6

Attentional Gestures in Dialogues between People and Robots

Candace Sidner and Christopher Lee

Abstract

Attentional gestures in dialogues provide critical cues to the focus of attention of the participants in the dialogues. Attentional gestures are evidence of the engagement process by which participants start, maintain and end their perceived connection to one another. This article investigates engagement and gestures that indicate engagement. It then applies these concepts to the development of a humanoid robot that converses with a human participant. Evaluations of conversations between people and the robot consider the effects of looking and nodding gestures.

6.1 Introduction

Gestures are fundamental to human interaction. When people are face-to-face at near or even far distance, they gesture to one another as a means of communicating their beliefs, intentions and desires. When too far apart or in too noisy an environment to converse, gestures can suffice, but in most human face-to-face encounters, conversation and gesture co-occur. According to the claims of (McNeill, 2002), they are tightly intertwined in human cognition. However cognitively entangled, people gesture freely, and the purposes of those gestures are fundamental to our work on human robot interaction.

Some gestures, especially those using the hands, provide semantic content; these gestures supplement the content of utterances in which the gestures occur (Cassell, 2000). However, some gestures indicate how the conversation is proceeding and how engaged the participants are in it. Attentional gestures are those that involve looking with head movements and eye gaze, and those involving body stance and position. Hand gestures by speakers also can be used to direct attention toward the gesture itself (Goodwin, 1981). All these gestures convey what the participants are or should be paying attention to. The function of these gestures is distinguished by their role in the ongoing interaction. Nodding gestures also can indicate that the participants are paying attention to one another. Nodding *grounds* (Clark 1996) previous comments from the other speaker, while looking conveys attention for what is coming.

We call these attentional gestures "engagement behaviors," that is, those behaviors by which interlocutors start, maintain and end their perceived connection to one another. The process by

which conversational participants undertake to achieve this connection is the process of *engagement*. These behaviors indicate how each participant is undertaking their perceived connection to one another. They also convey whether the participants intend to continue their perceived connection. Of course, linguistic gestures (that is, talking) are a significant signal of engagement. However, linguistic and non-linguistic gestures of engagement must be coordinated during face-to-face interactions. For example, a speaker's ongoing talk while he or she looks away from the hearer(s) into blank space for a long time conveys contradictory information about his or her connection to the hearer; she's talking but doesn't appear interested. While other researchers, notably Clark and his students, have been investigating how conversational participants come to share common ground, and have considered some aspects of non-verbal gesture in that light, the overall problem of how participants come to the attention of one another, how they maintain that attention, how they choose to bring their attention to a close, and how they deal with conflicts in these matters is an open matter of investigation. In this article, all of our observations concern gestural phenomena that are typical of North Americans interacting. While we believe that engagement is natural in human-human interactions in all cultures, the details of how this is signaled will vary and will require further study.

It is essential to understand the process by which engagement transpires. It is part and parcel of our everyday face-to-face interactions, and it may well be relevant even in non face-to-face circumstances such as phone conversations. Furthermore a clear picture of the engagement process is necessary to reproduce natural communication behavior for artificial agents, either those in embodied on-screen characters that interact with computer users or those in the form of humanoid robots.

6.2 Background and Related Research

Head nods are a well-studied behavior in human conversation (Duncan, 1973, Yngve, 1970). They are used for establishing common ground, that is, shared knowledge between conversational participants (Clark, 1996), and generally are accompanied by phrases such as “uh-huh, ok” or “yes.” They also occur in positive responses to yes/no questions and as assents to questions or agreements with affirmative statements. McClave (2000) observed that head nods serve as backchannels and as encouragement for the listener. Bickmore (2002) observed that people talking to people in an experiment with a hand-held device nodded 1.53 times per turn either as acknowledgements, agreements or, in the case of the speaker, for emphasis.

Argyle and Cook (1976) documented the function of gaze as an overall social signal and noted 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 hearers, in particular, how gestures direct gaze to the face and why gestures might direct it away from the face (Kendon, 1967; Duncan 1973; Goodwin, 1986). Nakano et al's (2003) work on grounding reported on the use of the hearer's gaze and the lack of negative feedback to determine whether the hearer has grounded the speaker’s turn. Our own studies of conversational tracking of the looks (using head directions only, not eye gaze) between a human speaker and human hearer observed that the hearer tracked the speaker about 55% of the time. The non-tracking times were characterized by attention to other objects, both ones that were relevant to their conversation and ones that were evidence of multi-tasking on the part of the hearer. However, sometimes for very short looks away on the part of the speaker, the hearer simply did not give evidence of attending to these looks (Sidner et al, 2005).

While we have investigated people talking to people in order to understand human behavior, we are also interested in how these behaviors translate to situations where people and robots converse. Robots that converse but have no gestural knowledge (either in production or in interpretation)

may miscue a human and miss important signals from the human about engagement. Nodding on the part of humans is so unconscious, that people nod in conversations with robots, even though the robot has no ability to recognize the nods, as we will show later in this article. A portion of the time, they also nod without accompanying verbal language, so the multi-modal aspects of gesture and language cannot be overlooked in human-robot interaction or in conversations with embodied conversational characters.

Previous work in human-robot interaction has largely explored gaze and basic interaction behavior. Breazeal's work (2002) on infantoid robots explored how the robot gazed at a person and responded to the person's gaze and prosodic contours in what might be called pre-conversational interactions. Other work on infantoid robot gaze and attention can be found in (Kozima et al, 2003). Minato et al (2002) explored human eye gaze during question answering with an android robot; gaze behavior differed from that found in human-human interaction. More recent work (Breazel et al, 2004) explores conversation with a robot learning tasks collaboratively, but the robot cannot interpret nods during conversation. Ishiguro et al (2003) report on development of Robovie with reactive transitions between its behavior and a human's reaction to it; they created a series of episodic rules and modules to control the robot's reaction to the human. However, no behaviors were created to interpret human head nods. Sakamoto et al (2005) have experimented with cooperative behaviors on the part of the robot (including robot nodding but not human nodding) in direction giving.

Thus we are interested in imbuing the robot with the ability to engage gesturally by looking, standing and nodding appropriately in interactions with humans and by properly interpreting such behaviors from people. This article reports on our progress thus far in such efforts.

6.3 A Conversational Robot

The robot we have developed interacts with a single user in a collaboration that involves: spoken language (both understanding and generation) using a mouth that opens and closes in coordination with the robot's utterances, gestures with its appendages, head gestures to track the user and to turn to look at objects of interest in the interaction, recognition of user head gestures in looking at objects, and recognition of user head nods. The robot also initiates interactions with users, and performs typical preclosings and goodbyes to end the conversation. All these capabilities increase the means by which the robot can engage the user in an interaction.

The robot, called Mel and embodied as a penguin, has the following hardware:

- 7 DOF in the body (1 in the beak/mouth, 2 in the head, 2 in each of the 2 wings),
- body mounted on a Pioneer II mobile robot platform for floor navigation and body positioning,
- stereo camera,
- 2 far distance microphones (one for speech recognition, one for speech detection),
- 2 onboard laptop computers and an onboard PC-104 computer for all software.

The robot is depicted in Figure 1, and the architecture for the robot is displayed in Figure 2.

The architecture of this robot is divided into a conversational subsystem and a sensorimotor subsystem. The conversational subsystem is based on the Collagen conversation and collaboration manager (Rich et al, 2001). The sensorimotor subsystem controls the robot's physical hardware and performs sensor fusion. Together, these subsystems maintain a model of

the dialog and of the user's engagement, and use this model to generate appropriate verbal and gestural behaviors.

Figure 1: Mel the robot penguin

The sensorimotor subsystem is based on a custom, task-based blackboard robot architecture. It detects and localizes human speech, and uses vision algorithms of (Viola and Jones, 2001) to detect and recognize faces, and (Morency et al, 2002) to track the 6-DOF position and orientation of one face. The subsystem fuses the sound and vision information to locate the robot's conversational partner, and roughly track that person's gaze from their head direction. In addition, it interprets certain of the human's head movements as head nods (Morency et al, [this volume] and 2005). The sensorimotor subsystem receives information from the conversational subsystem about the dialog state, commands for head and arm gestures, and commands to alter body stance (via the mobile base), and gaze.

Dialog state information from the conversation subsystem is vital to the sensorimotor subsystem, both for proper generation of robot motions and for interpretation of sensor data. For example, the robot's gaze typically tracks the face of the human conversational partner, but the robot glances away briefly when it takes the floor and starts to speak. Wing gestures of the penguin signal (1) new information in its utterances (so called *beat* gestures, see (Cassell et al, 2001)) and (2) provide pointing gestures to objects in the environment. The robot turns away from the conversational partner when it must point to an object in the demo, and it monitors the orientation of the person's head to ensure that the person follows the pointing motion. Proper tracking of the robot's pointing by the human is interpreted semantically, based on dialog context, by the conversational subsystem ("User looks at cup"). Failure to track within a reasonable amount of time may result in further prompting and gesturing ("The cup is here to my right").

Before the robot finishes speaking an utterance, it prepares for the human's response by activating speech recognition grammars tailored to the expected range of verbal responses. It also may adjust its priors for recognizing head motions as nods with respect to the content and timing of the robot's utterance (Morency et al, 2005). When the robot finishes speaking, it

activates the microphones for speech detection and recognition. If the robot detects a head nod at this time, it checks to see if it also detects speech. If no speech is detected, the nod is immediately interpreted as "yes" or an acknowledgment ("OK"). Otherwise, the robot waits a short time for a result from speech recognition, and only interprets the nod if no recognition result is generated or the speech was not understood.

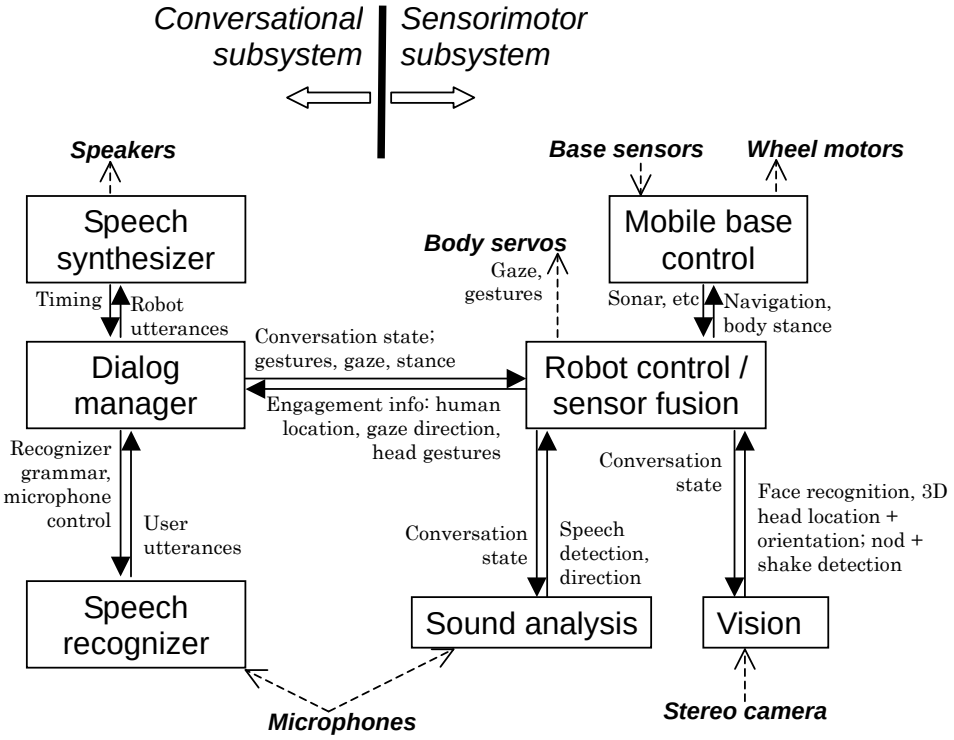

Figure 2: Architecture for a Conversational Robot

The robot participates in two different interactions with human participants. The first of these is a demonstration of IGlassware, a technology developed at Mitsubishi Electric Research Laboratories. IGlassware is a specially designed cup that senses fluid level using capacitance and communicates the level to a special table by inductance (Dietz et al, 2002). Mel was programmed to demonstrate this invention to a human visitor. A typical interaction is shown in Figure 3; the robot's gestures with its head towards the table are noted in the text. However, many different conversations are possible, including one that is not about the demo at all (if the human participant refuses to be shown a demo).

M: Hi, I'm Mel a robotic penguin.
U: Hi.
M: What's your name?
U: Sam.
M: Sam, I'd like to show you a demo. OK?
U: No.
M: But it's really interesting. Come on. Try it!
U: Okay.

M: It would be really nice to know what kind of progress people are making in their dining.
M: IGlassware stands for Instrumented glassware!
U: Okay.
M: Sam, right there <robot turns head to table> is the IGlassware cup <robot turns to user>.
U: Okay.
M: All right, now we'll see how to use the cup.
U: All right.
M: Please pour enough water from the pitcher into the glass to fill the cup.
Then put it on the table.
U: Okay.
M: I'm waiting for a reading from the table for the glass.
M: Good. Take a look at the bar on the readout.
U: Yes.
M: It shows that the cup is full, right?
U: Yes.
M: Would you like me to explain how this works?
U: Sure.
M: You see the copper in the glass? <robot turns head to glass>….
[continued conversation not illustrated]
M: Well, Sam, that's about all.
See Candy Sidner or Chris Lee for more about IGlassware. So long.
U: Good-bye.

Figure 3. Excerpts from the IGlassware Demo with Mel

The robot can also demonstrate its own capabilities in an interactive conversation. A portion of one such conversation can be found in Figure 4. During this conversation, the robot gestures at its own body parts when relevant, and in the case of its mobile base, moves around to demonstrate its locomotive abilities. We have observed that people interacting with the robot during this conversation maintain their gaze on the robot through most of the conversation with relatively few looks away (in contrast to the IGlassware conversations in which users direct much of their attention to the IGlassware cup and table). We surmise that gaze maintenance is due to the fact that there is nothing else that is relevant to the conversation, and due to the fact that the robot is pointing out parts of itself in the conversation.

M: Hello, I'm Mel, a robotic penguin.
U: Hello.
M: What's your name?
U: Mike.
M: Mike, I'd like to tell you about myself. Okay?
U: Ok.
M: First I'll explain my body.
U: Ok.
M: Well. I have 7 degrees of freedom in my body parts plus my mobile base can
turn wherever I want it to. <robot rotates base during conversation>
U: Ok. [continuing conversation not illustrated]
M: How about we discuss my vision and speech systems?
U: Ok.
M: Shall I explain my vision system, or my speech abilities first?
U: Your vision system.
M: Well, above my head is a black camera with two lenses. Do you see it?
<robot looks up above head during this exchange>

U: Yes.
M: I see with my two lens eyes.... [continuing conversation not illustrated]
M: Well, that's all my parts. Are you curious about what I can do with all that?
U: Yes.
M: Mostly I can talk about things. And I give demos. And I'd like to learn how to tell jokes well.
U: Ok.
M: So you wanna hear a joke?
U: No.
M: Well, Mike that's about all.
U: Ok.
M: See Candy Sidner or Chris Lee, for more about me! So long!
U: Good-bye.
M: Good-bye.

Figure 4: Excerpts from a Mel self demo

When initiating both interactions, Mel uses his vision system to find a conversational partner. Thereafter Mel tracks the conversational partner's face and adjusts his "gaze" towards the partner, even when the partner moves about. Mel has eyes in his head, but they do not see, and his camera is above his head, so his gaze serves only to communicate his focus of attention to the partner. Mel does not look at the human partner at all times. During the demos, he turns to look at the table and its contents or to parts of himself. Mel also prompts a partner who fails to look at the table to notice the objects there. After the demo and explanation conclude, Mel wishes the partner goodbye, waves and drops his head to his chest to indicate that he is no longer available.

6.4 Looking Behaviors for the Robot

Mel the robot is able to follow the looking gestures of a human partner using the vision technology discussed above. The robot also displays looking ability when he turns to relevant objects in the room, points at them and then turns back to look at his human interlocutor. Does this robot's non-verbal gestural behavior have an impact on the human partner? The answer is a qualified yes. We conducted an experiment with 37 human participants who were given an IGlassware demo by the robot. Twenty of these participants interacted with the robot with full gestural behaviors and with tracking of the human head movements (the robot indicated the tracking by its head movement), while seventeen participants interacted with a robot that only moved his beak but produced no other gestures (Sidner et al, 2005). This "wooden" robot looked straight at the position where it first saw the human's head and did not change its head position thereafter. Just as the moving robot did, the wooden robot also moved its beak with his utterances and noticed when the human looked at objects in the scene.

Our "wooden" robot allowed us to study the effects of just having a linguistic conversation without any looking behavior. In comparison, the moving robot performed looking gestures by tracking the person's face, looking at relevant objects (for a brief period of time) and then returning to look at the user. These behaviors permitted us to study the effects of looking in concert with conversation. Informally we observed that Mel, by talking to users and tracking their faces, gave users a sense of being paid attention to. However, whether this "sense" had any effect on human robot interaction lead us to the study with people interacting with the two versions of our robot.

Both "wooden" and moving robots assessed human engagement. Humans who did not respond to their turn in the conversation were judged as possibly wanting to disengage, so in both conditions, the robot would ask, "Do you want to continue the demo?" when humans failed to take a turn. Human looks away were not judged by themselves as expressing disengagement. In fact most looks away were to the demo objects, or in the case of IGlassware, to the area under the table to ascertain how the table worked. However, when humans did not perform shared attention to the demo objects when the robot introduced them into the discussion, the robot prompted the human with an additional utterance and looking gesture to encourage them to view the object. Thus our robot both makes decisions about where to look itself and assesses aspects of joint attention with its human conversational partner. As Nakano and Nishida (this volume) demonstrate, shared attention to objects in the environment is a critical attentional behavior in human-agent interactions.

In our experiment, each person had a conversation with the robot (before which the person was told that they would see a demonstration with a robot) and after filled out a questionnaire about their experience. From the questionnaires, we learned that all participants liked the robot. However, those who talked to the moving robot rated it as having more appropriate gestures than participants with the wooden robot participants did.

While questionnaires tell us something about people's response to a robot, videotapes of their interactions revealed more. We studied videotapes of each conversation between person and robot. From these studies, we discovered that in both conditions, about 70% of the time, the human participants who were looking away from the robot looked back to the robot when it was their turn to speak in the conversation. It seems that conversation alone is a powerful mechanism for keeping people engaged. In particular, in human-human interactions, people look at their conversational partners when they have the conversational turn. People behaved similarly when their conversational partner was a robot, even one whose looking gestures were non-existent.

However, two other aspects of human looking produced different effects in the "wooden" versus moving condition. Firstly, the human participants looked back at the moving robot significantly more often even when it was not their turn than their counterparts did with the "wooden" robot (Single-factor ANOVA, $F[1, 36] = 15.00$, $p < 0.001$). They determined that the robot was to be paid attention to and looked at it even when they were also trying to look at the demo objects.

Secondly, there was a weak effect of the robot's looking by turning its head towards objects relevant to the demo. We observed that human participants tracked very closely the turn of the robot's head by themselves turning to look at the IGlassware table or glass. This behavior was more common than for participants who only had utterances (such as "Right here is the IGlassware cup") to direct them to objects. Looking gestures of this type provide a strong signal of where to look and when to start looking, and our participants responded to those signals.

In sum, the moving robot's looking gestures affected how people participated in conversation. They paid closer attention to the robot, both in terms of what it said and in terms of where it looked at objects. All these behaviors are indicative of engagement in the interaction. Without overt awareness of what they were doing, people's gestures indicated that they were engaged with the moving robot.

6.5 Nodding at the Robot

Head nodding is one of the many means by which participants indicate that they are engaging in the conversation. In the experiments discussed in the previous section, we observed that the

participants nodded to the robot to acknowledge his utterances, to answer “yes” to questions and to assent to some statements. However, the robot had no ability to interpret such nods. Thus we were determined to see if this behavior would continue under other circumstances. Furthermore, because nodding is a natural behavior in normal human conversation, it behooved us to consider it as a behavior that a robot could interpret.

In our experiments (Sidner et al, 2006), human participants held one of two conversations with the robot, shown in the previous section, to demonstrate either its own abilities or the IGlassware equipment. During these conversations people nodded at the robot, either because it was their means for taking a turn after the robot spoke (along with phrases such as "ok" or "yes" or "uh-huh"), or because they were answering in the affirmative a yes/no question and accompanied their linguistic "yes" or "ok" with a nod. Participants also shook their heads to answer negatively to yes/no questions, but we did not study this behavior because too few instances of “no” answers and headshakes occurred in our data. Sometimes a nod was the only response on the part of the participant.

The participants were divided into three groups, called the MelNodsBack group, MelOnlyRecognizesNods group and the NoMelNods group. The NoMelNods group with 20 participants held conversations with a robot that had no head nod recognition capabilities. This group served as the control for the other two groups.

The MelNodsBack group with fifteen participants, who were told that the robot understood some nods during conversation, participated in a conversation in which the robot nodded back to the person every time it recognized a head nod. It should be noted that nodding back in this way is not something that people generally do in conversation. People nod to give feedback on what another has said, but having done so, their conversational partners only rarely nod in response. When they do, they are generally indicating some kind of mutual agreement. Nonetheless, by nodding back, the robot gives feedback to the user on their behavior. Due to mis-recognition of nods, this protocol meant that the robot sometimes nodded when the person did not nod.

The MelOnlyRecognizesNods group with fourteen participants held conversations without the robot ever nodding back, and without the knowledge that the robot could understand head nods although the nod recognition algorithms were operational during the conversation. We hypothesized that participants might be affected by the robot's nodding ability because 1) when participants responded only with a nod, the robot took another turn without waiting further, and 2) without either a verbal response or a nod, the robot waited a full second before choosing to go on. Hence participant nods caused the robot to continue talking. Therefore, we hypothesized that, over the whole conversation, the participants might have gotten enough feedback to rely on robot recognition of nods. This group provided an opportunity to determine whether the participants were so affected. Note also that as with the MelNodsBack group, nod mis-recognition occurred although the participants got no gestural feedback about it.

Every conversation with the robot in our total of 49 participants varied in the number of exchanges held. Hence every participant had a varying number of opportunities to give feedback with a nod depending on when a turn was taken or what question was asked. This variation was due to: different paths through the conversation (when participants had a choice about what they wanted to learn), the differences in the demonstrations of IGlassware and of the robot itself, speech recognition (in which case the robot would ask for re-statements), robot variations in pausing as a result of hearing the user say "ok," and instances where the robot perceived that the participant was disengaging from the conversation and would ask the participant if they wished to continue.

In order to normalize for these differences in conversational feedback, we coded each of the individual 49 conversations for feedback opportunities in the conversation. Opportunities were defined as the end of an exchange where the robot paused long enough to await a response from the participant before continuing, or exchange ends where it waited only briefly but the participant chose to interject a verbal response in that brief time.

So for each participant, our analysis used a "nod rate" as a ratio of total nods to feedback opportunities, rather than the raw number of nods in an individual conversation. Furthermore, the analysis made three distinctions: nod rates overall, nod rates where the participant also uttered a verbal response (nod rates with speech) and nod rates where no verbal response was uttered (nod only rates).

Our study used a between-subjects design with Feedback Group as our independent variable, and Overall Nod Rate, Nod with Speech Rate, and Nod Only Rate as our three dependent variables.

A one-way ANOVA indicates that there is a significant difference among the three feedback groups in terms of Overall Nod Rate ($F_{2,46} = 5.52$, $p < 0.01$). The mean Overall Nod Rates were 42.3%, 29.4%, and 20.8% for MelNodsBack, MelOnlyRecognizesNods, and NoMelNods groups respectively. A post-hoc LSD pairwise comparison between all possible pairs shows a significant difference between the MelNodsBack and the NoMelNods groups ($p=0.002$). No other pairings were significantly different. The mean Overall Nod Rates for the three feedback groups are shown in Figure 5.

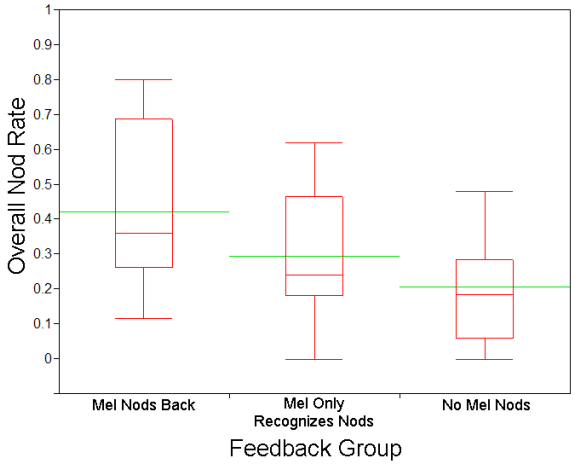

Figure 5: Overall Nod Rates by Feedback Group.

A one-way ANOVA indicates that there is also a significant difference among the three feedback groups in terms of Nod with Speech Rate ($F_{2,46} = 4.60$, $p = 0.02$). The mean Nod with Speech Rates was 32.6%, 23.5%, and 15.8% for the MelNodsBack, MelOnlyRecognizesNods, and NoMelNods groups respectively. Again, a LSD post-hoc pairwise comparison between all possible pairs of feedback groups shows a significant difference between the MelNodsBack and NoMelNods groups ($p=0.004$). Again, no other pairs were found to be significantly different. The mean Nod with Speech Rates for the three feedback groups is shown in Figure 6.

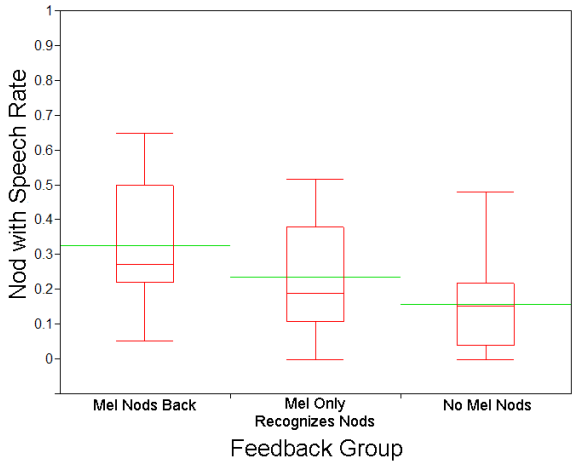

Figure 6: Nod with Speech Rates by Feedback Group

Finally, a one-way ANOVA found no significant differences among the three feedback conditions in terms of Nod Only Rate ($F_{2,46} = 1.08$, $p = 0.35$). The mean Nod Only Rates were more similar to one another than the other nod measurements, with means of 8.6%, 5.6 and 5.0% for the MelNodsBack, MelOnlyRecognizesNods, and NoMelNods groups respectively. The mean Nod Only Rates for the three feedback groups are shown in Figure 7.

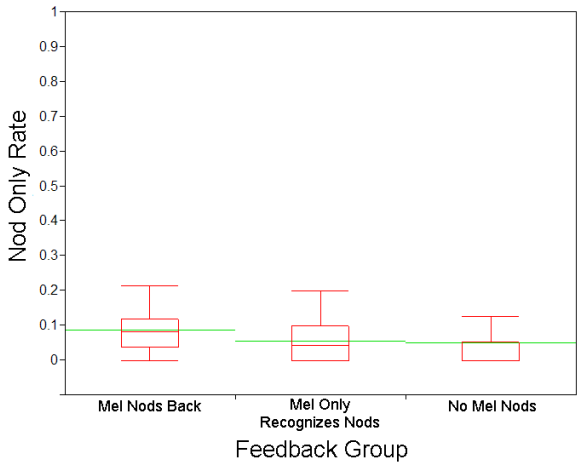

Figure 7: Nod Only Rates by Feedback Group

These results above indicate that under a variety of conditions people will nod at a robot as a conversationally appropriate behavior. Furthermore, these results show that even subjects who get no feedback about nodding do not hesitate to nod in a conversation with the robot. We conclude that conversation alone is an important feedback effect for producing human nods, regardless of the robot's ability to interpret it.

It is worthwhile noting that the conversations our participants had with robots were more than a few exchanges. While they did not involve the human participants having extended turns in terms of verbal contributions, they did involve their active participation in the purposes of the conversation. Future researchers exploring this area should bear in mind that the conversations in this study were extensive, and that ones with just a few exchanges might not see the effects reported here.

The two statistically significant effects for nods overall and nods with speech that were found between the NoMelNods group and the MelNodsBack group indicate that providing information to participants about the robot's ability to recognize nods and giving them feedback about it makes a difference in the rate at which they produce nods. This result demonstrates that adding perceptual abilities to a humanoid robot that the human is aware of and gets feedback from provides a way to affect the outcome of the human and robot's interaction.

The lack of statistical significance across the groups for nod rates without any verbal response (nod only rates) did not surprise us. The behavior of only nodding (without accompanying speech) in human conversation is a typical behavior, although there are no statistics that we are aware specifying rates of such nods. It is certainly not as common as nodding and making a verbal response as well. Again, it is notable that this behavior occurs in human-robot interaction and under varying conditions. By count of participants, in the NoMelNods group, 9 of 20 participants, nodded at least once without speech, in the MelOnlyRecognizesNods group, 10 of 14 did so and in the MelNodsBack group, 12 of 15 did so. However, when using normalized nod rates for this behavior for each group, there is no statistical difference. This lack is certainly partially due to the small number of times subjects did this: the vast majority participants (44/49) did a nod without speech only 1, 2, or 3 times during the entire conversation. A larger data set might show more variation, but we believe it would take a great deal of conversation to produce this effect. We conclude that nods without speech are a natural conversational behavior. Hence using vision techniques with a robot that capture this conversation behavior is valuable to any human-computer conversation, whether it’s with a robot or with an onscreen agent.

6.6 Lessons Learned

Looking at another person, looking at objects when one mentions them, especially initially, nodding appropriately in conversation and recognizing the nods of others are all gestures that people typically produce in interactions with others. While looking gestures are clearly indicative of engagement, nodding gestures may seem less so since their most obvious function is to supplement and sometimes replace utterances such as "okay," "yes," "uh-huh." In this respect nodding may be thought be akin to iconic gestures, which some researchers believe supplement or provide additional semantic content during speaking (McNeill, 1992).

However, nodding gestures are principally feedback to the speaker (cf. Nakano and Nishida [this volume]). They work even if the hearer is not looking at the speaker and does not say anything. In our observations of human-human interactions (Sidner et al, 2005), we observed many occasions where a hearer nodded (sometimes without speech) to provide feedback to the speaker that they were following the speaker's utterances. And as we discovered, people naturally nod at robots during conversation. In our human-robot studies, our human participants nodded both when looking at the robot and when looking at demo objects.

This feedback serves to signal engagement in subtle ways. While speakers will proceed without a nod or a "yes" when they have indicated that they expect such feedback[1], it is unlikely they typically will do so for a long stretch of conversation. Speakers find they must know if their conversational partners are attending to them. When a hearer does not provide that feedback on a regular basis, their behavior indicates a failure to be tracking the speaker and thereby failing to be engaged in the interaction. Conversely, there are speakers who disregard the lack of feedback from their hearers. While this behavior is not the normative way to behave, it still occurs. Importantly, speakers who speak for long stretches of conversation and ignore the lack of feedback from hearers are generally judged as holding the floor of the conversation in an unacceptable way.

The lack of feedback from a person to a robot as well as the robot's failure to recognize that feedback limits the robot's assessment of what is happening in the interaction. The studies undertaken so far provide only initial observations about how people engage with robots and how the interaction changes, as the robot knows more about engagement behaviors. However, these first steps indicate that human-machine conversations must include and account for these behaviors if robots are to function fully in conversation. Without the ability to perceive human engagement signals, robots will misinterpret what is occurring in the interaction. Without producing proper engagement signals, robots may misinform human partners about their role in the interaction.

6.7 Future directions

The research reported in this article addresses the maintenance of engagement during an interaction. Equally interesting to us is the process by which two interactors find each other and begin to interact. Research on “catching the eye of another person” (Miyauchi et al, 2004) illustrates ways that robots can undertake one part of this process. Much remains to be done to understand this phenomenon and provide algorithms for the robot to perform such activities.

As was mentioned earlier in this article, body stance serves to indicate one’s focus of attention. When a conversational participant stands so that his or her body is pointed away from the conversational partner, there is an apparent conflict between engaging the partner and focusing on something in the direction of the body. How this conflict is managed and how body stance is combined with looking and gazing are not well-understood matters. Yet for many activities one’s body must be pointed in another direction than the one of the conversational participant. We plan to investigate the issue of body stance as an attentional mechanism and its relation to engagement and to looking and gazing.

Acknowledgements

This work has been accomplished in collaboration with a number of colleagues over the years who have worked with us on Mel at various stages of the research. Our thanks to Myrosia Dzikovska, Clifton Forlines, Cory Kidd, Neal Lesh, Max Makeev, Louis-Philippe Morency, Chuck Rich and Kate Saenko. We are indebted to them for their contributions to the development of Mel, and to the studies reported here.

This work was undertaken and completed while the authors were at Mitsubishi Electric Research Laboratories, Cambridge, MA, 02139, www.merl.com.

[1] The linguistic means for signaling that feedback is called for is not completely understood. However, a typical means of signaling feedback is to bring an utterance to a close and to use end of utterance prosodic effects to indicate that the speaker is expecting a response from the hearer.

This work is based on an earlier work: *The effect of head-nod recognition in human-robot conversation*, in Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction, Pages: 290 -296, © ACM, 2006.
http://portal.acm.org/citation.cfm?doid=1121241.1121291

References

[1] Argyle, M., Cook, M. Gaze and Mutual Gaze. Cambridge University Press, New York, 1976.
[2] Bickmore, T. Towards the Design of Multimodal Interfaces for Handheld Conversational Characters, Proceedings of the CHI Extended Abstracts on Human factors in Computing Systems Conference, Minneapolis, MN, ACM Press, 788 – 789, 2002.
[3] Breazeal, C. and Aryananda, L. Recognizing affective intent in robot directed speech, Autonomous Robots, 12:1, pp. 83-104, 2002.
[4] Breazeal, C., Brooks, A., Gray, J., Hoffman, G., Kidd, C., Lee, H., Lieberman, J., Lockerd, A., and Chilongo, D. Tutelage and collaboration for humanoid robots. International Journal of Humanoid Robotics, Vol. 1, No. 2 ,pp. 315-348, 2004.
[5] Cassell, J. Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents, in Cassell, J., Sullivan, J., Prevost, S., Churchill, E. (Eds.), Embodied Conversational Agents. MIT Press, Cambridge, MA, pp. 1-18, 2000.
[6] Cassell, J., Vilhjlmsson, H.H., Bickmore, T. BEAT: The behavior expression animation toolkit. In: Fiume, E. (Ed.), SIGGRAPH 2001, Computer Graphics Proceedings. ACM Press / ACM SIGGRAPH, pp. 477-486, 2001.
[7] Clark, H.H., Using Language. Cambridge University Press, Cambridge, 1996.
[8] Dietz, P.H., Leigh, D.L., Yerazunis, W.S. Wireless liquid level sensing for restaurant applications, IEEE Sensors 1, 715-720, 2002.
[9] Duncan, S. On the structure of speaker-auditor interaction during speaking turns, Language in Society, 3 (161-180), 1973.
[10] Goodwin, C. Gestures as a resource for the organization of mutual attention., Semiotica 62 (1/2), 29-49, 1986.
[11] Ishiguro, H., Ono, T., Imai, M., and Kanda, T. Development of an interactive humanoid robot "Robovie"---an interdisciplinary approach. In: Jarvis, R.A., Zelinsky, A. (Eds.), Robotics Research. Springer, pp. 179—191, 2003.
[12] Kendon, A. Some functions of gaze direction in two person interactions, Acta Psychologica 26, 22-63, 1967.
[13] Kozima, H., Nakagawa, C., and Yano, H. Attention coupling as a prerequisite for social interaction. In: Proceedings of the 2003 IEEE International Workshop on Robot and Human Interactive Communication. IEEE Press, New York, pp. 109-114, 2003.
[14] McClave, E.Z. Linguistic functions of head movements in the context of speech. Journal of Pragmatics 32, 855-878, 2000.
[15] McNeill, D. Hand and Mind: What Gestures Reveal about Thought. University of Chicago Press, 1992.
[16] Minato, T., MacDorman, K., Simada, M., Itakura, S., Lee, K. and Ishiguro, H. Evaluating humanlikeness by comparing responses elicited by an android and a person, Proceedings of the Second International Workshop on Man-Machine Symbiotic Systems, pp. 373-383, 2002.
[17] Miyauchi, D., Sakurai, A., Makamura, A., Kuno, Y. Active eye contact for human-robot communication. In: Proceedings of CHI 2004--Late Breaking Results. Vol. CD Disc 2. ACM Press, pp. 1099-1104, 2004.
[18] Morency, L.-P., Sidner, C., and Darrell, T. [this volume].

[19] Morency, L.-P., Lee, C., Sidner, C., and Darrell, T. Contextual recognition of head gestures, Proceedings of the Seventh International Conference on Multimodal Interfaces (ICMI'05), pp.18-24, 2005.

[20] L.-P. Morency, A. Rahimi, N. Checka, and T. Darrell. Fast stereo-based head tracking for interactive environment, Proceedings of the International Conference on Automatic Face and Gesture Recognition, pp. 375–380, 2002.

[21] Nakano, Y. and Nishida, T. Recognition and Generation of Nonverbal Communicative Cues in Conversational Agents [this volume].

[22] Nakano, Y., Reinstein, G., Stocky, T., and Cassell, J. Towards a model of face-to-face grounding, Proceedings of the 41st meeting of the Association for Computational Linguistics. Sapporo, Japan, pp. 553—561, 2003.

[23] Rich, C.; Sidner, C.L., and Lesh, N.B. COLLAGEN: Applying Collaborative Discourse Theory to Human-Computer Interaction, Artificial Intelligence Magazine, Winter, Volume 22, Issue 4, pp. 15-25, 2001.