cassell, J., Nakano, Y. I., Bickmore , T. W. , Sidner, C. L. & Rich, C. (2001b). Non-Verbal Cues for Discourse Structure. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics,* Menlo Park, CA: Morgan Kaufman Publishers. pp. 106-115.

Clark, H.H. (1996). *Using Language*, Cambridge University Press, Cambridge.

Cohen, P. & Levesque, H. (1990). Persistence, commitment and intention. in *Intentions in Communication*, P. Cohen, J. Morgan and M.E. Pollack (eds.), Cambridge, MA: MIT Press.

Duncan, S. (1974). Some signals and rules for taking speaking turns in conversation. in *Nonverbal Communication*, S. Weitz (ed.), New York: Oxford University Press.

Grosz, B.J. & Sidner, C. L. (1990). Plans for discourse. in P. Cohen, J. Morgan, & M..Pollack (eds.), *Intentions and Plans in Communication and Discourse.* Cambridge: MIT Press.

Grosz, B. J. & Kraus, S. (1996). Collaborative plans for complex group action. *Artificial Intelligence*, 86(2): 269-357.

Johnson, W. L. , Rickel, J. W. & Lester, J.C. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments,. *International Journal of Artificial Intelligence in Education*, 11: 47-78.

Katagiri, Y., Takahashi, T. & Takeuchi, Y. (2001). Social persuasion in human-agent interaction*. Second IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems, IJCAI-2001*. Menlo Park, CA: Morgan Kaufman Publishers. pp. 64-69.

Lochbaum, K.E. (1998). A collaborative planning model of intentional structure. *Computational Linguistics*, 24(4): 525-572.

Luger, H.H. (1983). Some aspects of ritual communication. *Journal of Pragmatics.* 7(3): 695-711.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press.

Pelachaud, C., Badler, N., & Steedman, M. (1996). Generating facial expressions for speech. *Cognitive Science*, 20(1):1-46.

Rich, C. & Sidner, C. L. (1998). COLLAGEN: A collaboration manager for software interface agents," *User Modeling and User-Adapted Interaction*. 8(3/4):315-350.

Rich, C., Sidner, C.L., & Lesh, N. (2001). COLLAGEN: Applying collaborative discourse theory to human-computer interaction. *AI Magazine, Special Issue on Intelligent User Interfaces*. Menlo Park, CA: AAAI Press, 22(4): 15-25.

Rickel, J., Lesh, N. , Rich, C., Sidner , C. L. & Gertner, A. (2002). Collaborative discourse theory as a foundation for tutorial dialogue. *Proceedings of Intelligent Tutoring Systems*. New York: ACM Pres*s*.

Schegeloff, E. & Sacks, H. (1973). Opening up closing. *Semiotica*, 7(4): 289-327.

Sidner, C. L. (1994a). An artificial discourse language for collaborative negotiation. *Proceedings of the Twelfth National Conference on Artificial Intelligence*. Cambridge, MA : MIT Press. Vol. 1, pp. 814-819.

Sidner. C. L. (1994b). Negotiation in collaborative activity: a discourse analysis. *Knowledge-Based Systems*, 7(4): 265-167.

Sidner, C. L. & Dzikovska, M. (2002). Hosting activities: Experience with and future directions for a robot agent host. *Proceedings of the 2002 Conference on Intelligent User Interfaces*, New York: ACM Press. pp. 143-150.

Sidner, C.L.; Lee, C.; & Lesh, N.(2003). Engagement by Looking: Behaviors for Robots when Collaborating with People. In Kruiff-Korbayova and Kosny (eds.), *DiaBruck: The Proceedings of the Seventh Workshop on the Semantics and Pragmatics of Dialogue*. University of Saarland, 123-130.

Sidner, C.L.; Kidd, C., Lee, C.; & Lesh, N. (2004). Where to Look: A Study of Human-Robot Engagement, in Proceedings of the 2004 International Conference on Intelligent User Interfaces, ACM Press, (forthcoming).

11. AFFILIATIONS

Candace L. Sidner, Mitsubishi Electric Research Labs, 201 Broadway, Cambridge, MA 02139, email: sidner@merl.com.

Myroslava Dzikovska, Dept. of Computer Science, University of Rochester, Rochester, NY 14627, email: myros@cs.rochester.edu.