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

Engagement between Humans and Robots for Hosting Activities

Candace L. Sidner and Myrosia Dzikovska

TR2002-37 December 2002

Abstract

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

Presented at the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems, Copenhagen. A shortened version of this paper, joint with M. Dzikovska, will be presented at the International Conference on Multimodal Interfaces, October 2002.

Presented June 2002;

Engagement between Humans and Robots for Hosting Activities[1]

Candace L. Sidner
Mitsubishi Electric Res. Labs
201 Broadway
Cambridge,MA 02139
Sidner@merl.com

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

Abstract

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

Keywords: Human-robot interaction, hosting activities, engagement, conversation, collaborative interface agents, embodied agents.

1. INTRODUCTION

As a result of ongoing research on collaborative interface agents, including 3D robotic ones, I have begun exploring the problem of engagement in human interaction. Engagement is 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 interaction, evaluating staying involved, and deciding when to end connection.

To understand the engagement process I am studying human to human engagement interaction. Study of human to human engagement provides essential capabilities for human - robot interaction, which I view as a valid means to test theories about engagement as well as to produce useful technology results. My group has been experimenting with programming a (non-mobile) robot with engagement abilities.

2. HOSTING ACTIVITIES

My study of engagement centers on 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 human user undertake actions to support the fulfillment of those services. Hosting activities are situated or embedded activities, because they depend on the surrounding environment as well as the participants involved. They are social activities because, when undertaken by humans, they depend upon the social roles of humans to determine next actions, timing of actions, and negotiation among the choice of actions. Agents, 2D animated or physical robots, who serve as guides, are the hosts of the environment. This work hypothesizes that by creating computer agents that can function more like human hosts, the human participants will focus on the hosting activity and be less distracted by the agent interface. Tutoring applications require hosting activities; I have experimented with a robot host in tutoring, which is discussed in the next section.

Another hosting activity, which I am currently exploring, 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 becomes essential, and justifies the creation of physical agents. Other activities include hosting as part of their mission: sales activities of all sorts include hosting in order to make customers aware of types of products and features, locations, personnel, and the like. In these activities, hosting may be intermingled with selling or instructional tasks. Activities such as tour guiding or serving as a museum docent are primarily hosting activities (see [1] for a robot that can perform tour guide hosting).

Hosting activities are collaborative because neither party determines completely the goals to be undertaken. While the user's interests in the room are paramount in determining shared goals, the host's (private) knowledge of the environment also con-

---

[1] This paper will be given at the International CLASS Workshop on "Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems," Copenhagen, Denmark 28-29 June 2002.

strains the goals that can be achieved. Typically the goals undertaken will need to be negotiated between user and host. Tutoring offers a counterpart to room exploration because the host has a rather detailed private tutoring agenda that includes the user attaining skills. Hence the host must not only negotiate based on the user's interest but also based on its own (private) educational goals. Accordingly the host's assessment of the interaction is rather different in these two example activities.

3. WHAT'S ENGAGEMENT ABOUT?

Engagement is fundamentally a collaborative process (see [2], [3]), although it also requires significant private planning on the part of each participant in the engagement. Engagement, like other collaborations, consists of rounds of establishing the collaborative goal (the goal to be connected), which is not always taken up by a potential collaborator, maintaining the connection by various means, and then ending the engagement or opting out of it. The collaboration process may include negotiation of the goal or the means to achieve it [4], [5]. Described this way, engagement is similar to other collaborative activities.

Engagement is an activity that contributes centrally to collaboration on 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 gestural language 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 [6].

It is also possible to use language and just a few words to create and maintain connection with another, with no other intended goals. An exchange of hellos, a brief exchange of eye contact and a set of good-byes can accomplish a collaboration to be in connection to another, that is, to accomplish engagement. These are conversations for which one can reasonably claim that the only purpose is simply to be connected. The current work focuses on interactions, ones including conversations, where the participants wish to accomplish action in the world rather than just the relational connection that engagement can provide.

4. FIRST EXPERIMENT IN HOSTING: A POINTING ROBOT

In order to explore hosting activities and the nature of engagement, the work began with a well-delimited problem: appropriate pointing and beat gestures for a (non-mobile) robot, called Mel, while conducting a conversation. Mel's behavior is a direct product of extensive research on animated pedagogical agents [7]. It shares with those agents concerns about conversational signals and pointing as well. Unlike these efforts, Mel has greater dialogue capability, and its conversational signaling, including deixis, comes from combining the Collagen™ and Rea architectures [8]. Furthermore, while 2D embodied agents [9] can point to things in a 2D environment, 2D agents do not effectively do 3D pointing.

Building a robot host relied significantly on the Paco agent [10] built using Collagen™ [11,12] for tutoring a user on the operation of a gas turbine engine. Thus Mel took on the task of speaking all the output of the Paco system, a 2D application normally done with an on-screen agent, and pointing to the portions of the display, as done by the Paco agent. The user's operation of the display through a combination of speech input and mouse clicks remains unchanged. The speech understanding is accomplished with IBM ViaVoice™'s speech recognizer, the IBM JSAPI (see the ViaVoice SDK, at www4.ibm.com/software/ speech/dev/sdk_java.html) to parse utterances, and the Collagen middleware to provide interpretation of 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. Note that the agent is represented by a small window, where text, a cursor hand and a smiling face appear (the cursor hand, however, is pointing at a button at the bottom of the screen in the figure). 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.

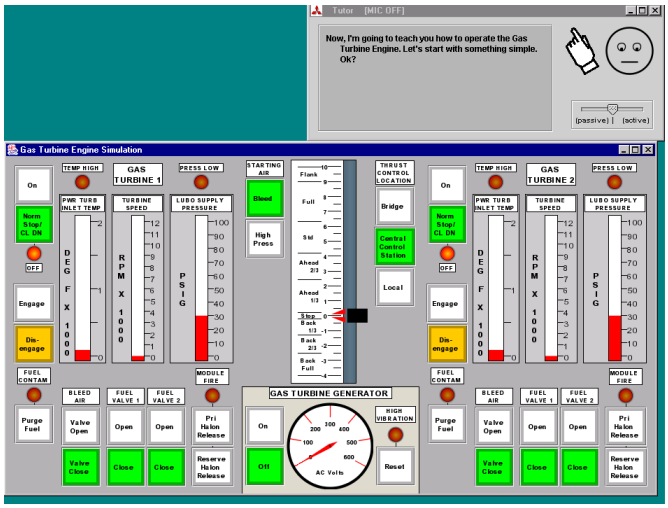

Figure 1:The Paco agent for gas turbine engine tutoring

Our robotic agent is a homegrown non-mobile robot created at Mitsubishi Electric Research Labs [Paul Dietz, personal communication], consisting of 5 servomotors to control the movement of the robot's head, mouth and two appendages. The robot takes the appearance of a penguin (called Mel). 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.

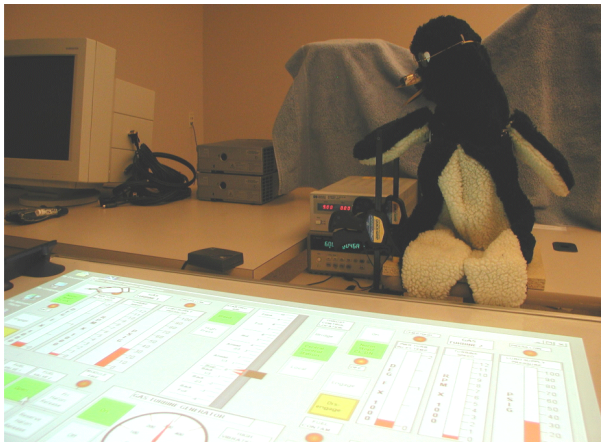

Figure 2: Mel pointing to the gas turbine control panel

While Mel's motor operations are extremely limited, they offer enough movement to undertake beat gestures, which indicate new and old information in utterances [13], and a means to point deictically at objects with its beak. For gas turbine tutoring, Mel sits in front of a large (2 foot x 3 foot) horizontal flat-screen display on which the gas turbine display panel is projected.

All speech activities normally done by the on-screen agent, as well as pointing to screen objects, are instead performed by Mel. With his wings, Mel can convey beat gestures, which the on-screen agent does not. Mel does not however change his face as the onscreen agent does. Mel points with his beak and turns his head towards the user to conduct the conversation when he is not pointing.

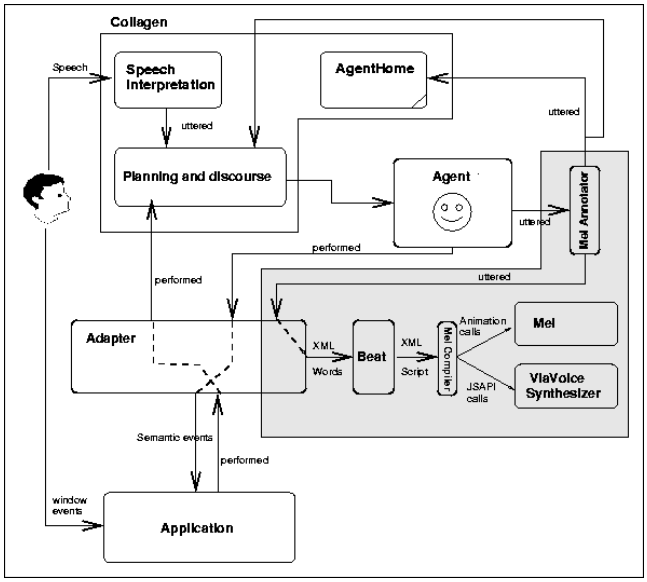

Figure 3: Architecture of Mel

The architecture of a Collagen agent and an application using Mel is shown in figure 3. Specifics of Collagen internal organization and the way it is generally connected to the applications are beyond the scope of this paper; see [11] for more information. Basically, he application is connected to the Collagen system through the application adapter. The adapter translates between the semantic events Collagen 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 when a user performs actions on the application. Collagen 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 for showing the agent's words on screen as well as generating speech in a speech-enabled system. The shaded area highlights the components and events that were added to the basic Collagen middleware. With these additions, utterance events go through the Mel annotator and BEAT system [13] 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 [14].

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 working) locates the appropriate buttons and dials on the screen. However, the means by which one begins a conversation with Mel and ends it are unsatisfactory. Furthermore, Mel has only two weak means of checking on engagement during the conversation: to ask "okay?" and await a response from the user after every explanation it offers, and to await (including indefinitely) a user response (utterance or action) after each time it instructs the user to act.

To expand these capabilities I am studying human-human scenarios to determine what types of engagement strategies humans use effectively in hosting situations.

Figure 4 provides a constructed engagement scenario that illustrates a number of features of the engagement process for room hosting. These 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 assessing and planning, and execution of end-of-engagement activities. There are also collaborative behaviors that support the action in the world activities (called 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. 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 this scenario in part 1 the visitor in the room hosting activity 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 world. Both the engagement and the joint activity are not 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, the visitor at the next turn finally chooses to engage the host in several rounds of questioning, a simple form of collaboration for touring. Questioning also maintains the engagement by its very nature, but also because the visitor performs such activities as going where the host requests in part 2. While the scenario does not stipulate gaze and tracking, 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 over the initiative in the conversation and offers to demonstrate a device in the room; this is another 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 a different outcome. The host, a sophisticated negotiator, provides a solution to the visitor's objection, and the demonstration is undertaken. Here, negotiation of collaboration on the domain task keeps the engagement happening.

However, 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 [14], a disengagement event, is, in fact, odd given the overall interaction that has preceded it because the visitor does not follow the American cultural convention of expressing appreciation or at least offering a simple thanks for the activities performed by the host.

Part 0

<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?

Part 1

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

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.

Part 2

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 can provide us with some features of engagement, a more solid basis of study of human hosting is needed. To that end I am currently collecting several videotaped interactions between human hosts and visitors in a natural hosting situation. In each session, the host is a lab researcher, while the visitor is a guest invited by the author to come and see the work going on in the lab. The host demonstrates new technology in a research lab to the visitor for between 28 and 50 minutes, with variation determined by the host and the equipment available.

6. ENGAGEMENT AMONG HUMAN HOSTS AND VISITORS

This section discusses engagement among people in hosting settings and draws on videotaped interactions collected at MERL.

Engagement is a collaboration that largely happens together with collaboration on a domain task. In effect, at every moment in the hosting interactions, there are two collaborations happening, one to tour a lab and the other to stay engaged with each other. While the first collaboration provides evidence for ongoing process of the second, it is not enough. Engagement appears to depend 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 not domain tasks being undertaken. Filling out this story is one of my ongoing research tasks.

In the hosting situations I have observed, engagement begins with two groups of actions. The first is the approach of the two participants accompanied by gaze at the other. Each notices the other. Then, the second group of actions takes place, namely those for opening ritual greetings [15], name introductions and hand shakes. Introductions and hand shakes 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 approach is that it 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 parts 0 and 1. However, in the videotaped sessions, no such failures occur, in large part, I surmise, due to the circumstances of the pre-agreement 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 a 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 starting hosting 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 collaboration know that the engagement process is succeeding, that the participants are continuing to engage each other? When participants follow the shared recipes for a domain collaboration, they have evidence that the engagement is ongoing by virtue of the domain collaboration. However, many additional behaviors provide signals between the participants that they are still engaged. These signals are not necessary, but without them, the collaboration is a

slow and inefficient enterprise and likely to breakdown because their actions can be interpreted as not continuing to be engaged or to participating in the domain task. Some of these signals are also essential to conversation for the same reason. The signals include:

- talking about the task,
- turn taking,
- timing of uptake of a turn,
- use of gaze at the speaker, gaze away for taking turns[17],
- use of gaze at speaker to track speaker gestures with objects,
- use of gaze by speaker or non-speaker to check on attention of other,
- hand gestures for pointing, iconic description, beat gestures, (see [19], [7]), and in the hosting setting, gestures associated with domain objects,
- head gestures (nods, shakes, sideways turns)
- body stance (facing at other, turning away, standing up when previously sitting and sitting down),
- facial gestures (not explored in this work but see [20]),
- non-linguistic auditory responses (snorts, laughs),
- social relational activities (telling jokes, role playing, supportive rejoinders).

Several of these signals have been investigated by other researchers, and hence only a few are noteworthy here. The timing of uptake of a turn concerns the delay between the end of one speaker's utterances and the next speaker'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 at least of conversational difficulties. Uptake delay may only be a signal of disengagement when other cues also indicate disengagement: looking away, walking away, or body stance away from the other participant.

In hosting situations, among many other circumstances, 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 so that he can proceed with a demo. The visitors interpret these extended turns of attention to something as part of the domain 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 gestures related to operating equipment and object display when relevant to the domain task indicate that the collaboration is happening 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.

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. 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 transcript of one hosting session in which 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 highly coordinated and interactive and is not discussed before it occurs. Role playing begins at 00 in the figure and ends at 17. [The host P has shown the visitor C how restaurant customers order food in an imaginary restaurant using an actual electronic table, and is just finishing an explanation of how wait staff might use the new electronic table to assist customers.] Note that utterances by P and C are labeled with their letter and a colon, while other material describes their body actions.

52: P left hand under table, right hand working table, head and eyes to table, bent over

C watching P.

P: so that way they can have special privileges to make different things happen

C nods at "privileges" and at "happen"

54: P turns head/eyes to C, raises hands up

C's had down, eyes on table

55: P moves away from C and table, raises hands and shakes them; moves totally away full 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 point at each phrase and looks at table

P laughs; does 3 pointings while looking at table

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 and looks at C

C: yes

C: looks at P

17: P: and 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 looking at C and puts hands in lap

Figure 5 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, in addition to achieving some task domain goals, many researchers claim that participants are managing their social encounters, their "social face," or their trust [21,22] in each other. Social relational activities may occur in support these concerns. This claim seems quite likely to this author. However, one need not take a stand 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 activities observed in the videotaped sessions contribute to maintaining the connection between the participants. Social relational activities such as the role playing one in Figure 5 allow participants to demonstrate they are socially connected to one another in a strong way. They are more than just paying attention to one another, especially to accomplish their domain goals. They actively seek ways to indicate to the other that they have some 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 relational connection.

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 I am studying, my subjects seem very facile at accomplishing the integration of relational connection and the domain collaboration.

All collaborations have an end condition either because the participants give up on the goal (c.f. [23]), or because the collaboration succeeds in achieving the desired goals. When collaboration on a domain task ends, participants can elect to negotiate an additional collaboration or refrain from doing so. When they so refrain, they then undertake to close the engagement. Their means to do so is presumably as varied as the rituals to begin engagement, but I observe the common patterns of preclosing, expressing appreciation, saying goodbye, with an optional handshake, and then moving away from one another. Preclosings [24] 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 my 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. One might want to argue that if that is the case, then more complex machinery is needed than that so far suggested in conversational models of collaboration (cf. [2],[3],[25]). I believe this is not the case because much of the collaboration on engagement is non-verbal behavior that simply conveys that collaboration is happening. For much of the collaboration to be engaged, no complex recipes are needed. The portions of engagement that require complex recipes are those of beginning and ending the engagement. Once some domain collaboration begins, engagement is maintained by the engagement signals discussed above, and while these signals must be planned for by the individual participants and recognized by each counterpart, they do not require much computational mechanism to keep going. In particular, no separate stack is needed to compute the effects of engagement because the engagement itself is not discussed as such once a domain task collaboration begins.

How does one account for the social relational behaviors discussed above in this way? 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 can function with the mechanisms for that purpose. Intermixing relational connection 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 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 the ongoing engagement collaboration as well as the domain col-

laboration. 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 participants jointly. The host accepts this participation by simply continuing his part in it. It appears on the surface that they are just jointly participating in the hosting goal, but at the same time they are also participating jointly in a social interaction. Working out the details of how hosting agents and visitors accomplish this second collaboration remains to be done.

Presumably not all social behaviors cannot be interpreted in the context of the domain task. Sometimes participants interrupt their collaboration to tell a story that is either not pertinent to the collaboration or while pertinent, somehow out of order. These stories are interruptions of the current collaboration and are understood as having some other conversational purpose. As interruptions, they also 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 of uptake of the interruption that signals disengagement possibilities. Thus, failure to uptake the interruption is clearly one means to signal a start towards disengagement.

Open Questions

The discussion above raises a number of questions that must be addressed in my ongoing work. First, in my 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 SharedPlans collaboration model does not account for the kind of fine detail required to explain gaze changes, and nothing in the standard models of turn taking does either. 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 my data, participants do not always acknowledge or accept what another participant has said via linguistic expressions. Sometimes they use laughs 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. I am 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 as part of domain collaboration. 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 I pursue theory of human-human engagement, I am also interested in building new capabilities for Mel that are founded on human communication. To accomplish that, I will be combining hosting conversations with other research at MERL on face tracking and face recognition. These will make it possible to greet visitors in ways similar to human experience and may also allow us to 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 disinterest. Building a robot that can detect faces and track them and notice when the face disengages for a brief or extended period of time provides a piece of the interactive behavior.

Another challenge for a robot host is to experiment with techniques for dealing with unexpected speech input. People, it is said, say that darndest things. Over time I plan to be able to collect data for what people say to a robot host and use it to train speech recognition engines. However, at the beginning, and every time the robot's abilities improve dramatically, I do not have reliable data for conversational purposes. To operate in these conditions, I will make some rough predictions of what people say and then need to use techniques for behaving when the interpretation of the user's utterances falls below a threshold of reliability. Techniques I have used in spoken-language systems in onscreen application [16] are not appropriate for 3D agents because they cannot be effectively presented to the human visitor. Instead I expect to use techniques that (1) border on Eliza-like behavior, and (2) use the conversational models in Collagen [12] to recover when the agent is not sure what has been said.

8. SUMMARY

Hosting activities are a natural and common interaction among humans and one that can be accommodated by human-robot interaction. Making the human-machine experience natural requires attention to engagement activities in conversation. Engagement is a collaborative activity that is accomplished in part through gestural means. Previous experiments with a non-mobile robot that can converse and point provide a first level 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 Myroslava Dzikovska and Paul Dietz on Mel, Neal Lesh, Charles Rich, and Jeff Rickel on Collagen and PACO.

10. REFERENCES

1. W. Burgard and A. B. Cremes, "The Interactive Museum Tour Guide Robot," *Proceedings of AAAI-98*, 11-18, AAAI Press, Menlo Park, CA, 1998.
2. B.J. Grosz and C. L. Sidner. "Plans for discourse," in *Intentions and Plans in Communication and Discourse*. P. Cohen, J. Morgan, and M.Pollack (eds.), MIT Press, 1990.
3. B. J. Grosz and S. Kraus. "Collaborative Plans for Complex Group Action," *Artificial Intelligence*, 86(2): 269-357, 1996.
4. C. L. Sidner. "An Artificial Discourse Language for Collaborative Negotiation," in *Proceedings of the Twelfth National Conference on Artificial Intelligence*, MIT Press, Cambridge, MA, Vol.1: 814-819, 1994.
5. C. L. Sidner. "Negotiation in Collaborative Activity: A Discourse Analysis," *Knowledge-Based Systems*, Vol. 7, No. 4, 1994.
6. D. McNeill. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, Chicago, 1992.
7. W. L. Johnson, J. W. Rickel and J.C. Lester, "Animated Pedagogical Agents: Face-to-Face Interaction in Interactive Learning Environments," *International Journal of Artificial Intelligence in Education*, 11: 47-78, 2000.
8. J. Cassell, Y. I. Nakano, T. W. Bickmore ,C. L. Sidner and C. Rich. "Non-Verbal Cues for Discourse Structure," *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pp. 106-115, Toulouse, France, July 2001.
9. J. Cassell, J. Sullivan, S. Prevost and E. Churchill, *Embodied Conversational Agents*. MIT Press, Cambridge, MA, 2000.
10. J. Rickel, N. Lesh, C. Rich, C. L. Sidner and A. Gertner, "Collaborative Discourse Theory as a Foundation for Tutorial Dialogue," To appear in the *Proceedings of Intelligent Tutorial Systems 2002*, July 2002.
11. C. Rich, C. L. Sidner and N. Lesh, "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.
12. C. Rich and C. L. Sidner, "COLLAGEN: A Collaboration Manager for Software Interface Agents," *User Modeling and User-Adapted Interaction*, Vol. 8, No. 3/4, 1998, pp. 315-350.
13. J. Cassell, H. Vilhjálmsson, and T. W. Bickmore, "BEAT: the Behavior Expression Animation Toolkit" *Proceedings of SIGGRAPH 2001*,pp. 477-486, ACM Press, New York, 2001.
14. C. L. Sidner and M. Dzikovska, "Hosting Activities: Experience with and Future Directions for a Robot Agent Host," in *Proceedings of the 2002 Conference on Intelligent User Interfaces*, ACM Press, New York, pp. 143-150, 2002.
15. H.H. Luger, "Some Aspects of Ritual Communication," *Journal of Pragmatics.* Vol. 7: 695-711, 1983.
16. C. L. Sidner and C. Forlines, "Subset Languages For Conversing With Collaborative Interface Agents," submitted to the *2002 International Conference on Spoken Language Systems*.
17. S. Duncan, "Some Signals and Rules for Taking Speaking Turns in Conversation," in *Nonverbal Communication*, S. Weitz (ed.), Oxford University Press, New York, 1974.
18. J. Cassell, T. Bickmore, L. Campbell, H. Vilhjálmsson, and H. Yan, "Human Conversation as a System Framework: Designing Embodied Conversational Agents," in *Embodied Conversational Agents*, J. Cassell, J. Sullivan, S. Prevost, and E. Churchill (eds.), MIT Press, Cambridge, MA, 2000.

19. J. Cassell, , “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.), MIT Press, Cambridge, MA, 2000.
20. C. Pelachaud, N. Badler, and M. Steedman, “Generating facial expressions for speech,” *Cognitive Science*, 20(1):1-46, 1996.
21. Bickmore, T. and Cassell, J. "Relational Agents: A Model and Implementation of Building User Trust". *Proceedings of CHI-2001*, pp. 396-403, ACM Press, New York, 2001.
22. Katagiri, Y., Takahashi, T. and Takeuchi, Y. Social Persuasion in Human-Agent Interaction, *Second IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems, IJCAI-2001*, Seattle, pp. 64-69, August, 2001.
23. P.Cohen and H. Levesque, “Persistence, Commitment and Intention,” in *Intentions in Communication*, P. Cohen, J. Morgan and M.E. Pollack (eds.), MIT Press, Cambridge, MA, 1990.
24. E. Schegeloff and H. Sacks, “Opening up closings,” *Semiotica*, 7:4, pp 289-327, 1973.
25. K. E. Lochbaum, , “A Collaborative Planning Model of Intentional Structure,” *Computational Linguistics*, 24(4): 525-572, 1998.