Recognizing Engagement in Human-Robot Interaction

Charles Rich, Brett Ponsler, Aaron Holroyd and Candace L. Sidner
Computer Science Department
Worcester Polytechnic Institute
Worcester, MA 01609
(rich | bponsler | aholroyd | sidner)@wpi.edu

***Abstract*—Based on a study of the engagement process between humans, we have developed and implemented an initial computational model for recognizing engagement between a human and a humanoid robot. Our model contains recognizers for four types of connection events involving gesture and speech: directed gaze, mutual facial gaze, conversational adjacency pairs and backchannels. To facilitate integrating and experimenting with our model in a broad range of robot architectures, we have packaged it as a node in the open-source Robot Operating System (ROS) framework. We have conducted a preliminary validation of our computational model and implementation in a simple human-robot pointing game.**

***Keywords*—dialogue, conversation, nonverbal communication**

I. Introduction

Engagement is "the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake" [1] To elaborate,

> ...when people talk, they maintain conscientious psychological connection with each other and each will not let the other person go. When one is finished speaking, there is an acceptable pause and then the other *must* return something. We have this set of unspoken rules that we all know unconsciously but we all use in every interaction. If there is an unacceptable pause, an unacceptable gaze into space, an unacceptable gesture, the cooperating person will change strategy and try to re-establish contact. Machines do none of the above, and it will be a whole research area when people get around to working on it. (Biermann, invited talk at User Modeling Conference, 1999)

In the remainder of this paper, we first review the results of a video study of the engagement process between two humans. Based on this and prior studies, we have codified four types of events involving gesture and speech that contribute to the perceived connection between humans: directed gaze, mutual facial gaze, conversational adjacency pairs and backchannels.

Next we analyze the relationship between engagement recognition and other processes in a typical robot architecture, such as vision, planning and control, with the goal of designing a reusable human-robot engagement recognition module to coordinate and monitor the engagement process. We then describe our implementation of a Robot Operating System (ROS, see ros.org) node based on this design and its validation in a simple human-robot pointing game.

A. Motivation

We believe that engagement is a fundamental process that underlies all human interaction and has common features across a very wide range of interaction circumstances. At least for humanoid robots, this implies that modeling engagement is crucial for constructing robots that can interact effectively with humans without special training.

This argument motivates the main goal of our research, which is to develop an engagement module that can be *reused* across different robots and applications. There is no reason that every project should need to reimplement the engagement process. Along with the creators of ROS and others, we share the vision of increasing code reuse in the robotics research and development community.

Closer to home, we recently experienced first-hand the difference between simply implementing engagement behaviors in a human-robot interaction and having a reusable implementation. The robot's externally observable behavior in the first version of the pointing game [2] is virtually indistinguishable from our current demonstration (see Fig. 13). However, internally the first version was implemented as one big state machine in which the pointing game logic, engagement behaviors and even some specifics of our robot configuration were all mixed together. In order to make further research progress, however, we needed to pull out a reusable engagement recognition component, which caused us, among other things, to go back and more carefully analyze our video data. This paper is in essence a report of that work.

B. Related Work

In the area of human studies, Argyle and Cook [3] documented that failure to attend to another person via gaze is evidence of lack of interest and attention. Other researchers have offered evidence of the role of gaze in coordinating talk between speakers and listeners, in particular, how gestures direct gaze to the face and why gestures might direct gaze away from the face [4], [5], [6]. Nakano *et al.* [7]) reported on the use of the listener's gaze and the lack of negative feedback to determine whether the listener has grounded [8] the speaker's turn. We rely upon the background of all of this work in the analysis of our own empirical studies.

In terms of computational applications, the most closely related work is that of Peters [9], which involves agents in virtual environments, and Bohus and Horvitz [10], [11], which involves a realistically rendered avatar head on a desktop display. We share a similar theoretical framework with both

Fig. 1. Two camera views of participants in human engagement study (during directed gaze event).

of these efforts, but differ in dealing with a humanoid robot and in our focus on building a reusable engagement module.

Mutlu *et al.* [12] have studied the interaction of gaze and turn-taking [15] using a humanoid robot. Flippo *et al.* [13] have developed a similar architecture (see Section III) with similar concerns of modularity and the fusion of verbal and nonverbal behaviors, but for multimodal interfaces rather than robots. Neither of these efforts, however use the concepts of engagement or connection events.

II. Human Engagement Study

Holroyd [2] conducted a study of human engagement behavior in which pairs of humans sat across an L-shaped table from each other and prepared canapés together (see Fig. 1). Each of four sessions involved an experimenter (confederate) and two study participants and lasted about 15–20 minutes. In the first half of each session, the experimenter instructed the study participant in how to make several different kinds of canapés using combinations of the different kinds of crackers, spreads and toppings arrayed on the table. The experimenter then left the room and was replaced by a second study participant, who was then taught to make canapés by the first participant.[1] The eight study participants consisted of six males and two females, all college students at Worcester Polytechnic Institute (WPI). All sessions were videotaped using two cameras.

In our current analysis of the videotapes, we only looked at the engagement maintenance process. We did not analyze the participants' behaviors for initiating engagement (meeting, greeting, sitting down, etc.) or terminating engagement (ending the conversation, getting up from the table, leaving the room, etc.) These portions of the videotapes will be fruitful for future study.

During the periods of maintained engagement, we coded where each person was looking at each moment (i.e., at the other person's face, at a specific object or group of objects on the table, or "away"), when they pointed at a specific object or objects on the table, and the beginning and end of each person's speaking turn. Based on this analysis and the literature on engagement cited above, we have identified four types of what we call *connection events*, namely directed gaze, mutual facial gaze and adjacency pairs and backchannels. Our hypothesis is that these events, occuring at some minimum frequency, are the process mechanism for maintaining engagement.

[1]The second half of one of the sessions is missing due to camera failure.

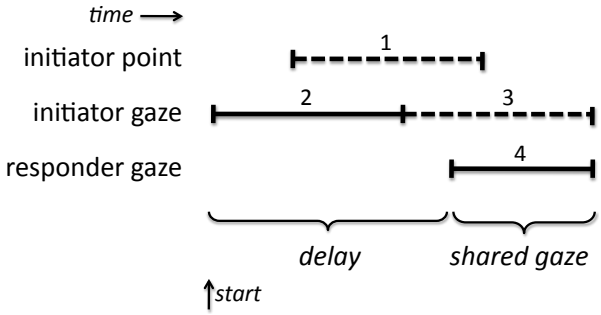

Fig. 2. Time line for directed gaze (numbers for reference in text).

A. Connection Event Types

Figures 2 through 5 shows the time lines for the four types of connection events we have analyzed and TABLE I shows some summary statistics. In our discussion of each event type below, we will both describe the objectively observeable behavior components of the event type and hypothesize regarding the accompanying intentions of the participants. It is also important to note that the two gestural event types, directed gaze and mutual facial gaze, can and often do overlap with adjacency pairs, which involve speech, and backchannels are by definition overlapping communications. Dotted lines indicate optional behavior.

1) Directed Gaze: In directed gaze [4], one person (the *initiator*) looks and optionally points at some object or group of objects in the immediate environment, following which the other person (the *responder*) looks at the same object(s). We hypothesize that the initiator intends to bring the indicated object(s) to the responder's attention, i.e., to make the object(s) more salient in the interaction. This event is often synchronized with the initiator referring to the object(s) in speech, as in "now spread the *cream cheese* on the cracker." By turning his gaze where directed, the responder intends to

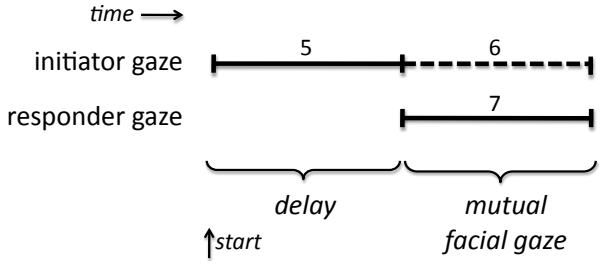

Fig. 3. Time line for mutual facial gaze (numbers for reference in text).

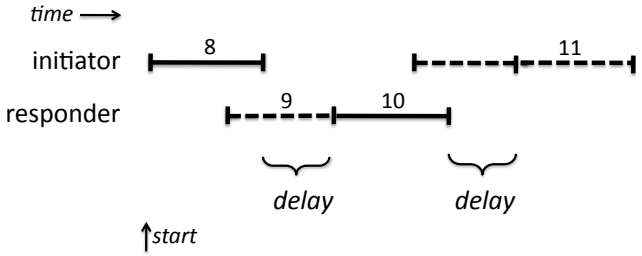

Fig. 4. Time line for adjacency pair (numbers for reference in text).

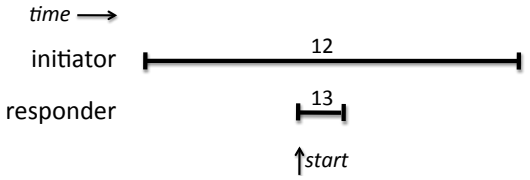

Fig. 5. Time line for backchannel (numbers for reference in text).

be cooperative and thereby signals his desire to continue the interaction (maintain engagement).

In more detail (see Fig. 2), notice first that the pointing behavior (1), if it is present, begins after the initiator starts to look (2) at the indicated object(s). This is likely because it is hard to accurately point at something without looking to see where it is located.[2] Furthermore, we observed several different configurations of the hand in pointing, such as extended first finger, open hand (palm up or palm down—see Fig. 1), and a circular waving motion (typically over a group of objects). An interesting topic for future study (that will contribute to robot generation of these behaviors) is to determine which of these configurations are individual differences and which serve different communicative functions.

After some delay, the responder looks at the indicated object(s) (4). The initiator usually maintains the pointing (1), if it is present, at least until the responder starts looking at the indicated object(s). However, the initiator may stop looking at the indicated object(s) (2) before the responder starts looking (4), especially when there is pointing. This is often because the initiator looks at the responder's face, assumedly to check whether the responder has directed his gaze yet. (Such a moment is captured in Fig. 1.)

Finally, there may be a period of shared gaze, i.e., a period when both the initiator (3) and responder (4) are looking at the same object(s). Shared gaze has been documented [14] as an important component of human interaction.

2) Mutual Facial Gaze: Mutual facial gaze [3] has a time line (see Fig. 3) similar to directed gaze, but simpler, since it does not involve pointing. The event starts when the initiator looks at the responder's face (5). After a delay, the responder looks at the initiator's face, which starts the period of mutual facial gaze (6,7). Notice that the delay can be zero, which occurs when both parties simultaneously look at each other.

The intentions underlying mutual facial gaze are less clear than those for directed gaze. We hypothesize that both the initiator and responder in mutual facial gaze engage in this behavior because they intend to maintain the engagement process. Mutual facial gaze does however have other interaction functions. For example, it is typical to establish mutual facial gaze at the end of a speaking turn.

[2]Given the extreme flexibility of human behavior and the complexity of the world, it is usually possible to creatively imagine an exception to almost any rule such as this. For example, suppose a person is standing with his back to a mountain range. One could imagine him quite naturally pointing over his shoulder to "the mountains" without turning around to look at them. We will not bother continuing to point out the possibility of such exceptions below.

Finally, what we are calling mutual facial gaze is often referred to informally as "making eye contact." This latter term is a bit misleading since people do not normally stare continously into each other's eyes, but rather their gaze roams around the other person's face, coming back to the eyes from time to time.

3) Adjacency Pair: In linguistics, an adjacency pair [15] consists of two utterances by two speakers, with minimal overlap or gap between them, such that the first utterance provokes the second utterance. A question-answer pair is a classic example of an adjacency pair. We generalize this concept slightly to include both verbal (utterances) and non-verbal communication acts. So for example, a nod could be the answer to a question, instead of a spoken "yes." Adjacency pairs, of course, often overlap with the gestural connection events, directed gaze and mutual facial gaze.

The simple time line for an adjacency pair is shown in Fig. 4. First the initiator communicates what is called the *first turn* (8). Then there is a delay, which could be zero if the responder starts talking before the the initiator finishes (9). Then the responder communicates what is called the *second turn* (9,10). In some conversational circumstances, this could also be followed by a *third turn* (11) in which the initiator, for example, repairs the responder's misunderstanding of his original communication.

4) Backchannel: A backchannel [15] is an event (see Fig. 5) in which one party (the responder) directs a brief verbal or gestural communication (13) back to the initiator *during* the primary communication (12) from the initiator to the responder. Typical examples of backchannels are nods and/or saying "uh, huh." Backchannels are typically used to communicate the responder's comprehension of the initiator's communication (or lack thereof, e.g., a quizzical facial expression) and/or desire for the initiator to continue. Unlike the other three connection event types, the start of a backchannel event is defined as the start of the responder's behavior and this event has no concept of delay.

TABLE I
SUMMARY STATISTICS FOR HUMAN ENGAGEMENT STUDY

		count	delay (sec)		
			min	mean	max
directed gaze	succeed	13	0	0.3	2.0
	fail	1	1.5	1.5	1.5
mutual facial gaze	succeed	11	0	0.7	1.5
	fail	13	0.3	0.6	1.8
adjacency pair	succeed	30	0	0.4	1.1
	fail	14	0.1	1.2	7.4
backchannel		15	n/a	n/a	n/a
mean time between connection events (MTBCE) = 5.7 sec max time between connection events = 70 sec					

B. Summary Statistics

Summary statistics from a detailed quantitive analysis of approximately nine minutes of engagement maintenance time are shown in TABLE I. The time between connection events is defined as the time between the *start* of successive events, which properly models overlapping events. We hypothesize that the mean time between connection events (MTBCE) captures something of what is informally called the "pace" of an interaction [16]:

$$\text{pace} \quad \propto \quad \frac{1}{\text{MTBCE}}$$

In other words, the faster the pace, the less the time between connection events. Furthermore, our initial implementation of an engagement recognition module (see Section IV) calculates the MTBCE on a sliding window and considers an increase as evidence for the weakening of engagement.

Two surprising observations in TABLE I are the relatively large proportion of failed mutual facial gaze (13/24) and adjacency pair (15/45) events and the 70 second maximum time between connection events. Since we do not believe that engagement was seriously breaking down anywhere during the middle of our sessions, we take these observations as an indication of missing factors in our model of engagement. In fact, reviewing the specific time intervals involved, what we found was that in each case the (non-)responder was busy with a detailed task on the table in front of him.

III. HUMAN-ROBOT ARCHITECTURE

In general in software development, the key to making a reusable component is careful attention to the setting in which it will be used and the "division of labor" between the component and the rest of the computational environment in which it is embedded.

A. Human-Robot Setting

Fig. 6 shows the setting of our current architecture and implementation, which mirrors the setting of the human engagement study, namely a human and a humanoid robot with a table of objects between them. Either the robot or the human can be the initiator (or responder) in the connection event time lines shown in the previous section.

Like the engagement maintenance part of the human study, mobility is not part of this setting. Unlike the human study, we are not dealing here with manipulation of the objects

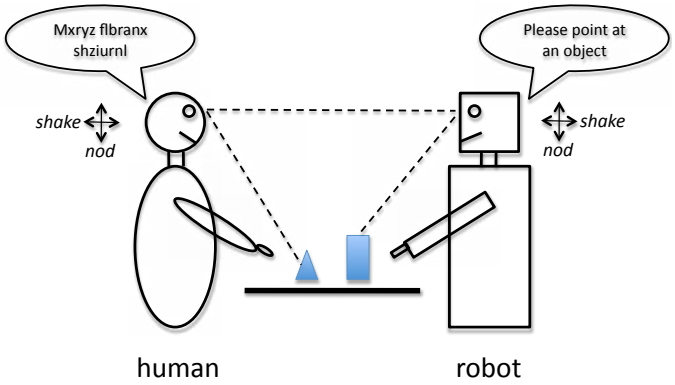

Fig. 6. Setting of human-robot interaction.

or changes in stance (e.g., turning the body to point to or manipulate objects on the side part of the L-shaped table).

Both the human and the robot can perform the following behaviors and observe them in the other:

- look at the other's face, objects on the table or "away"
- point at objects on the table
- nod the head (up and down)
- shake the head (side to side)

The robot can generate speech that is understood by the human. However, since our demonstration system (see Section IV) does not include natural languge understanding, the robot can only detect the beginning and end of the human's speech.

B. Information Flow

Fig. 7 shows the information flow between the engagement recognition module and rest of the software that operates the robot. In ROS, this information flow is implemented via message passing, as described in the next section. The next section also specifies the state machine for recognizing each connection event type.

Notice first in Fig. 7 that the rest of the robot architecture, not including the engagement recognition module, is shown as a big cloud. This vagueness is intentional in order to maximize the reusability of the engagement module. This cloud typically contains sensor processing, such as computer vision and speech recognition, cognition, including planning and natural language understanding, and actuators that control the robot's arms, head, eyes, etc. However, the exact organization of these components does not matter to the engagement module. Instead we focus on the solid arrows in the diagram, which specify what information the rest of the robot architecture must supply to the engagement module.

Starting with arrow (1), the engagement module needs to receive information about where the human is looking and pointing in order to recognize human-initiated directed gaze and mutual facial gaze events. It also needs to be notified of the human's head nods and shakes in order to recognize human backchannel events and human gestural turns in adjacency pair events.

The engagement module also needs to be notified (2) of where the robot is looking (in order to recognize the completion of a human-initiated directed gaze or mutual facial

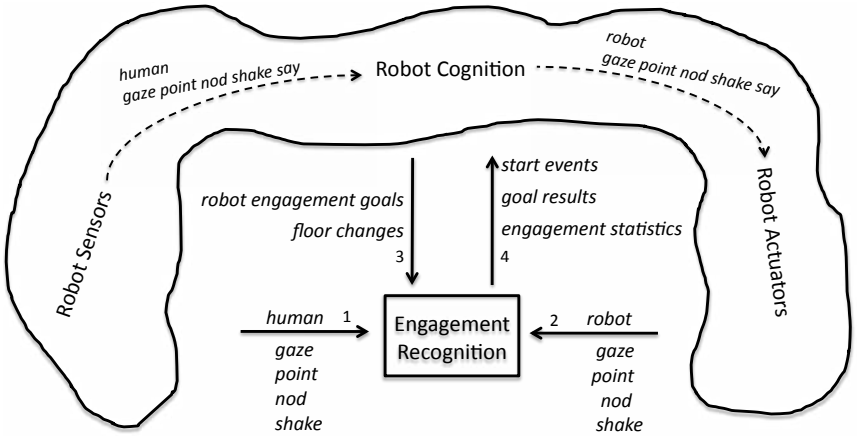

Fig. 7. Information flow between engagement recognition and the rest of robot architecture (numbers for reference in text).

gaze), pointing and when the robot nods or shakes. This may seem a bit counterintuitive at first. For example, would not the engagement module be more useful if it took responsibility for making the robot automatically look where the human directs it to look? The problem with this potential modularity is that the decision of where to look can depend on a deep understanding of the current task context. You may sometimes ignore an attempt to direct your gaze—suppose you are in the midst of a very delicate manipulation on the table in front of you when your partner points and says "look over here." Such decisions need to be made in the cognitive components of the robot. Similarly, only the cognitive components can decide when the robot should point and whether it should backchannel comprehension (nod) or the lack thereof (shake).

Robot engagement goals (3) trigger the engagement recognition module to start waiting for the human response in all robot-initiated event types, except backchannel (which does not have a delay structure). For example, suppose the (cognitive component of the) robot decides to direct the human's gaze to a particular object. After appropriately controlling the robot's gaze and point, a directed-gaze engagement goal is then sent to the engagement component.

The *floor* in a conversational interaction simply refers to who is the (primary) person currently speaking (communicating). Floor change information (3) is needed to support recognition of adjacency pair events. In natural spoken conversation, people signal that they are done with their turn via a combination of intonation (dropping tone), gesture (mutual facial gaze) and utterance semantics (e.g., a question). The engagement module thus relies on the rest of the robot architecture, such as speech recognition and natural language understanding, to decide when the human is beginning and ending his/her turn. Similarly, only the cognitive component of the robot can decide when/whether to take and/or give up the robot's turn.

Arrow (4) summarizes the information that the engagement recognition module provides to the rest of the robot architecture to coordinate and monitor the engagement process. First, the module provides notification of the start of human-initiated connection events, so that the robot can respond. The

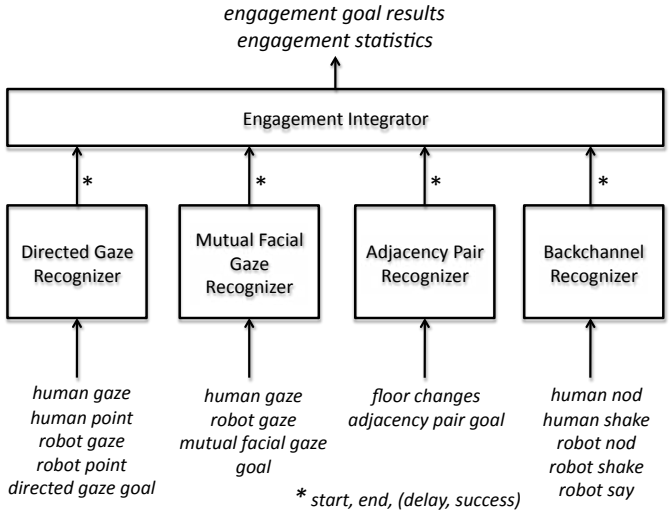

Fig. 8. Internal architecture of engagement recognition module.

module also provides real-time feedback on the successful or unsuccessful completion of robot-initiated connection events (engagement goals). For example, if the robot directs the user's gaze to an object and the user does not look, the engagement module notifies the rest of the architecture, so that the robot can try again, if necessary. Finally, the engagement module provides various ongoing statistics, similar to those in TABLE I, which the robot can use to gauge the health of the engagement process and decide, for example, to initiate more connection events.

C. Engagement Recognition Module

Fig. 8 shows the internal architecture of the engagement recognition module, which consists of four parallel recognizers that feed information to an integrator process. More than one recognizer may be active at one time (i.e., overlapping connection events), but only one event of each type may be in progress at any time. As shown in the figure, each recognizer responds to a subset of the information coming into the recognition module. The state machine for each recognizer is shown in the next section.

Each recognizer reports its start time, end time and, except for backchannel, its delay duration and whether it successfully

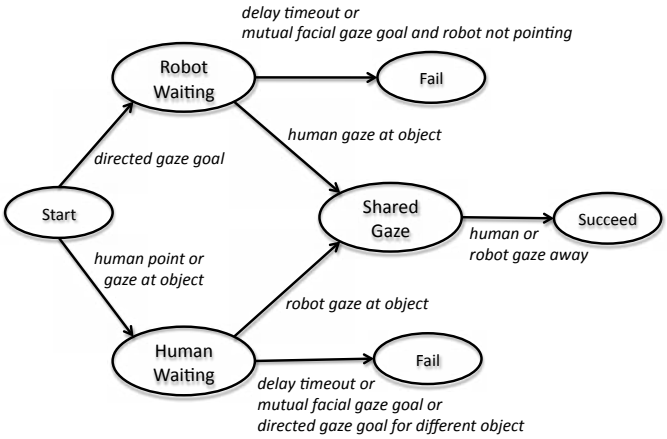

Fig. 9. Recognizer for directed gaze (see Fig. 2).

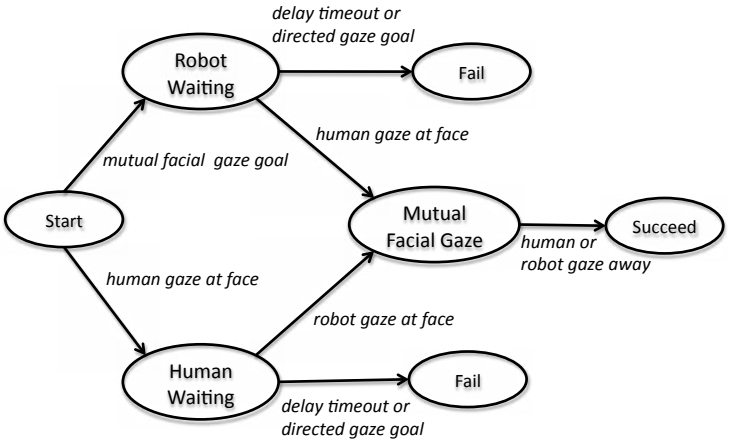

Fig. 10. Recognizer for mutual facial gaze (see Fig. 3).

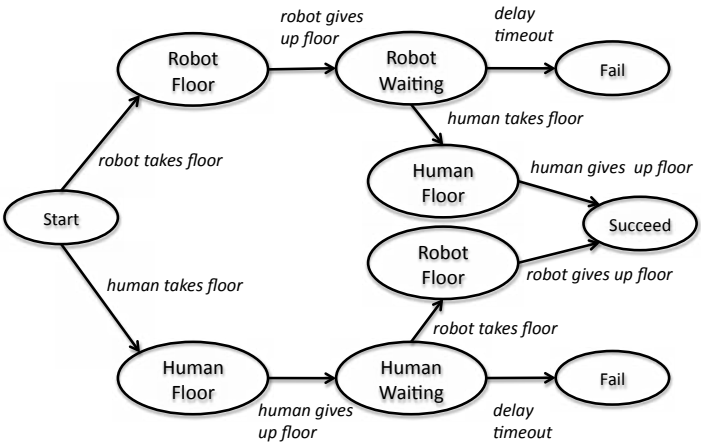

Fig. 11. Recognizer for adjacency pair (see Fig. 4).

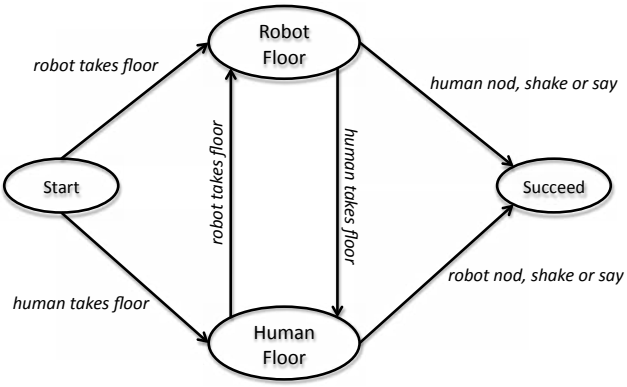

Fig. 12. Recognizer for backchannel (see Fig. 5).

completed its time line or failed (typically because the delay exceeded a threshold).

The integrator process incrementally calculates the mean and maximum time between connection events, the mean and maximum delay times and the number of failed events per unit time, over both a recent time window and the whole interaction (baseline). All of these statistics are available to the rest of the robot architecture to provide an adaptive estimate of the current strength of engagement. For example, increases in recent versus baseline time between connection events, delay time and/or failure rate may indicate the human's desire to disengage. Exactly how to weigh these factors along with other information, such as the content of what the human says, is beyond the scope of the engagement recognition module. Future experimentation with the system may yield further insight into this issue.

IV. Human-Robot Implementation

To implement the architecture described in the preceding section, we chose the ROS framework, because it offered the highest likelihood that our work could be easily shared with other robot researchers and developers. Each of the recognizers in Fig. 8 is implemented as a finite state machine which follows the time line of the corresponding connection event type. These recognizers, together with code implementing the statistics described above are then packaged into what is called a *node* in ROS.

A. Recognizers

Figures 9 through 12 show pseudocode-level state machine diagrams for recognizing each of the four connection event types described by the corresponding time lines in Figures 2 through 5. Notice that the time lines, because they are descriptions from an outside observer's point of view, are symmetric when applied to human-robot interaction, i.e., either the human or the robot can be the initiator or responder for a particular event occurrence. The state machines, however, because they are computations from the point of view of the robot, are asymmetric, i.e., they follow different state transition paths for human-initiated versus robot-initiated events.

Each state machine starts in the state labeled Start and terminates in either a Succeed or Fail state, at which point the relevant statistics for the event occurrence are provided to the statistics process (see Fig. 8). State transitions occur in response to messages coming into the engagement recognition module. Some anomalous transitions have been suppressed for readability, but are included in the ROS documentation.

1) Directed Gaze: In Fig. 9, recognition of a human-initiated directed gaze event is triggered by the human looking and optionally pointing at an object. After transitioning to the

Human Waiting state, the recognizer waits until either the robot decides to respond by looking at the same object (in which case the Shared Gaze state is entered), or time runs out, or the robot decides it wants to make eye contact or direct the human's gaze to another object instead (in which case the event fails). The Shared Gaze state always transitions to Succeed, which occurs when either the human or robot stops looking at the directed object.

The state transition path for recognizing a robot-initiated directed gaze event is similar, except that the directed gaze goal (robot intention) triggers the transition from Start to Robot Waiting. At this point, the robot is supposed to already be looking and optionally pointing at the directed object. As before, the recognizer waits, in this case until the human looks at the directed object, before entering the Shared Gaze state. If time runs out or the robot decides to make eye contact and is not also pointing, then the event fails.

2) Mutual Facial Gaze: Fig. 10 has a similar state structure to directed gaze, with a Mutual Facial Gaze state instead of Shared Gaze. The Mutual Facial Gaze state transitions to Succeed when either the robot or the human breaks eye contact. As in directed gaze, the Human Waiting and Robot Waiting states correspond to the recognition of human-initiated and robot-initiated events, respectively, and each of these states may lead to failure due to timeout. Also, at the point that the mutual facial gaze goal message arrives, the robot is supposed to already be looking at the human's face. Finally, if the robot decides to look at another object (directed gaze goal) during either the Human Waiting or Robot Waiting state, the event fails (because the robot cannot both make eye contact and look at an object at the same time).

3) Adjacency Pair: The state machine in Fig. 11 for recognizing adjacency pair events also has Human Waiting and Robot Waiting states (with timeouts to failure), on the human-initiated and robot-initiated recognition paths, respectively. All the other transitions in this recognizer depend on floor change messages, which come in two forms: taking the floor and giving up the floor. Unlike the previous two recognizers, this state machine could in fact be written more compactly in terms of an initiator and responder, but for consistency of understanding we have expanded out separate paths for the human and robot.

We have not yet implemented the handling of third turns or barge-in (when one party starts taking a turn—not just a backchannel—without the other party first yielding the floor).

4) Backchannel: The state machine in Fig. 12 has no delays or failure states. Basically, the machine keeps track of who has the floor so that it can recognize a backchannel nod or shake by the other party.

B. ROS Node

ROS is a distributed framework of processes (called *nodes*) that communicate via message passing. Nodes are grouped into *packages*, which can be easily shared and distributed. We have contributed a package called "engagement," which currently contains a single node called "recognition." (We eventually expect to add a "generation" node—see Future Work.)

Fig. 13. The pointing game.

TABLE II
SUMMARY STATISTICS FOR HUMAN-ROBOT DEMONSTRATION

		count	delay (sec)		
			min	mean	max
directed gaze	succeed	19	0	0.4	2.3
	fail	50	0	1.6	3.0
mutual facial gaze	succeed	43	0	0.3	1.6
	fail	36	0.1	0.7	1.8
adjacency pair	succeed	21	0	1.0	2.4
	fail	12	3.1	3.1	3.1
mean time between connection events (MTBCE) = 3.0 sec max time between connection events = 9.9 sec					

Information flows into and out of an ROS node via messages (called *topics*) and *services*. Services are a higher-level abstraction that uses messages to implement return values (similar to remote procedure call). Each type of information flowing into the engagement recognition node (see Fig. 7), except for the robot engagement goals, is a separate ROS topic (message type).

C. Preliminary Validation

As a preliminary validation of our computational model and implementation, we developed a simple human-robot demonstration, which we call the "pointing game" (see Fig. 13), that naturally includes the three main engagement behaviors we are studying (no backchannels). Our humanoid robot was built by Michaud *et al.* at U. Sherbrooke (Canada). We used Morency's Watson system [17] for face and gaze tracking and detecting head nods and shakes, and OpenCV to implement plate and hand tracking. Since the focus of our research is on engagement and collaboration, we have simplified the robot's vision problem as much as possible. We used Collagen [18] for the cognitive component of the robot.

In the pointing game, several plates of different colors are place randomly on the table between the human and robot. The robot starts the game by saying "Please point at a plate." The human is then expected to respond by pointing at any plate. The robot identifies the chosen plate by pointing to it and saying, for example, "You pointed at the red plate." If the

human does not respond within a certain amount of time, the robot asks "Do you want to stop now?" If the human nods yes, the robot says "Thank you for playing"; if he shakes no, then the robot repeats its last request.

Our first step was to choose values for the single adjustable parameter of each state machine in Figures 9 through 12, namely the *delay timeout*. We did this subjectively by testing different values starting with the minimum, mean and maximum delays observed in the human study for the corresponding failed event types (see TABLE I). For this testing, we used a simple programming loop in which the robot repeatedly initiated the same event type over and over and waited for the human to respond. The subjectively best delay timeout values were 3.0 sec. for directed gaze, 1.8 sec. for mutual facial gaze and 3.1 sec for adjacency pair. When the timeouts were less than these values, the robot tended to go on before we had time to react; when the timeouts were greater, it felt like we were waiting for the robot a lot.

Next we had three WPI students play the pointing game and collected the aggregated statistics shown in TABLE II. Comparing this data overall with the human data in TABLE I provides a positive preliminary validation. In more detail, notice that the overall pace (MTBCE) was faster in the pointing game than in the human study. We believe this is because the task content in the human study (making canapés) required more thinking time compared to the trivial pointing game. Also, the anomalous (less than timeout) values for minimum delay in failed directed gaze and mutual facial gaze events are due to the fact that, in the current pointing game generation code, the robot sometimes proceeds without responding to human-initiated connection events.

V. Future Work

The most immediate future work is a larger, controlled human-robot study to further validate the engagement recognition model and implementation, using a collaborative task that is more similar in complexity to making canapés. The study should compare conditions in which the delay timeouts are more systematically varied and in which various parts of the recognizer state machines are disabled. Comparisons will include objective measures, such as time and quality of task completion, and subjective post-study questions to the participants about the robot, such as how attentive it seemed, how easy it was to collaborate with, etc.

We have also started working on the problem of how to factor the *generation* of engagement behaviors into a separate reusable module with abstract interfaces to the rest of a generic robot architecture. The decisions in this module concern when the robot should initiate connection events and when/whether it should respond to human-initiated events.

Finally, as mentioned in Section II, we do not believe that the current taxonomy of connection events provides a complete account of the engagement process, particularly as we move beyond the maintenance phase to formalize the initiation and termination of engagement. For example, the effect on engagement of many kinds of nonverbal social and emotional transactions between people, such as laughing, smiling, waving, fidgeting in your seat, etc., need to be studied (even though it may be a while until robotic technology is capable of recognizing or producing all of these). Also, although Bohus and Horvitz [10] have started to model engagement in multiperson interactions, further development of detailed behavioral models, such as those in this paper, is needed.

Acknowledgments

The authors wish to thank Sarah Judd for her work in analyzing connection events in the human engagement study. This work is supported in part by the National Science Foundation under award IIS-0811942.

References

[1] C. L. Sidner, C. Lee, C. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots," *Artificial Intelligence*, vol. 166, no. 1-2, pp. 104–164, 2005.

[2] A. Holroyd, B. Ponsler, and P. Koakiettaveechai, "Hand-eye coordination in a humanoid robot," Major Qualifying Project, Worcester Polytechnic Institute, 2009.

[3] M. Argyle and M. Cook, *Gaze and Mutual Gaze*. New York: Cambridge University Press, 1976.

[4] A. Kendon, "Some functions of gaze direction in two person interaction," *Acta Psychologica*, vol. 26, pp. 22–63, 1967.

[5] S. Duncan, "Some signals and rules for taking speaking turns in conversations," *Journal of Personality and Social Psychology*, vol. 23, no. 2, pp. 283–292, 1972.

[6] C. Goodwin, "Gestures as a resource for the organzation of mutual attention," *Semiotica*, vol. 62, no. 1/2, pp. 29–49, 1986.

[7] Y. Nakano, G. Reinstein, T. Stocky, and J. Cassell, "Towards a model of face-to-face grounding," in *Proc 41st Meeting of Assoc. for Computational Linguistics*, Sapporo, Japan, 2003, pp. 553–561.

[8] H. H. Clark, *Using Language*. Cambridge: Cambridge Univ. Press, 1996.

[9] C. Peters, "Direction of attention perception for conversation initiation in virtual environments," in *Proc. 5th Int. Conf. Intelligent Virtual Agents*, Kros, Greece, 2005, pp. 215–218.

[10] D. Bohus and E. Horvitz, "Models for multiparty engagement in open-world dialog," in *Proceedings of the SIGDIAL 2009 Conference*. London, UK: Ass. for Computational Linguistics, Sept. 2009, pp. 225–234.

[11] ——, "Learning to predict engagement with a spoken dialog system in open-world settings," in *Proceedings of the SIGDIAL 2009 Conference*. London, UK: Association for Computational Linguistics, September 2009, pp. 244–252.

[12] B. Mutlu, T. Shiwa, T. Kanda, H. Ishiguro, and N. Hagita, "Footing in human-robot conversations: How robots might shape participant roles using gaze cues," in *Proc. ACM Conf. on Human-Robot Interaction*, San Diego, CA, 2009.

[13] F. Flippo, A. Krebs, and I. Marsic, "A framework for rapid development of multimodal interfaces," in *Proc. 5th Int. Conf. Multimodal Interfaces*, Nov. 2003, pp. 109–116.

[14] S. Brennan, "How conversation is shaped by visual and spoken evidence," in *Approaches to Studying World-Situated Language Use*, J. Trueswell and M. Tanenhaus, Eds. Cambridge, MA: MIT Press, 1999, pp. 95–129.

[15] D. Crystal, *The Cambridge Encyclopedia of Language*. Cambridge, England: Cambridge University, 1997.

[16] A. Dix, "Pace and interaction," in *Proc. of HCI'92: People and Computers VII*. Cambridge University Press, 1992, pp. 193–207.

[17] L.-P. Morency, A. Rahami, and T. Darrell, "Adaptive view-based appearance model," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Madison, WI, June 2003, pp. 803–810.

[18] C. Rich, C. Sidner, and N. Lesh, "Collagen: Applying collaborative discourse theory to human-computer interaction," *AI Magazine*, vol. 22, no. 4, pp. 15–25, 2001.