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

An Architecture for Engagement in Collaborative Conversations between a Robot and Humans

Candace L. Sidner and Christopher Lee

TR2003-12 March 2003

Abstract

This paper reports on our research on developing the ability for robots to engage with humans in a collaborative conversation. Engagement is the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake. The paper makes two contributions: an architecture for human-robot collaborative conversation with engagement, and a set of rules and associated algorithm for the engagement process.

An Architecture for Engagement in Collaborative Conversations between a Robot and Humans

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

Abstract

This paper reports on our research on developing the ability for robots to engage with humans in a collaborative conversation. Engagement is the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake. The paper makes two contributions: an architecture for human-robot collaborative conversation with engagement, and a set of rules and associated algorithm for the engagement process.

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. 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 [1].

Conversational gestures generally concern gaze at/away from the conversational partner, pointing behaviors (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 are not about where the eye camera gets its input from, but what the eyes, hands, and body tell conversational participants about their interaction.

Not only must the robot produce these behaviors, but it must interpret similar behaviors from its conversational 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], [3]), none of them have attempted to model human-robot interaction to the degree that involves the numerous aspects of engagement and collaborative conversation that we have set out above. Robotics researchers interested in collaboration and dialogue [4] 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, [5],[6]) 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 make two contributions: an architecture for human-robot collaborative conversation with engagement, and a set of rules and associated algorithm for the engagement process. We discuss the architecture and software used in our robot, and we detail the rules and their evaluation for several tested human-robot conversations.

2. Creating a robot with engagement in conversational interaction

To create a robot that can converse, collaborate, and engage a human interactor, a number of different capabilities must be included in the robot's repertoire. The key communicative capabilities include:

- Engagement behaviors: initiate, maintain or disengage in interaction;

- Conversation management: turn taking [7], interpreting the intentions of the conversational participants, establishing the relations between intentions and goals of the participants and relating utterances to the attentional state [8] 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 [9,10]) that CPs produce and observe in their partners. These are capabilities that are also significant to robotic participation in conversation.

In addition to these capabilities, the robot must fuse data gathered from its visual and auditory sensors to determine human gestures, and it must plan and carry out its own appropriate gestures in coordination with spoken utterances during the interaction.

Our current robot, which looks like a penguin, as shown in Figure 1, has limited physical capabilities. While we are developing rules that model gestures from several different physical devices (head turns, gaze, arm movements, body movements, pointing gestures), the robot we currently use only performs head turns (to gaze with fixed eyes), arm movements to indicate beat gestures and head turns to point at objects with its beak. Our robot cannot adjust its body direction at present although a mobile base is forthcoming. Mobility will be crucial for allowing the robot to adjust its body position to address human participants in conversation. 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.

Furthermore, our robot can only perceive a portion of the gestures that humans produce due to limits in the vision algorithms at our disposal. Thus while our architecture is quite general, the robot uses only limited algorithms within that architecture.

3. Architecture

Figure 2 presents the architecture for our robot. The modules of the architecture divide linguistic decisions from sensor and motor decisions. However, information from sensor fusion can cause new tasks to be undertaken by the conversational model. These tasks concern changes in engagement that are signaled by behaviors detected by sensor fusion.

Modules of our architecture:

- Robot motors with 5 degrees of freedom: 3 DOF for head and mouth movement (with pointing via the beak), 1 DOF in appendages, a distance microphone for speech recognition.

Figure 1: Mel, the penguin robot

- Input to Data Fusion comes from two (OrangeMicro iBot) cameras and a pair of microphones.
- Speech and collaborative conversation (Conversation Model) using the Collagen middleware for collaborative agents (11,12) and commercially available speech recognition software (IBM ViaVoice).
- Agent decision-making software in the Conversation Model that determines the overall set of gestures to be generated by the robot motors.
- Sensor Fusion and Robot Control collect data from cameras and microphones, process them (see details below), and provide higher level information to the Conversation Model. The Robot Control synchronizes the set of gestures from the Conversation Model and controls the robot motors.

3.1 Architectural Details

Data Fusion uses the face location algorithm of [13] to find faces, notice when a face disappears, and notice the change of a face from full face to profile. It uses an object tracking algorithm [14] to locate an object to point to and track as the object moves in the visual field. A sound location algorithm detects the source of spoken utterances. The sound location is tuned for speech as opposed to office sounds, such as air conditioning. Results of these three algorithms are used to (1) choose the human CP and his/her location from among the faces detected, (2) pass information about changes in faces and objects to the agent decision-maker in the Conversation Model.

Issues related to spoken language input and output are treated by the speech recognition engine and the Conversation Model. These components interpret sounds and produce syntactic and semantic models of utterances, relate utterances to dialogue context, extract intentions of either human or robot as the speaker and provide possible next spoken actions for the robot. The Conversation Model makes extensive use of the Collagen middleware for collaborative interface agents.

The Collagen system is tailored so that our robot acts as a conversational partner who is hosting a human visitor in our laboratory. Rather than manipulate GUI objects in an interface, our robot is aimed at a collaboration with a human on tasks with objects in the physical world. The Collagen model is based on extensive theory of collaboration [15] and conversation [8,16] and involves direct human-robot interaction rather than teleoperated situations. Our work is complementary to efforts such as [17], who have focused on sharpening the navigational skills of robots with limited human-robot interaction. Our current work extends our first effort [18] to make a robot that could simply talk about a collaborative task and point to objects on a horizontally positioned computer interface.

The Collagen system provides an agenda of next moves. This agenda is expanded by the Collagen agent, which serves to make decisions given the agenda (the agent interface is provided in the Collagen system). It uses engagement rules (discussed in the next section) to determine gestures for the robot, and to assess engagement information about the human CP from the Robot Control and Sensor Fusion module. Decisions by the agent are passed to the Robot Control module for generation of behaviors by robot motors.

Some robotic gestures must be synchronized with spoken language. For example, the beak movement must be timed closely to the start and end of speech synthesis. Likewise, the robot must look at the CP when it passes off the turn. It must also produce beat gestures (with its wings) at the phrases in an utterance that represent new information. To capture this need for synchrony, the robot responds to events generated when the speech synthesis engine reaches certain embedded meta-text markers in the speech text (a method inspired by [19]). The conversation state information from the Conversation model module is thus crucial for the correct operation of the Robot Control module. In addition to synchronizing robot movement with speech and turn state, the module must vary its sensor fusion depending on the conversation state. For example, fusion of visual face location and speech localization information (for determining the location of the human CP) must only be performed when the conversational model indicates the human has the turn.

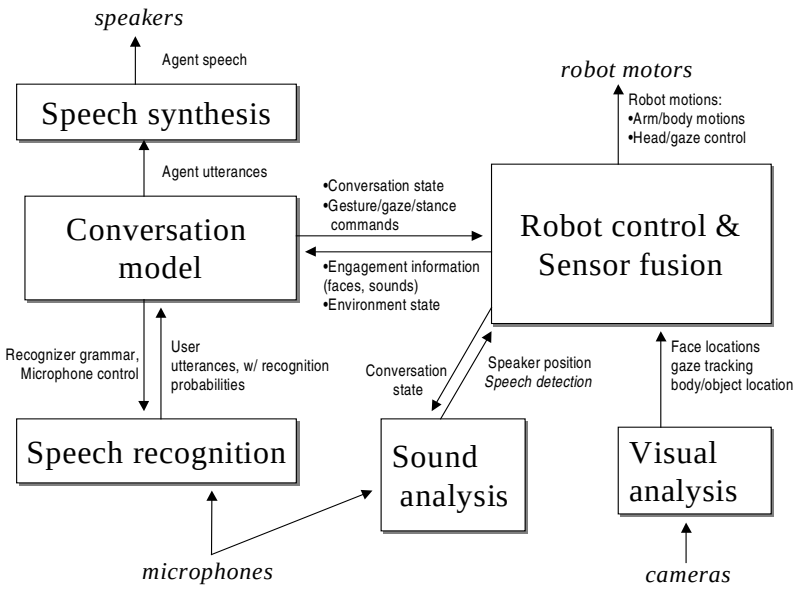

Figure 2: Architecture for Robot with Engagement

4. Engagement Rules and Evaluation

To determine gestures, we have developed a set of rules for engagement in the interaction. These rules are gathered from the linguistic and psycholinguistic literature as well as from 3.5 hours of videotape of humans performing a hosting activity (that is, human host guiding a human guest on tour of laboratory artifacts).

These gestures, we hypothesize, reflect US standard cultural rules for subjects in the US speaking English (that is, the default cultural norms used by persons dwelling in the US in English language interactions with any other persons there, even if they are from cultural groups with other cultural norms).

Our first set of rules, which we have tested in scenarios described below, are a small and relatively simple set. These scenarios do not involve pointing to or manipulating objects. Rather they are focused on engagement in more basic conversation. These rules direct the robot:

(1) to gaze at a human to begin to engage the human in interacting, followed by a conversational greeting,

(2) once the human has responded to the initiation of engagement, to look at the human when he or she takes the turn to speak in the conversation,

(3) when the robot begins its speaking turn, to briefly look away,

(5) during the robot's speaking turn, to mainly look at the human who is the CP, but also to look at onlookers,

(6) to notice but not disengage when the human either fails to take the turn to speak or moves his/her head out of the robot's visual field,

(6) to ascertain, via dialogue, the human's desire to disengage when both loss of the human's face and human failure to take a turn occur,

(7) to disengage quickly when in (5) the human desires to end the conversation (disengagement is a brief "goodbye" and a long look away from the human),

(8) to end the conversation with a normal conversational closing [20,21] when the robot runs out of things to discuss.

While this is a fairly small repertoire of engagement behaviors, it was sufficient to test the robot's behavior in a number of scenarios. Since evaluation of an engaging, conversational robot is uncharted territory, we choose to undertake evaluation by simple criteria. We wished to determine which robot gestures were either counter to known rules of gesture in conversation or seemed unnatural to us, as we observed the robot in conversation. Furthermore, while humans speak to the robot in English, we have limited the variety of utterances people can speak because we did not want to focus our research on the problem for speech research of interpreting fluent conversational utterances.

Our scenarios include:

1. Single human CP and robot in uneventful conversation about visiting the research lab in which the robot is situated.

2. Single human CP and robot, 2 onlookers in same type of conversation above.

3. Single human CP and robot, no onlookers where the CP departs unexpectedly.

4. Single human CP and robot, with onlookers where CP departs unexpectedly.

5. Single CP and robot, with and without onlookers where the human CP terminates interaction before task is complete.

6. Single CP and robot, with and without onlookers where interaction ends because the robot has run out of things to discuss with human.

Figure 3 provides a portion of the transcript of a videotape clip (to be available with the paper) that illustrates scenario 2. The robot is called Mel in this dialogue. Figure 4 depicts the situation of scenario 1.

<Robot Mel finds a face belonging to a human, David, who will be an onlooker in the conversation; it turns its head to gaze at David. It is also aware of other humans, Chris and Candy, from the face detection algorithm.>

Mel: hello there

Chris: hi Mel

<Mel turns to gaze at Chris and decides Chris is the conversational partner>

Mel: Would you please tell me your name? If you are visiting Merl, just say so.

Chris: Chris.

Mel: well, Chris as you can see, I'm a conversational penguin. I can also track your face, but I can't recognize your face or see <Robot turns head to onlooker David at this moment> objects in the room.

Chris: ok.

<Robot turns to gaze back to Chris> ……..

Figure 3: A portion of the transcript for of Scenario 2

Figure 4: Robot and Human Partner in Scenario 1

The results of our evaluations have indicated three sets of problems:

- incorrect gaze gestures with onlookers at end of turn (incorrect behavior given known rules of conversation),
- incorrect uptake of conversation with onlookers when the human CP disappears (unnatural behavior),
- failure to deal with the human CP's random glance behavior that is of long enough duration to be noticed by the face detection algorithm (incorrect behavior of known rules for gesture).

As a result, we are now developing a rule framework with statistical weightings as well as rule-like constraints. This framework will also be useful for scenarios involving laboratory demos of objects, which must be discussed, pointed to and looked at.

5. Future Directions

In addition to the combined statistical and rule-based engagement paradigm, we will be adding mobility to our robot, so that it can change its position of address, and move about to point at objects. These changes will also require a larger collection of rules than our initial efforts, so the new engagement paradigm will be valuable in light of these changes.

We also plan a new means of evaluating the engagement model using a pair of simulated and animated robots, both of which are run from the same architecture and rule set. These simulations will be tested on an expanded set of scenarios. Finally we want to test the robot using the standard training set/test set model used in many other parts of computational linguistics.

6. Summary

This paper has discussed the nature of engagement in human-robot interaction, provided rules for engagement, and detailed an architecture for robots engaged in collaborative conversation with humans.

7. Acknowledgements

The authors wish to acknowledge the work of Neal Lesh and Charles Rich on aspects of Collagen critical to this effort.

8. References

1. McNeill. D. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, Chicago, 1992.
2. Breazeal, C. "Affective interaction between humans and robots", *Proceedings of the 2001 European Conference on Artificial Life (ECAL2001)*. Prague, Czech Republic, (2001).
3. Kanda, T. Ishiguro, H. Imai. M. Ono, T. and Mase, K. "A constructive approach for developing interactive humanoid robots. *Proceedings of IROS 2002*, IEEE Press, NY, 2002.
4. Fong, T., Thorpe, C. and Baur, C. Collaboration, Dialogue and Human-Robot Interaction, *10th International Sumposium of Robotics Research*, Lorne, Victoria, Australia, November, 2001.
5. J. Cassell, J. Sullivan, S. Prevost and E. Churchill, *Embodied Conversational Agents*. MIT Press, Cambridge, MA, 2000.
6. Johnson, W. L., Rickel, J. W. and Lester, J.C. "Animated Pedagogical Agents: Face-to-Face Interaction in Interactive Learning Environments," *International Journal of Artificial Intelligence in Education*, 11: 47-78, 2000.
7. Duncan, S. (1974). Some signals and rules for taking speaking turns in conversation. in *Nonverbal Communication*, S. Weitz (ed.), New York: Oxford University Press.
8. Grosz, B. J. and Sidner, C. L. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175--204, 1986.
9. Halliday, M.A.K., *Explorations in the Functions of Language*, London: Edward Arnold, 1973.
10. Cassell, J. (2000a). Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents. in *Embodied Conversational Agents*, J. Cassell, J. Sullivan, S. Prevost, and E. Churchill (eds.), Cambridge, MA: MIT Press.

11. Rich, C., Sidner, C. L. and Lesh, N. "COLLAGEN: Applying Collaborative Discourse Theory to Human-Computer Interaction," *AI Magazine, Special Issue on Intelligent User Interfaces*, AAAI Press, Menlo Park, CA, Vol. 22: 4: 15-25, 2001.
12. Rich, C. and Sidner, C. L. "COLLAGEN: A Collaboration Manager for Software Interface Agents," *User Modeling and User-Adapted Interaction*, Vol. 8, No. 3/4, 1998, pp. 315-350.
13. 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.
14. Beardsley, P.A. *Piecode Detection*, Mitsubishi Electric Researchs Labs TR2003-11, Cambridge, MA, February, 2003.
15. Grosz, B. J. and Kraus, S. "Collaborative Plans for Complex Group Action," *Artificial Intelligence*, 86(2): 269-357, 1996.
16. Lochbaum, K. E. A Collaborative Planning Model of Intentional Structure. *Computational Linguistics*, 24(4): 525-572, 1998.
17. Burgard, W., Cremes, A. B., Fox, D., Haehnel, D., Lakemeyer, G., Schulz, D., Steiner, W. & Thrun, S. "The Interactive Museum Tour Guide Robot," *Proceedings of AAAI-98*, 11-18, AAAI Press, Menlo Park, CA, 1998.
18. Sidner, C. and Dzikovska, M. Hosting activities: Experience with and future directions for a robot agent host. *Proceedings of the 2002 Conference on Intelligent User Interfaces*, New York: ACM Press. pp. 143-150, 2002.
19. Cassell, J., Vilhjálmsson, H., & Bickmore, T. W. (2001a). BEAT: the behavior expression animation toolkit. *Proceedings of SIGGRAPH 2001*. New York: ACM Press. pp. 477-486.
20. Schegeloff, E. & Sacks, H. (1093). Opening up closing. *Semiotica*, 7(4): 289-327.
21. Luger, H.H. "Some Aspects of Ritual Communication," *Journal of Pragmatics.* Vol. 7: 695-711, 1983.