# Engagement Rules for Human-Robot Collaborative Interactions*

Candace L. Sidner
Mitsubishi Electric Research Laboratories
201 Broadway
Cambridge, MA 02139
Sidner@merl.com

Christopher Lee
Mitsubishi Electric Research Laboratories
201 Broadway
Cambridge, MA 02139
Lee@merl.com

**Abstract** - *This paper reports on research on developing the ability for robots to engage with humans in a collaborative conversation for hosting activities. It defines the engagement process in collaborative conversation. The paper then presents the analysis of a study of human-human "look tracking" and discusses rules that will allow a robot to track and fail to track humans so that engagement is maintained.*

**Keywords:** Human-robot interaction, engagement, conversation, collaboration, collaborative interface agents, gestures in conversation.

## 1 Introduction

This paper reports on our research on developing the ability for robots to engage with humans in a collaborative conversation for hosting activities. Engagement is the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake. Engagement is supported by the use of conversation (that is, spoken linguistic behavior), ability to collaborate on a task (that is, collaborative behavior), and gestural behavior that conveys connection between the participants. While it might seem that conversational utterances alone are enough to convey connectedness (as is the case on the telephone), gestural behavior in face-to-face conversation conveys much about connection between the participants.

Collaborative conversations cover a vast range of interactions activities from call centers to auto repair to physician-patient dialogues. In order to narrow our research efforts, we have focused on hosting activities. Hosting activities are a class of collaborative activity in which an agent provides guidance in the form of information, entertainment, education or other services in the user's environment and may also request that the user undertake actions to support the fulfillment of those services. Hosting activities are situated or embedded activities, because they depend on the surrounding environment as well as the participants involved. 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 our research, physical robots, who serve as guides, are the hosts of the environment, which is a real-world, physical place.

Conversational gestures generally concern gaze at/away from the conversational partner, pointing behaviors, hand and facial gestures, (bodily) addressing the conversational participant and other persons/objects in the environment, all in appropriate synchronization with the conversational, collaborative behavior. These gestures are culturally determined, but every culture has some set of behaviors to accomplish the engagement task. These gestures in some cases play a dual role of getting sensory input (for the eyes and ears) and providing engagement cues to conversational participants. Here we focus on the engagement issues.

Not only must the robot produce these behaviors, but also it must interpret similar behaviors from its collaborative partner (hereafter CP). Proper gestures by the robot and correct interpretation of human gestures dramatically affect the success of conversation and collaboration. Inappropriate behaviors can cause humans and robots to misinterpret each other's intentions. For example, a robot might look away for an extended period of time from the human, a signal to the human that it wishes to disengage from the conversation and could thereby terminate the collaboration unnecessarily. Incorrect recognition of the human's behaviors can lead the robot to press on with a conversation in which the human no longer wants to participate.

While other researchers in robotics are exploring aspects of gesture (for example, [2], [12]), none of them have attempted to model human-robot interaction to the degree that involves the numerous aspects of engagement and

* 0-7803-7952-7/03/$17.00 © 2003 IEEE

collaborative conversation that we have set out above. Robotics researchers interested in collaboration and dialogue [9] have not based their work on extensive theoretical research on collaboration and conversation, as we will detail later. Our work is also not focused on emotive interactions, in contrast to Breazeal among others. For 2D conversational agents, researchers (notably, [6],[11]) have explored agents that produce gestures in conversation. However, they have not tried to incorporate recognition as well as production of these gestures, nor have they focused on the full range of these behaviors to accomplish the maintenance of engagement in conversation.

In this paper we discuss our research agenda for creating a robot with collaborative conversational abilities, including gestural capabilities in the area of hosting activities. We will also discuss the results of a study of human-human hosting and how we are using the results of that study to determine rules and associated algorithms for the engagement process in hosting activities. We will also critique our current engagement rules, and discuss how our study results might improve our robot's future behavior.

## 2 Communicative capabilities for collaborative robots

To create a robot that can converse, collaborate, and engage with a human interactor, a number of different communicative capabilities must be included in the robot's repertoire. Most of these capabilities are linguistic, but some make use of physical gestures as well. These capabilities are:

- Engagement behaviors: initiate, maintain or disengage in interaction;
- Conversation management: turn taking [8], interpreting the intentions of the conversational participants, establishing the relations between intentions and goals of the participants and relating utterances to the attentional state [10] of the conversation.
- Collaboration behavior: choosing what to say or do next in the conversation, to foster the shared collaborative goals of the human and robot, as well as how to interpret the human's contribution (either spoken acts or physical ones) to the conversation and the collaboration.

Turn taking gestures also serve to indicate engagement because the overall choice to take the turn is indicative of engagement, and because turn taking involves gaze/glance gestures. There are also gestures in the conversation (such as beat gestures, which are used to indicate old and new information [5]) that CPs produce and observe in their partners. These capabilities are significant to robotic participation in conversation because they allow the robot to communicate using the same strategies and techniques that are normal for humans, so that humans can quickly perceive the robot's communication.

Humans do not turn off their own conversational and engagement capabilities when talking to robots. Hence robots can and must make use of these capabilities to successfully communicate with humans. To use human linguistic and gestural behavior, a robot must fuse data gathered from its visual and auditory sensors to determine the human gestures, and infer what the human intends to convey with these gestures.

In our current work, we have developed (1) a model of engagement for human-human interaction and adapted it for human-robot interaction, and (2) an architecture for a collaborative non-mobile conversational robot.

Our engagement model describes an engagement process in three parts, (1) initiating a collaborative interaction with another, (2) maintaining the interaction through conversation, gestures, and, sometimes, physical activities, and (3) disengaging, either by abruptly ending the interaction or by more gradual activities upon completion of the goals of the interaction. The rules of engagement, which operate within this model, provide choices to a decision-making algorithm for our robot about what gestures and utterances to produce.

Our robot, which looks like a penguin, as shown in Figure 1, uses its head, wings and beak for gestures that help manage the conversation and also expresses engagement with its human interlocutor (3 DOF in head/beak, 2 in wings).

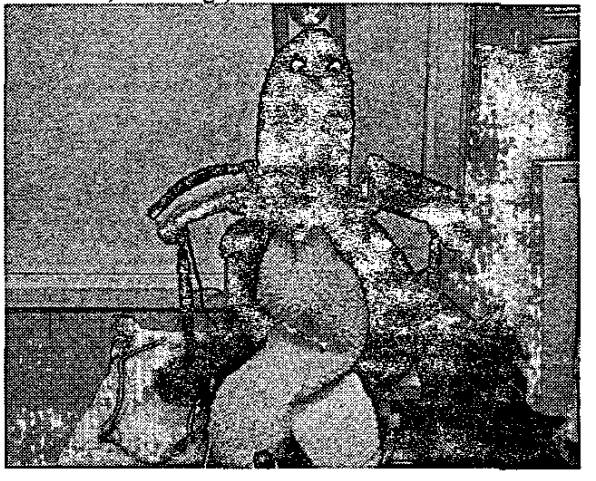

Figure 1: Mel, the penguin robot

While the robot can only converse with one person at a time, it gazes at the onlookers in the conversation in the conversation to acknowledge their presence. Gaze for our

robot is determined by the position of its head, since its eyes do not move. Since our robot cannot turn its whole body, it does not make use of rules we have already created concerning addressing with the body. Because bodily addressing (in US culture) is a strong signal for whom a CP considers the main other CP, change of body position is a significant engagement signal. However, we will be mobilizing our robot in the near future and expect to test these rules following that addition.

To create an architecture for collaborative interactions, we use several different systems, largely developed at MERL. The conversational and collaborative capabilities of our robot are provided by the Collagen ™ middleware for collaborative agents [16,17], and commercially available speech recognition software (IBM ViaVoice). We use a face detection algorithm [20], a sound location algorithm, and an object recognition algorithm [1] and fuse the sensory data before passing results to the Collagen system. The robot's motor control algorithms use the engagement rule decisions, which make use of information from the conversational state to decide what gestures to output. Further details about the architecture and current implementation can be found in [19].

# 3 Current engagement capabilities

The greatest challenge in our work on engagement is determining rules governing the maintenance of engagement. Our first set of rules, which we have tested in scenarios as described in [19], are a small and relatively simple set. The test scenarios do not involve pointing to or manipulating objects. Rather they are focused on engagement in more basic conversation. These rules direct the robot to initiate engagement with gaze and conversational greetings, to use gaze to indicate ongoing engagement when the human interlocutor is speaking and to use to gaze at both the human interlocutor and onlookers when the robot speaks, to use gaze away behaviors in turn taking, and to use failure of the interlocutor to take an expected turn together with loss of the face of the human as indication of the human's desire to end the interaction. In those cases where the human stays engaged until the robot has run out of things to say, the robot closes the conversation using known rules of conversational closing [14,18].

While this is a fairly small repertoire of engagement behaviors, it was sufficient to test the robot's behavior in a number of scenarios involving a single CP and robot, with and without onlookers. Much of the robot's behavior is quite natural. However, we have observed oddities in its gaze at the end of its turn (e.g., it will incorrectly look at an onlooker instead of its CP when ending its turn) as well as confusion about where to look when the CP leaves the scene. These behaviors are easy to correct. However, our rules are too limited for more complex interactions.

To explore these conversations, we have provided our robot with some additional gestural rules and new recipes for action (in the Collagen framework), so that our penguin robot now undertake a hosting conversation to demo an invention created at MERL. Mel greets a visitor (other onlookers can also be present), convinces the visitor to participate in a demo, and proceeds to show the visitor the invention. Mel points to demo objects (a kind of electronic cup sitting on a table), talks (with speech) the visitor through the use of the cup, asks the visitor questions, and interprets the spoken answers, and turns to onlookers during its turn. The robot also expects the visitor to say and do certain activities, as well as gaze at objects, and will await or insist on such gestures if they are not performed. The entire interaction lasts about five minutes.

Not all of Mel's behaviors in this interaction appear acceptable to us. For example, Mel often looks away for too long (at the cup and table) when explaining them, it (Mel is "it" since it is not human) fails to make sure it is looking at the visitor when it calls the visitor by name, and it sometimes fails to look for a long enough when it turns to look at objects. To make Mel perform more effectively, we are investigating gesture in human-human interactions.

# 4 Evidence for engagement in human behavior

Our first efforts in building a collaborative, engaged robot for hosting involved conversations of the how-are-you-and-welcome-to-our-lab format. However, our goal is hosting conversations, which involve much more activity. While other researchers have made considerable progress on the navigation involved for a robot to host visitors [e.g. 4] and gestures needed to begin conversation [e.g. 3], many aspects of conversation are open for investigation. These involve extended explanations, pointing at objects, manipulating them (on the part of the humans), moving around in a physical environment to access objects and interact with them. This extended repertoire of tasks requires many more gestures than our current set. In addition, some of these gestures could alternatively be taken to indicate that the human participant is disengaging from the conversation (e.g. looking away from the human speaker). So many behaviors appear to be contextually sensitive.

Our task, then, is to identify and interpret the range of gestures that humans in US culture typically produce and respond to in hosting situations. Much of the available literature (e.g. [8], [13]) provides a base for determining what gestures to consider, but does not provide enough detail about how gestures are used to maintain conversational engagement, that is, to signal that the participants are interested what the other has to say and in keeping the interaction going.

The significance of gestures for human-robot interaction can be understood by considering the choices that the robot has at every point in the conversation for its head movement, its gaze, and its use of pointing. The robot must also determine whether the CP has changed its head position or gaze and what objects the CP points to or manipulates. Head position and gaze are indicators of engagement. Looking at the speaking CP is evidence of engagement, while looking around that room, for more than very brief moments, is evidence of disinterest in the interaction and possibly the desire to disengage. However, looking at objects relevant to the conversation is not evidence of disengagement. Furthermore, the robot needs to know that the visitor has or has not paid attention to what it points at or looks at. If visitor fails to look at what the robot looks at, the visitor might miss something crucial to the interaction.

A simple hypothesis for maintaining engagement (for each CP) is: Do what the speaking CP does: look wherever the CP looks; look at him if he looks at you, and look at whatever objects are relevant to the discussion when he does. When the robot is the speaking CP, this hypothesis means it will expect perfect tracking of its looking by the human interlocutor. The hypothesis leaves does not constrain the robot's decisions for what to look and point at when the robot is the speaking CP.

Note that there is evidence that the type of utterances that occur in conversation affect the gaze of the non-speaking CP. [15] provides evidence that in direction giving tasks, the non-speaking CP will gaze more often at the speaking CP when utterance pairs are assertion followed by elaboration and more often at a map when the utterance pairs are assertion followed by the next map direction.

To evaluate the simple hypothesis for engagement, we have been evaluating interactions in videotapes of human-human hosting activities, which were recorded in our laboratory. In these interactions, a member of our lab hosted a visitor who was shown various new inventions and computer software systems. The host and visitor were followed by video camera as they experienced a typical tour of our lab and its demos. The host and visitor were not given instructions to do anything except to give/take the lab tour. We have obtained about 3.5 hours of video, with three visitors, who are, on separate occasions, each given a tour by the same host.

Our method of evaluating this data reflects our experiences in the computational linguistic, psychological and ethno-methodological communities. There are other approaches that might be brought to bear to provide gestural knowledge for our robot. For example, [2] uses built-in learning algorithms to learn in real-time from features of perceptual input how to respond to a human partner in an empathetic way. However, this method requires many hours of human training, and for conversational interactions, it is unclear just which features must be best trained over. Alternatively, one could train over the videotapes themselves, but vision algorithms are not yet sophisticated enough to perceive the subtleness of gesture in human interaction. Hence we have been transcribing portions of the video for all the utterances made and the gestures (head, hands, body position, body addressing) that occur during portions of the video. We have not transcribed facial gestures. We report here on our results in observing gestural data (and its corresponding linguistic data) for just over five minutes of one of the host-visitor pairs.

The subject of the investigated portion of the video is a demonstration of an "Iglassware" cup which P (a male) demos and explains to C (a female). P produces a great many types of gestures, with his hands, face, and body. P gazes at C and changes gaze to other objects. P sometimes also points to the cup and holds it. He also interacts with a table to which the cup transfers data. He also uses his hands to produce iconic and metaphorical gestures [5], and he uses the cup as a presentation device as well. We do not discuss iconic, metaphorical or presentation gestures, in large part because our robot does not have hands with which to perform similar actions.

We report here on C's tracking of P's gaze gestures (since P speaks the overwhelming majority of the utterances in their interaction). Gaze in this analysis is expressed in terms of head movements (looking). We did not code eye movements due to video quality.

There are 81 occasions on which P changes his gaze by moving his head with respect to C. Seven additional gaze changes occurred that are not counted in this analysis because it is unclear to where P changed his gaze. Of the counted gaze changes, C tracks 45 of them (55%). The remaining failures to track gaze (26, or 45% of all gazes) can be subclassed into 3 groups: quick looks, nods (glances followed by gestural or verbal feedback), and uncategorized failures (see Table 1).

Table 1: Failures to Track Gaze Change

| | Count | %of tracking failures | %of total tracking |
|---|---|---|---|
| Quick looks | 11 | 31 | 14 |
| Nods | 13 | 36 | 16 |
| Uncategorized | 12 | 33 | 15 |

These tracking failures indicate that our simple hypothesis for maintaining engagement is incorrect. Of these tracking failures, the *nod failures* can be explained because they occur when P looks at C even though C is

looking at something else (usually the cup or the table, depending on which is relevant to the discussion). However, in all these cases, C immediately nods and often articulates with "Mm-hm," "Wow" or other phrases to indicate that she is following her conversational partner. In grounding terms [7], P is attempting to ascertain by looking at C that she is following his utterances and actions. When C cannot look, she provides feedback by nods and comments. She grounds his comments and thereby indicates that she is still engaged. In the nod cases, she also is not just looking around the room, but paying attention to an object that is highly relevant to the demo.

The quick looks and uncategorized failures represent 64% of the failures and 28% of the gaze changes in total. Closer study of these cases reveals significant information for our robot vision detection algorithms.

In the *quick look* cases, P looks quickly (including moving his head) at something without C tracking his behavior. Is there a reason to think that C is not paying attention or has lost interest? Does P not care that C is not following his actions? He never stops to check for her feedback in theses cases. Has C lost interest or is she merely not required to follow? In all of these cases, C does not track P for the simple reason that P's gazes away (usually to the cup or the table, but occasionally to something else he is commenting on) happen very quickly. It's a quick up-and-down or down-and-up head movement. There simply is insufficient time for C to do so. In every case C is paying attention to P, so C is able to see what's happening. It appears she judges on the velocity of the movement not to track it.

Of the *uncategorized failures*, the majority (8 instances) occur when C has other actions or goals to undertake. For example, in three instances C is finishing a nod and not be able to track P while she's nodding; 3 instances are goal conflicts where C must at something on the table rather than P; two instances are goal conflicts for personal actions (drinking water C has brought with her). One uncategorized failure is similar to a quick look but longer in duration.

Of the remaining 3 tracking failures, we note that each one lasts between 1 to 2 seconds and hence are not quick looks. Each occurs for seemingly good reasons to us as observers, but may not have been known to the participants at the time of occurrence. For example, one failure occurs at the start of the demo when C is looking at the new (to her) object that P displays and does not track P when he looks up at her. The second failure is similar to the nod failures, except there is a longer delay before the feedback utterance. P has been looking at C and over a two second time period, and he looks down at the cup as he says "and there's alike a sixteen bit id number;" he also points in a circling motion at the object he holds. So C misses his face looking down and also the pointing that he does. However, she does convey that she's paying attention, because at the conclusion of his intonational phrase, she nods and says "Mm-hm." She has not lost interest, and still conveys understanding of his linguistic utterance. P does not attempt to re-do his action because the pointing and circling is not critical to his statement, and C has indicated that the statement is understood. We can conclude that interlocutors use a variety of means of keeping connected, and sometimes forego, for reasons we cannot discern, complete tracking. They do however convey their interest by other means.

The third tracking failure, unlike the second, does not result in C missing gestural information. P has been looking at C, and looks down at the cup as he says "And um, I'm going to make the argument, that you know there's" after which he looks at C. C does not track his look down at the cup. She is clearly looking at him, so if he can see her peripherally, she is conveying her interest in him. Because P does not make any motions with the cup (he just holds it presented to C), C is not missing any gesturally conveyed information with the cup (which has been under discussion during the whole conversation). Again, it would seem that not completely tracking the other's movements can occur, as long as some means of conveying connection is undertaken.

This data clearly indicates that our rules for maintaining engagement must be more complex than the "do whatever the speaking CP does" hypothesis. In fact, tracking is critical to the interaction. It is not completely essential because very quick glances away can be disregarded, and one can avoid tracking the interlocutor by providing verbal feedback. In those cases where arguably one should track the speaking CP, failure to do so may lead the CP to restate an utterance or to just go on because the lost information is not critical to the interaction. The data in fact suggest that for our robot engagement rules, tracking the speaking CP in general is the best move, but when another goal interferes, providing verbal feedback maintains engagement. Furthermore, the robot as tracker can ignore head movements of brief duration, as these quick looks do not need tracking.

One final observation from our data set can be offered here that is pertinent to the decision rules for where the robot looks during its turn. Before either P or C pointed to an object, they always looked in that direction first. This look makes sense because it is required for a directionally accurate pointing gesture. Similar behavior is natural and obvious for our robot as well.

## 5 Future Directions

An expanded set of rules for engagement requires a means to evaluate them. We want to use the standard training and testing set paradigm common in computational linguistics and speech processing, but training and test sets are hard to come by for interaction with robots. Our solution has been to approach the process with a mix of techniques. We have developed a graphical display of simulated, animated robots running our engagement rules. We plan to observe the interactions in the graphic display for a number of scenarios to check and tune our engagement rules. Following that, we will undertake something akin to a testing set by testing the robot's interactions with human visitors as the robot demonstrates a piece of hardware developed in our lab.

## 6 Summary

This paper has discussed the nature of engagement in human-robot interaction, and outlined our methods for investigating rules for engagement for the robot. We report on analysis of human-human look tracking where the humans do not always track the changes in looks by their conversational interlocutors. We conclude that such tracking failures indicate both the default behavior for a robot and when it can fail to track without its human conversational partner inferring that it wishes to disengage from the interaction.

## References


1. P.A. Beardsley, "Piecode Detection," Mitsubishi Electric Research Labs TR2003-11, Cambridge, MA, February, 2003.
2. C. Breazeal, "Affective interaction between humans and robots", Proceedings of the 2001 European Conference on Artificial Life (ECAL2001). Prague, Czech Republic, (2001).
3. A. Bruce, I. Nourbakhsh, R. Simmons. "The Role of Expressiveness and Attention in Human Robot Interaction," In Proceedings of the IEEE International Conference on Robotics and Automation, Washington DC, May 2002.
4. W. Burgard, A.B. Cremes, D. Fox, D. Haehnel, G. Lakemeyer, D. Schulz, W. Steiner, & S. Thrun, "The Interactive Museum Tour Guide Robot," Proceedings of AAAI-98, 11-18, AAAI Press, Menlo Park, CA, 1998.
5. 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.), Cambridge, MA: MIT Press, 2000.
6. J. Cassell, J. Sullivan, S. Prevost and E. Churchill, *Embodied Conversational Agents*. MIT Press, Cambridge, MA, 2000.
7. H.H. Clark, *Using Language*, Cambridge University Press, Cambridge, 1996.
8. S. Duncan. "Some signals and rules for taking speaking turns in conversation," in *Nonverbal Communication*, S. Weitz (ed.), New York: Oxford University Press, 1974.
9. T. Fong, C. Thorpe, and C. Baur, "Collaboration, Dialogue and Human-Robot Interaction,"10th International Symposium of Robotics Research, Lorne, Victoria, Australia, November, 2001.
10. B.J. Grosz and C.L. Sidner, "Attention, intentions and the structure of discourse," *Computational Linguistics*, 12(3): 175-204, 1986.
11. W.L. Johnson, J.W. Rickel and J.C. Lester, "Animated Pedagogial Agents: Face-to-Face Interaction in Interactive Learning Evnironments," *International Journal of Artificial Intelligence in Education*, 11:47-78, 2000.
12. T. Kanda, H. Ishiguro, M. Imai, T. Ono, and K. Mase. "A constructive approach for developing interactive humanoid robots. Proceedings of IROS 2002, IEEE Press, NY, 2002.
13. A. Kendon. "Some functions of gaze direction in social interaction," *Acta Psychologica*, 26: 22-63, 1967.
14. H.H. Luger. "Some Aspects of Ritual Communication," *Journal of Pragmatics*. Vol. 7: 695-711, 1983.
15. Y. Nikano, G. Reinstein, T. Stocky, J. Cassell. "Towards a Model of Face-to-Face Grounding," in submission, 2003.
16. 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, 1998.
17. 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.
18. E. Schegeloff, & H. Sacks. "Opening up closings," *Semiotica*, 7(4): 289-327, 1973.
19. C.L. Sidner and C. Lee, "An Architecture for Engagement in Collaborative Conversations between a Robot and Humans," Mitsubishi Electric Research Labs TR2003-13, Cambridge, MA, April, 2003.
20. C. Viola, P. and Jones, M. "Rapid Object Detection Using a Boosted Cascade of Simple Features," IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, pp. 905-910, 2001.