Proceedings of the 41st Annual Meeting of the Association
for Computational Linguistics, July 2003, pp. 553-561.

Towards a Model of Face-to-Face Grounding

Yukiko I. Nakano[†/††] **Gabe Reinstein**[†] **Tom Stocky**[†] **Justine Cassell**[†]

[†]MIT Media Laboratory
E15-315
20 Ames Street
Cambridge, MA 02139 USA
{yukiko, gabe, tstocky, justine}@media.mit.edu

[††]Research Institute of Science and
Technology for Society (RISTEX)
2-5-1 Atago Minato-ku,
Tokyo 105-6218, Japan
nakano@kc.t.u-tokyo.ac.jp

Abstract

We investigate the verbal and nonverbal means for *grounding*, and propose a design for embodied conversational agents that relies on both kinds of signals to establish common ground in human-computer interaction. We analyzed eye gaze, head nods and attentional focus in the context of a direction-giving task. The distribution of nonverbal behaviors differed depending on the type of dialogue move being grounded, and the overall pattern reflected a *monitoring* of *lack of negative feedback*. Based on these results, we present an ECA that uses verbal and nonverbal grounding acts to update dialogue state.

1 Introduction

An essential part of conversation is to ensure that the other participants share an understanding of what has been said, and what is meant. The process of ensuring that understanding – adding what has been said to the common ground – is called grounding [1]. In face-to-face interaction, nonverbal signals as well as verbal participate in the grounding process, to indicate that an utterance is grounded, or that further work is needed to ground. Figure 1 shows an example of human face-to-face conversation. Even though no verbal feedback is provided, the speaker (S) continues to add to the directions. Intriguingly, the listener gives no explicit nonverbal feedback – no nods or gaze towards S. S, however, is clearly monitoring the listener's behavior, as we see by the fact that S looks at her twice (continuous lines above the words). In fact, our analyses show that maintaining focus of attention on the task (dash-dot lines underneath the words) is the listener's public signal of understanding S's utterance sufficiently for the task at hand. Because S is manifestly attending to this signal, the signal allows the two jointly to recognize S's contribution as grounded. This paper provides empirical support for an essential role for nonverbal behaviors in grounding, motivating an architecture for an embodied conversational agent that can establish common ground using eye gaze, head nods, and attentional focus.

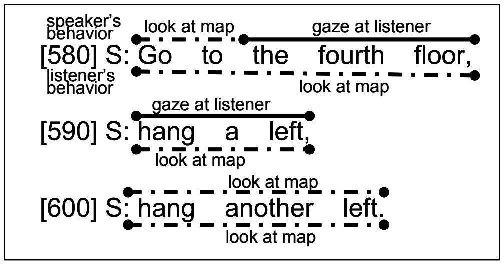


Figure 1: Human face-to-face conversation

Although grounding has received significant attention in the literature, previous work has not addressed the following questions: (1) what predictive factors account for how people use nonverbal signals to ground information, (2) how can a model of the face-to-face grounding process be used to adapt dialogue management to face-to-face conversation with an embodied conversational agent. This paper addresses these issues, with the goal of contributing to the literature on discourse phenomena, and of building more advanced conversational humanoids that can engage in human conversational protocols.

In the next section, we discuss relevant previous work, report results from our own empirical study and, based on our analysis of conversational data, propose a model of grounding using both verbal and nonverbal information, and present our implementation of that model into an embodied conversational agent. As a preliminary evaluation, we compare a user interacting with the embodied conversational agent with and without grounding.

2 Related Work

Conversation can be seen as a collaborative activity to accomplish information-sharing and to pursue joint goals and tasks. Under this view, agreeing on what has been said, and what is meant, is crucial to conversation. The part of what has been said that the interlocutors understand to be mutually shared is called the *common ground*, and the process of establishing parts of the conversation as shared is called *grounding* [1]. As [2] point out, participants in a conversation attempt to minimize the effort expended in grounding. Thus, interlocutors do not always convey all the information at their disposal; sometimes it takes less effort to produce an incomplete utterance that can be repaired if needs be.

[3] has proposed a computational approach to grounding where the status of contributions as provisional or shared is part of the dialogue system's representation of the "information state" of the conversation. Conversational actions can trigger updates that register provisional information as shared. These actions achieve grounding. Acknowledgment acts are directly associated with grounding updates while other utterances effect grounding updates indirectly, because they proceed with the task in a way that presupposes that prior utterances are uncontroversial.

[4], on the other hand, suggest that actions in conversation give probabilistic evidence of understanding, which is represented on a par with other uncertainties in the dialogue system (e.g., speech recognizer unreliability). The dialogue manager assumes that content is grounded as long as it judges the risk of misunderstanding as acceptable.

[1, 5] mention that eye gaze is the most basic form of positive evidence that the addressee is attending to the speaker, and that head nods have a similar function to verbal acknowledgements. They suggest that nonverbal behaviors mainly contribute to lower levels of grounding, to signify that interlocutors have access to each other's communicative actions, and are attending. With a similar goal of broadening the notion of communicative action beyond the spoken word, [6] examine other kinds of multimodal grounding behaviors, such as posting information on a whiteboard. Although these and other researchers have suggested that nonverbal behaviors undoubtedly play a role in grounding, previous literature does not characterize their precise role with respect to dialogue state.

On the other hand, a number of studies on these particular nonverbal behaviors do exist. An early study, [7], reported that conversation involves eye gaze about 60% of the time. Speakers look up at grammatical pauses for feedback on how utterances are being received, and also look at the task. Listeners look at speakers to follow their direction of gaze. In fact, [8] claimed speakers will pause and restart until they obtain the listener's gaze. [9] found that during conversational difficulties, mutual gaze was held longer at turn boundaries.

Previous work on embodied conversational agents (ECAs) has demonstrated that it is possible to implement face-to-face conversational protocols in human-computer interaction, and that correct relationships among verbal and nonverbal signals enhances the naturalness and effectiveness of embodied dialogue systems [10], [11]. [12] reported that users felt the agent to be more helpful, lifelike, and smooth in its interaction style when it demonstrated nonverbal conversational behaviors.

3 Empirical Study

In order to get an empirical basis for modeling face-to-face grounding, and implementing an ECA, we analyzed conversational data in two conditions.

3.1 Experiment Design

Based on previous direction-giving tasks, students from two different universities gave directions to campus locations to one another. Each pair had a conversation in a **(1) Face-to-face condition (F2F):** where two subjects sat with a map drawn by the direction-giver sitting between them, and in a **(2) Shared Reference condition (SR):** where an L-shaped screen between the subjects let them share a map drawn by the direction-giver, but not to see the other's face or body.

Interactions between the subjects were video-recorded from four different angles, and combined by a video mixer into synchronized video clips.

3.2 Data Coding

10 experiment sessions resulted in 10 dialogues per condition (20 in total), transcribed as follows.

Coding verbal behaviors: As grounding occurs within a turn, which consists of consecutive

Combinations of NVs		Listener's behavior			
		gP	gM	gMwN	gE
Speaker's behavior	gP	gP/gP	gP/gM	gP/gMwN	gP/gE
	gM	gM/gP	gM/gM	gM/gMwN	gM/gE
	gMwN	gMwN/gP	gMwN/gM	gMwN/gMwN	gMwN/gE
	gE	gE/gP	gE/gM	gE/gMwN	gE/gE

Table 1: NV statuses

	Shift to	
	within UU	pause
Acknowledgement	*gMwN/gM* (0.495)	*gM/gM* (0.888)
Answer	*gP/gP* (0.436)	*gM/gM* (0.667)
Info-req	*gP/gM* (0.38)	*gP/gP* (0.5)
Assertion	*gP/gM* (0.317)	*gM/gM* (0.418)

Table 2: Salient transitions

utterances by a speaker, following [13] we tokenized a turn into utterance units (UU), corresponding to a single intonational phrase [14]. Each UU was categorized using the DAMSL coding scheme [15]. In the statistical analysis, we concentrated on the following four categories with regular occurrence in our data: *Acknowledgement*, *Answer*, *Information request (Info-req)*, and *Assertion*.

Coding nonverbal behaviors: Based on previous studies, four types of behaviors were coded:

Gaze At Partner (gP): Looking at the partner's eyes, eye region, or face.

Gaze At Map (gM): Looking at the map

Gaze Elsewhere (gE): Looking away elsewhere

Head nod (Nod): Head moves up and down in a single continuous movement on a vertical axis, but eyes do not go above the horizontal axis.

By combining Gaze and Nod, six complex categories (ex. *gP* with nod, *gP* without nod, etc) are generated. In what follows, however, we analyze only categories with more than 10 instances. In order to analyze dyadic behavior, 16 combinations of the nonverbal behaviors are defined, as shown in Table 1. Thus, *gP/gM* stands for a combination of speaker gaze at partner and listener gaze at map.

Results

We examine differences between the **F2F** and **SR** conditions, correlate verbal and nonverbal behaviors within those conditions, and finally look at correlations between speaker and listener behavior.

Basic Statistics: The analyzed corpus consists of 1088 UUs for **F2F**, and 1145 UUs for **SR**. The mean length of conversations in **F2F** is 3.24 minutes, and in **SR** is 3.78 minutes (t(7)=-1.667 p<.07 (one-tail)). The mean length of utterances in **F2F** (5.26 words per UU) is significantly longer than in **SR** (4.43 words per UU) (t(7)=3.389 p< .01 (one-tail)). For the nonverbal behaviors, the number of shifts between the statuses in Table 1 was compared (eg. NV status shifts from *gP/gP* to *gM/gM* is counted as one shift). There were 887 NV status shifts for **F2F**, and 425 shifts for **SR**. The number of NV status shifts in **SR** is less than half of that in **F2F** (t(7)=3.377 p< .01 (one-tail)).

These results indicate that visual access to the interlocutor's body affects the conversation, suggesting that these nonverbal behaviors are used as communicative signals. In **SR,** where the mean length of UU is shorter, speakers present information in smaller chunks than in **F2F**, leading to more chunks and a slightly longer conversation. In **F2F**, on the other hand, conversational participants convey more information in each UU.

Correlation between verbal and nonverbal behaviors: We analyzed NV status shifts with respect to the type of verbal communicative action and the experimental condition (**F2F**/**SR**). To look at the continuity of NV status, we also analyzed the amount of time spent in each NV status. For gaze, transition and time spent gave similar results; since head nods are so brief, however, we discuss the data in terms of transitions. Table 2 shows the most frequent target NV status (shift *to* these statuses from others) for each speech act type in **F2F**. Numbers in parentheses indicates the proportion to the total number of transitions.

<Acknowledgement> Within an UU, the dyad's NV status most frequently shifts to *gMwN/gM* (eg. speaker utters "OK" while nodding, and listener looks at the map). At pauses, a shift to *gMgM* is most frequent. The same results were found in **SR** where the listener could not see the speaker's nod. These findings suggest that *Acknowledgement* is likely to be accompanied by a head nod, and this behavior may function introspectively, as well as communicatively.

<Answer> In **F2F**, the most frequent shift within a UU is to *gP/gP*. This suggests that speakers and listeners rely on mutual gaze (*gP/gP*) to ensure an answer is grounded, whereas they cannot use this strategy in **SR**. In addition, we found that speakers frequently look away at the beginning of an answer, as they plan their reply [7].

<Info-req> In **F2F**, the most frequent shift within a UU is to *gP/gM*, while at pauses between UUs shift to *gP/gP* is the most frequent. This suggests that speakers obtain mutual gaze after asking a question to ensure that the question is clear, before the turn is transferred to the listener to reply. In **SR**, however, rarely is there any NV status shift, and participants continue looking at the map.

<Assertion> In both conditions, listeners look at the map most of the time, and sometimes nod. However, speakers' nonverbal behavior is very different across conditions. In **SR**, speakers either look at the map or elsewhere. By contrast, in **F2F**, they frequently look at the listener, so that a shift to *gP/gM* is the most frequent within an UU. This suggests that, in **F2F**, speakers check whether the listener is paying attention to the referent mentioned in the *Assertion*. This implies that not only listener's gazing at the speaker, but also paying attention to a referent works as positive evidence of understanding in **F2F**.

In summary, it is already known that eye gaze can signal a turn-taking request [16], but turn-taking cannot account for all our results. Gaze direction changes within as well as between UUs, and the usage of these nonverbal behaviors differs depending on the type of conversational action. Note that subjects rarely demonstrated communication failures, implying that these nonverbal behaviors represent positive evidence of grounding.

Correlation between speaker and listener behavior: Thus far we have demonstrated a difference in distribution among nonverbal behaviors, with respect to conversational action, and visibility of interlocutor. But, to uncover the *function* of these nonverbal signals, we must examine how listener's nonverbal behavior affects the speaker's following action. Thus, we looked at two consecutive *Assertion* UUs by a direction-giver, and analyzed the relationship between the NV status of the first UU and the direction-giving strategy in the second UU. The giver's second UU is classified as *go-ahead* if it gives the next leg of the directions, or as *elaboration* if it gives additional information about the first UU, as in the following example:

```
[U1]S: And then, you'll go
   down this little corridor.
[U2]S: It's not very long.
```

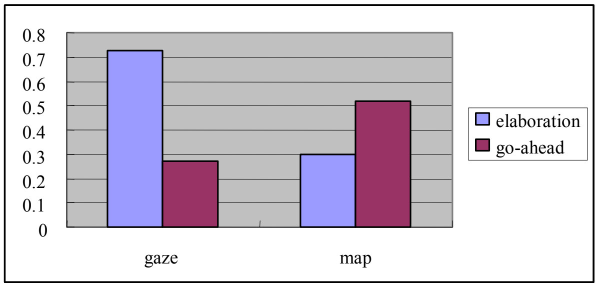


Figure 2: Relationship between receiver's NV and giver's next verbal behavior

Results are shown in Figure 2. When the listener begins to gaze at the speaker somewhere within an UU, and maintains gaze until the pause after the UU, the speaker's next UU is an *elaboration* of the previous UU 73% of the time. On the other hand, when the listener keeps looking at the map during an UU, only 30% of the next UU is an *elaboration* (z = 3.678, p<.01). Moreover, when a listener keeps looking at the speaker, the speaker's next UU is go-ahead only 27% of the time. In contrast, when a listener keeps looking at the map, the speaker's next UU is go-ahead 52% of the time (z = -2.049, p<.05)[1]. These results suggest that speakers interpret listeners' continuous gaze as evidence of not-understanding, and they therefore add more information about the previous UU. Similar findings were reported for a map task by [17] who suggested that, at times of communicative difficulty, interlocutors are more likely to utilize all the channels available to them. In terms of floor management, gazing at the partner is a signal of giving up a turn, and here this indicates that listeners are trying to elicit more information from the speaker. In addition, listeners' continuous attention to the map is interpreted as evidence of understanding, and speakers go ahead to the next leg of the direction[2].

3.3 A Model of Face-to-Face Grounding

Analyzing spoken dialogues, [18] reported that grounding behavior is more likely to occur at an

[1] The percentage for map does not sum to 100% because some of the UUs are cue phrases or tag questions which are part of the next leg of the direction, but do not convey content.

[2] We also analyzed two consecutive *Answer* UUs from a giver, and found that when the listener looks at the speaker at a pause, the speaker elaborates the Answer 78% of the time. When the listener looks at the speaker during the UU and at the map after the UU (positive evidence), the speaker elaborates only 17% of the time.

intonational boundary, which we use to identify UUs. This implies that multiple grounding behaviors can occur within a turn if it consists of multiple UUs. However, in previous models, information is grounded only when a listener returns verbal feedback, and acknowledgement marks the smallest scope of grounding. If we apply this model to the example in Figure 1, none of the UU have been grounded because the listener has not returned any spoken grounding clues.

In contrast, our results suggest that considering the role of nonverbal behavior, especially eye-gaze, allows a more fine-grained model of grounding, employing the UU as a unit of grounding.

Our results also suggest that speakers are actively monitoring positive evidence of understanding, and also the absence of negative evidence of understanding (that is, signs of miscommunication). When listeners continue to gaze at the task, speakers continue on to the next leg of directions.

Because of the incremental nature of grounding, we implement nonverbal grounding functionality into an embodied conversational agent using a process model that describes steps for a system to judge whether a user understands system contribution: (1) Preparing for the next UU: according to the speech act type of the next UU, nonverbal positive or negative evidence that the agent expects to receive are specified. (2) Monitoring: monitors and checks the user's nonverbal status and signals during the UU. After speaking, the agent continues monitoring until s/he gets enough evidence of understanding or not-understanding represented by user's nonverbal status and signals.(3) Judging: once the agent gets enough evidence, s/he tries to judge groundedness as soon as possible. According to some previous studies, length of pause between UUs is in between 0.4 to 1 sec [18, 19]. Thus, time out for judgment is 1 sec after the end of the UU. If the agent does not have evidence then, the UU remains ungrounded.

This model is based on the information state approach [3], with update rules that revise the state of the conversation based on the inputs the system receives. In our case, however, the inputs are sampled continuously, include the nonverbal state, and only some require updates. Other inputs indicate that the last utterance is still pending, and allow the agent to wait further. In particular, task attention over an interval following the utterance triggers grounding. Gaze in the interval means that the contribution stays provisional, and triggers an obligation to elaborate. Likewise, if the system times-out without recognizing any user feedback, the segment remains ungrounded. This process allows the system to keep talking across multiple utterance units without getting verbal feedback from the user. From the user's perspective, explicit acknowledgement is not necessary, and minimal cost is involved in eliciting elaboration.

4 Face-to-face Grounding with ECAs

Based on our empirical results, we propose a dialogue manager that can handle nonverbal input to the grounding process, and we implement the mechanism in an embodied conversational agent.

4.1 System

MACK is an interactive public information ECA kiosk. His current knowledgebase concerns the activities of the MIT Media Lab; he can answer questions about the lab's research groups, projects, and demos, and give directions to each.

On the input side, MACK recognizes three modalities: (1) speech, using IBM's ViaVoice, (2) pen gesture via a paper map atop a table with an embedded Wacom tablet, and (3) head nod and eye gaze via a stereo-camera-based 6-degree-of-freedom head-pose tracker (based on [20]). These inputs operate as parallel threads, allowing the Understanding Module (UM) to interpret the multiple modalities both individually and in combination.

MACK produces multimodal output as well: (1) speech synthesis using the Microsoft Whistler Text-to-Speech (TTS) API, (2) a graphical figure with synchronized hand and arm gestures, and head and eye movements, and (3) LCD projector highlighting on the paper map, allowing MACK to reference it.

The system architecture is shown in Figure 3. The UM interprets the input modalities and converts them to dialogue moves which it then passes on to the Dialogue Manager (DM). The DM consists of two primary sub-modules, the Response Planner, which determines MACK's next action(s) and creates a sequence of utterance units, and the Grounding Module (GrM), which updates the Discourse Model and decides when the Response Planner's next UU should be passed on to the Generation module (GM). The GM converts the UU into speech, gesture, and projector output, sending

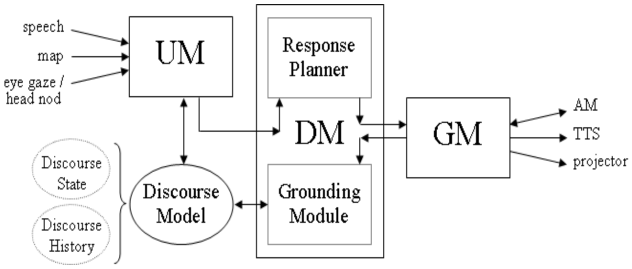


Figure 3: MACK system architecture

these synchronized modalities to the TTS engine, Animation Module (AM), and Projector Module.

The Discourse Model maintains information about the state and history of the discourse. This includes a list of grounded beliefs and ungrounded UUs; a history of previous UUs with timing information; a history of nonverbal information (divided into gaze states and head nods) organized by timestamp; and information about the state of the dialogue, such as the current UU under consideration, and when it started and ended.

4.2 Nonverbal Inputs

Eye gaze and head nod inputs are recognized by a head tracker, which calculates rotations and translations in three dimensions based on visual and depth information taken from two cameras [20]. The calculated head pose is translated into "look at MACK," "look at map," or "look elsewhere." The rotation of the head is translated into head nods, using a modified version of [21]. Head nod and eye gaze events are timestamped and logged within the nonverbal component of the Discourse History. The Grounding Module can thus look up the appropriate nonverbal information to judge a UU.

4.3 The Dialogue Manager

In a kiosk ECA, the system needs to ensure that the user understands the information provided by the agent. For this reason, we concentrated on implementing a grounding mechanism for *Assertion,* when the agent gives the user directions, and *An swer,* when the agent answers the user's questions

Generating the Response

The first job of the DM is to plan the response to a user's query. When a user asks for directions, the DM receives an event from the UM stating this intention. The Response Planner in the DM, recognizing the user's direction-request, calculates the directions, broken up into segments. These segments are added to the DM's Agenda, the stack of UUs to be processed.

At this point, the GrM sends the first UU (a direction segment) on the Agenda to the GM to be processed. The GM converts the UU into speech and animation commands. For MACK's own nonverbal grounding acts, the GM determines MACK's gaze behavior according to the type of UU. For example, when MACK generates a direction segment (an Assertion), 66% of the time he keeps looking at the map. When elaborating a previous UU, 47% of the time he gazes at the user.

When the GM begins to process the UU, it logs the start time in the Discourse Model, and when it finishes processing (as it sends the final command to the animation module), it logs the end time. The GrM waits for this speech and animation to end (by polling the Discourse Model until the end time is available), at which point it retrieves the timing data for the UU, in the form of timestamps for the UU start and finish. This timing data is used to look up the nonverbal behavior co-occurring with the utterance in order to judge whether or not the UU was grounded.

Judgment of grounding

When MACK finishes uttering a UU, the Grounding Module judges whether or not the UU is grounded, based on the user's verbal and nonverbal behaviors during and after the UU.

Using verbal evidence: If the user returns an acknowledgement, such as *"OK"*, the GrM judges the UU grounded. If the user explicitly reports failure in perceiving MACK's speech (ex. *"what?"*), or not-understanding (ex. *"I don't understand"*), the UU remains ungrounded. Note that, for the moment, verbal evidence is considered stronger than nonverbal evidence.

Using nonverbal evidence: The GrM looks up the nonverbal behavior occurring during the utterance, and compares it to the model shown in Table 3. For each type of speech act, this model specifies the nonverbal behaviors that signal positive or explicit negative evidence. First, the GrM compares the within-UU nonverbal behavior to the model. Then, it looks at the first nonverbal behavior occurring during the pause after the UU. If these two behaviors ("within" and "pause") match a pattern that signals positive evidence, the UU is grounded. If they match a pattern for negative evidence, the UU is not grounded. If no pattern has yet been

Target UU Type	Evidence Type	NV Pattern	Judgment of ground	Suggested next action
Assertion	positive	within: map pause: map /nod	grounded	go-ahead: 0.7 elaboration: 0.30
	negative	within: gaze pause: gaze	ungrounded	go-ahead: 0.27 elaboration:0.73
Answer	positive	within: gaze pause: map	grounded	go-ahead: 0.83 elaboration: 0.17
	negative	pause: gaze	ungrounded	go-ahead: 0.22 elaboration: 0.78

Table 3: Grounding Model for MACK

matched, the GrM waits for a tenth of a second and checks again. If the required behavior has occurred during this time, the UU is judged. If not, the GrM continues looping in this manner until the UU is either grounded or ungrounded explicitly, or a 1 second threshold has been reached. If the threshold is reached without a decision, the GrM times out and judges the UU ungrounded.

Updating the Dialogue State

After judging grounding, the GrM updates the Discourse Model. The Discourse State maintained in the Discourse Model is similar to TRINDI kit [3], except that we store nonverbal information. There are three key fields: (1) a list of grounded UUs, (2) a list of pending (ungrounded) UUs, and (3) the current UU. If the current UU is judged grounded, its belief is added to (1). If ungrounded, the UU is stored in (2). If an UU has subsequent contributions such as elaboration, these are stored in a single discourse unit, and grounded together when the last UU is grounded.

Determining the Next Action

After judging the UU's grounding, the GrM decides what MACK does next. (1) MACK can continue giving the directions as normal, by sending on the next segment in the Agenda to the GM. As shown in Table 3, this happens 70% of the time when the UU is grounded, and only 27% of the time when it is not grounded. Note, this happens 100% of the time if verbal acknowledgement (e.g. “Uh huh”) is received for the UU.

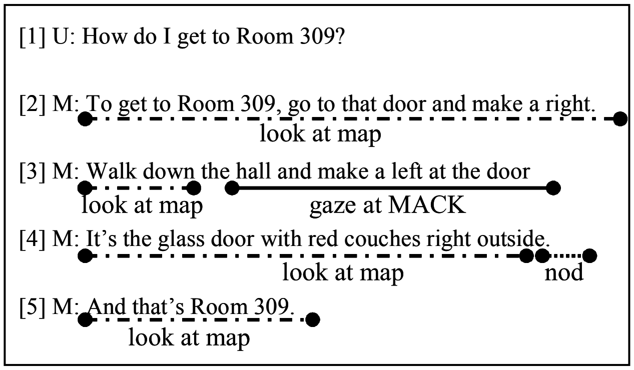
(2) MACK can elaborate on the most recent stage of the directions. Elaborations are generated 73% of the time when an Assertion is judged ungrounded, and 78% of the time for an ungrounded Answer. MACK elaborates by describing the most recent landmark in more detail. For example, if the directions were “Go down the hall and make a right at the door,” he might elaborate by saying “The big blue door.” In this case, the GrM asks the Response Planner (RP) to provide an elaboration for the current UU; the RP generates this elaboration (looking up the landmark in the database) and adds it to the front of the Agenda; and the GrM sends this new UU on to the GM.

Finally, if the user gives MACK explicit verbal evience of not understanding, MACK will simply repeat the last thing he said, by sending the UU back to the GM.

4.4 Example

Figure 4 shows an example of a user's interaction with MACK. The user asks MACK for directions, and MACK replies using speech and pointing (using a projector) to the shared map.

When the GrM sends the first segment in the Agenda to the GM,the starting time of the UU is noted and it is sent to the AM to be spoken and animated. During this time, the user's nonverbal signals are logged in the Discourse Model. When the UU has finished, the GrM evaluates the log of the UU and of the very beginning of the pause (by waiting a tenth of a second and then checking the nonverbal history). In this case, MACK noted that the user looked at the map during the UU, and continued to do so just afterwards. This pattern matches the model for Assertion. The UU is judged as grounded, and the grounded belief is added to the Discourse Model.



[1] U: How do I get to Room 309?

[2] M: To get to Room 309, go to that door and make a right.

[3] M: Walk down the hall and make a left at the door

[4] M: It's the glass door with red couches right outside.

[5] M: And that's Room 309.

Figure 4: Example of user (U) interacting with MACK (M). User gives negative evidence of grounding in [3], so MACK elaborates [4].

MACK then utters the second segment as before, but this time the GrM, finds that the user was looking up at MACK during most of the UU as well as after it, which signals that the UU is not grounded. Therefore, the RP generates an elaboration (line 4). This utterance is judged to be

Figure 5: MACK with user

grounded both because the user continues looking at the map, and because the user nods, and so the final stage of the directions is spoken. This is also grounded, leaving MACK ready for a new inquiry.

5 Preliminary Evaluation

Although we have shown an empirical basis for our implementation, it is important to ensure both that human users interact with MACK as we expect, and that their interaction is more effective than without nonverbal grounding. The issue of effectiveness merits a full-scale study and thus we have chosen to concentrate here on whether MACK elicits the same behaviors from users as does interaction with other humans.

Two subjects were therefore assigned to one of the following two conditions, both of which were run as Wizard of Oz (that is, "speech recognition" was carried out by an experimenter):

(a) MACK-with-grounding: MACK recognized user's nonverbal signals for grounding, and displayed his nonverbal signals as a speaker.

(b) MACK-without-grounding: MACK paid no attention to the user's nonverbal behavior, and did not display nonverbal signals as a speaker. He gave the directions in one single turn.

Subjects were instructed to ask for directions to two places, and were told that they would have to lead the experimenters to those locations to test their comprehension. We analyzed the second direction-giving interaction, after subjects became accustomed to the system.

Results: In neither condition, did users return verbal feedback during MACK's direction giving. As shown in Table 4, in MACK-with-grounding 7 nonverbal status transitions were observed during his direction giving, which consisted of 5 Assertion UUs, one of them an elaboration. The transition patterns between MACK and the user when MACK used nonverbal grounding are strikingly similar to those in our empirical study of human-to-human communication. There were three transitions to *gM/gM* (both look at the map), which is a normal status in map task conversation, and two transitions to *gP/gM* (MACK looks at the user, and the user looks at the map), which is the most frequent transition in Assertion as reported in Section 3. Moreover, in MACK's third UU, the user began looking at MACK at the middle of the UU and kept looking at him after the UU ended. This behavior successfully elicited MACK's elaboration in the next UU.

		with-grounding	w/o-grounding
num of UUs		5	4
Shift to	*gMgM*	3	2
	gPgM	2	0
	gMgP	1	0
	gPgP	1	0
	gMgMwN	0	1
	total	7	3

Table 4: Preliminary evaluation

On the other hand, in the MACK-without-grounding condition, the user never looked at MACK, and nodded only once, early on. As shown in Table 4, only three transitions were observed (shift to *gMgM* at the beginning of the interaction, shift to *gMgMwN*, then back to *gMgM*).

While a larger scale evaluation with quantitative data is one of the most important issues for future work, the results of this preliminary study strongly support our model, and show MACK's potential for interacting with a human user using human-human conversational protocols.

6 Discussion and Future Work

We have reported how people use nonverbal signals in the process of grounding. We found that nonverbal signals that are recognized as positive evidence of understanding are different depending on the type of speech act. We also found that maintaining gaze on the speaker is interpreted as evidence of not-understanding, evoking an additional explanation from the speaker. Based on these empirical results, we proposed a model of nonverbal grounding and implemented it in an embodied conversational agent.

One of the most important future directions is to establish a more comprehensive model of face-to-face grounding. Our study focused on eye gaze and head nods, which directly contribute to grounding. It is also important to analyze other types of nonverbal behaviors and investigate how they interact with eye gaze and head nods to achieve common ground, as well as contradictions between verbal and nonverbal evidence (eg. an interlocutor says, “OK”, but looks at the partner).

Finally, the implementation proposed here is a simple one, and it is clear that a more sophisticated dialogue management strategy is warranted, and will allow us to deal with back-grounding, and other aspects of miscommunication. For example, it would be useful to distinguish different levels of miscommunication: a sound that may or may not be speech, an out-of-grammar utterance, or an utterance whose meaning is ambiguous. In order to deal with such uncertainty in grounding, incorporating a probabilistic approach [4] into our model of face-to-face grounding is an elegant possibility.

Acknowledgement

Thanks to Candy Sidner, Matthew Stone, and 3 anonymous reviewers for comments that improved the paper. Thanks to Prof. Nishida at Univ. of Tokyo for his support of the research.

References

1.Clark, H.H. and E.F. Schaefer, *Contributing to discourse.* Cognitive Science, 1989. **13,**: p. 259-294.

2.Clark, H.H. and D. Wilkes-Gibbs, *Referring as a collaborative process.* Cognition, 1986. **22**: p. 1-39.

3.Matheson, C., M. Poesio, and D. Traum. *Modelling Grounding and Discourse Obligations Using Update Rules*. in *1st Annual Meeting of the North American Association for Computational Linguistics (NAACL2000)*. 2000.

4.Paek, T. and E. Horvitz, *Uncertainty, Utility, and Misunderstanding*, in *Working Papers of the AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems*, S.E. Brennan, A. Giboin, and D. Traum, Editors. 1999, AAAI: Menlo Park, California. p. 85-92.

5.Clark, H.H., *Using Language*. 1996, Cambridge: Cambridge University Press.

6.Traum, D.R. and P. Dillenbourg. *Miscommunication in Multimodal Collaboration*. in *AAAI Workshop on Detecting, Repairing, and Preventing Human-Machine Miscommunication*. 1996. Portland, OR.

7.Argyle, M. and M. Cook, *Gaze and Mutual Gaze*. 1976, Cambridge: Cambridge University Press.

8.Goodwin, C., *Achieving Mutual Orientation at Turn Beginning*, in *Conversational Organization: Interaction between speakers and hearers*. 1981, Academic Press: New York. p. 55-89.

9.Novick, D.G., B. Hansen, and K. Ward. *Coordinating turn-taking with gaze*. in *ICSLP-96*. 1996. Philadelphia, PA.

10.Cassell, J., et al. *More Than Just a Pretty Face: Affordances of Embodiment*. in *IUI 2000*. 2000. New Orleans, Louisiana.

11.Traum, D. and J. Rickel. *Embodied Agents for Multiparty Dialogue in Immersive Virtual Worlds*. in *Autonomous Agents and Multi-Agent Systems*. 2002.

12.Cassell, J. and K.R. Thorisson, *The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents.* Applied Artificial Intelligence, 1999. **13**: p. 519-538.

13.Nakatani, C. and D. Traum, *Coding discourse structure in dialogue (version 1.0)*. 1999, University of Maryland.

14.Pierrehumbert, J.B., *The phonology and phonetics of english intonation*. 1980, Massachusetts Institute of Technology.

15.Allen, J. and M. Core, *Draft of DMSL: Dialogue Act Markup in Several Layers*. 1997, http://www.cs.rochester.edu/research/cisd/resources/damsl/RevisedManual/RevisedManual.html.

16.Duncan, S., *On the structure of speaker-auditor interaction during speaking turns.* Language in Society, 1974. **3**: p. 161-180.

17.Boyle, E., A. Anderson, and A. Newlands, *The Effects of Visibility in a Cooperative Problem Solving Task.* Language and Speech, 1994. **37**(1): p. 1-20.

18.Traum, D. and P. Heeman. *Utterance Units and Grounding in Spoken Dialogue*. in *ICSLP*. 1996.

19.Nakajima, S.y. and J.F. Allen. *Prosody as a cue for discourse structure*. in *ICSLP*. 1992.

20.Morency, L.P., A. Rahimi, and T. Darrell. *A View-Based Appearance Model for 6 DOF Tracking," Proceed-ings of.* in *IEEE conference on Computer Vision and Pattern Recognition*. 2003. Madison, Wisconsin.

21.Kapoor, A. and R.W. Picard. *A Real-Time Head Nod and Shake Detector*. in *Workshop on Perceptive User Interfaces*. 2001. Orlando FL.