Why and how to model multi-modal interaction for a mobile robot companion

Shuyin Li and Britta Wrede
Applied Computer Science, Faculty of Technology
Bielefeld University,
33594 Bielefeld, Germany
Email: shuyinli, bwrede@techfak.uni-bielefeld.de

Abstract

Verbal and non-verbal interaction capabilities for robots are often studied isolated from each other in current research trend because they largely contribute to different aspects of interaction. For a robot companion that needs to be both useful and social, however, these capabilities have to be considered in a unified, complex interaction context. In this paper we present two case studies in such a context that clearly reveal the strengths and limitations of these modalities and advocate their complementary benefits for human-robot interaction. Motivated by this evidence we propose a powerful interaction framework which addresses common features of interactional and propositional information instead of their differences, as popular in many other works in this field, and models them using one single principle: grounding.

Introduction

Humanoid robots are viewed as unique platforms for developing human-like capabilities for machines. Based on their appearance and their hardware equipment these robots are likely to exhibit believable, human-like behaviors which facilitate a smooth and intuitive interaction with human users. At present, two main research strands can be observed in the field of interaction study. The first strand emphasizes the emotional and social expressivity of *non-verbal behaviors* and targets at the modeling of human-like behaviors for robots whose sole purpose is to engage human users in social interaction (Breazeal 2002) (Kanda *et al.* 2002). Their evaluation focus is often how enjoyable or human-like the interaction is for the user. The second strand concerns the development of *conversational* systems for robots (Matsui *et al.* 1999), (Lemon *et al.* 2001), that provide services for human users and communicate with them mainly using speech. In their evaluations mainly the efficiency and success ratio of the (mostly) task-related interaction are tested. Although the issues of these two strands contribute both to the human-robot interaction (HRI) research they are largely studied in an isolated manner.

The focus of our work is to develop an interaction framework for our robot companion BIRON (The Bielefeld Robot Companion). A robot companion is a robot that should be able to both accomplish useful tasks and behave socially (Dautenhahn *et al.* 2005). It is thus essential for us to take into account both task-related and social aspects of interaction which requires a combination of the two research strands. To achieve this goal we first have to clarify the advantages and disadvantages of verbal and non-verbal modalities in a complex interaction context so that we can use the appropriate modality for different purposes. Researchers in the first strand tend to argue for the general importance of non-verbal behaviors. For example in case of facial expressions, it is argued in (Breazeal 2002) that human-like expressions are easy to understand for human users. The author of (Billard 2005) holds the view that a face can give a user clues as to the ability of the robot. For example, a two-year-old face implies less cognitive capability than an adult face. In contrast, researchers in the second strand emphasize more the importance of speech which is, almost indisputable, the main channel to communicate task-related information between the user and the robot. In our view, the issue of strengths and limitations of modalities should not be discussed in an isolated context. Instead, we should draw conclusions from real, complex human-robot interaction where both task-related and social aspects are manifested. More specifically, the robot can be configured to communicate these aspects *only* in either verbal or non-verbal modality because this may result in an suboptimal interaction which reveals limitations of the involved modality and also possible complementary advantages of other modalities. In this paper we report our observations made during a user study with BIRON which, at the time of the study, could only generate verbal feedback. From these observations we draw conclusions clearly in favor of developing additional non-verbal feedback capabilities for BIRON because of benefits that can not be derived only by speech.

Having realized the importance of both verbal and non-verbal robot behaviors in HRI we need an interaction framework that can handle these behaviors in a flexible way. Researchers in the field of virtual agents have been developing such systems since decades and have proposed many sophisticated approaches. Cassell (Cassell 2000) focuses on the optimal animation of synchronized multi-modal behavior and proposed a generic architecture for virtual agents. This architecture represents the central idea of her FMTB model: “multiple (*interactional* and *propositional*) commu-

nicative goals are conveyed by conversational *functions* that are expressed by conversational *behaviors* in one or several modalities" (Cassell *et al.* 2000). In this model, the interactional goals regulate the state of the conversation, e.g., establishing contact with the user or releasing turn, while the propositional goals are driven by the needs of discourse. In the architecture, interactional and propositional information that is communicated during the interaction is processed by two different modules. The processing results of these two modules are conversational functions that are subsequently converted into different conversational behaviors by another module. Traum (Traum & Rickel 2002) proposed a multi-modal conversational model based on the *information state theory* (Larsson & Traum 2000) which is a powerful mechanism to model complex conversation. His model consists of 5 main layers representing different aspects of a conversation: contact, attention, conversation, obligation and negotiation. The state of these layers can be changed by dialog moves that the dialog participants perform in verbal or/and non-verbal form. As can be seen, both Cassell and Traum tackle the issue of multi-modality by grouping information involved in the conversation into categories (interactional and propositional information in Cassell's model and the layers in Traum's model) and then handle them separately. However, it is not always easy to make a clear distinction between these groups and one single interaction contribution often has to be analyzed in terms of its relevance in several groups, as will be discussed in more detail later. In this paper, we present a computational model of multi-modal grounding which does not differentiate distinct types or layers of the interaction but model them using one single principle: grounding. This model is motivated by the understanding that conversation is a *reciprocal* process in a variety of aspects that largely follow the principle of grounding.

The rest of the paper is organized as follows: section Observations reports the observations we made during a user study with BIRON and section A Multi-modal Interaction Framework" presents our grounding-based interaction framework. Finally, section Conclusion summarizes the necessity of incorporating verbal and non-verbal interaction capabilities into HRI and the strength of our approach.

Observations

Our robot BIRON is a personal robot with social learning abilities. It can detect and follow persons, focus on objects (according to detected deictic gestures of the human user), and store collected multi-modal information into a memory. Our implementation scenario is the so-called *home tour*: a user bought a new robot from a shop and shows it her home to prepare it for future tasks. For example, if the user says "This is my favorite cup." and points to it, the robot should be able to understand the user's speech, track her deictic gesture, detects the object that the user is pointing to and remember features of that object, i.e., name, color, images etc. With this knowledge, BIRON is able to use this favorite cup of the user to serve her a tea if she wishes. This scenario is also a learning scenario for the user because she interacts with a complex robot system for the first time. Lacking knowledge about its capabilities the user has to totally rely on the robot's ability to explain things itself.

The user study we conducted was originally intended to test various features of two new modules that we integrated into the system. Although the interaction system with the capability to handle input and output of different modalities was already integrated, the robot face, which was intended to be the main non-verbal feedback media, was not yet completed so that it was excluded from the study. We recruited 14 subjects aged from 20 to 37 from the Bielefeld University. None of the subjects had interacted with BIRON before. They were just told to try out BIRON's functionalities without any knowledge about BIRON's technical limitations. With this setting we intended to simulate the situation of the "naive user" in the home tour. In average, each subject interacted with BIRON more than 7 minutes. After the interaction the subjects were asked to fill out a questionnaire. The quantitative results of this study are discussed in (Li, Wrede, & Sagerer 2006b). For the purpose of this paper, we only focus on the observations that we made as a third person and report two cases that illustrate the limitation of the speech modality, which should be compensated by non-verbal modalities.

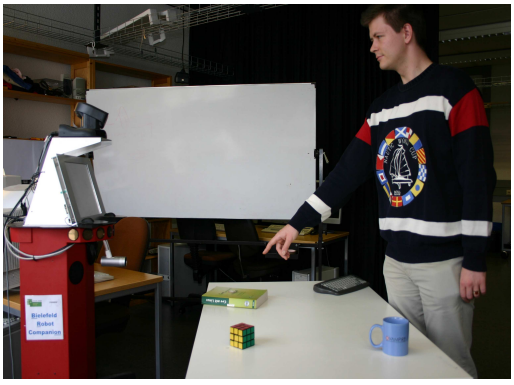

Figure 1: The setup of the user study

Case one: meta-commentators

The interaction system of BIRON is able to assess the performance of the robot in the runtime and take initiative (, which happens approximately every four user-robot-exchanges) to generate output to praise or comfort the user according to the assessment results. This feature was designed to show users some level of social awareness of BIRON which should motivate them and also make the interaction more enjoyable.

During the experiment we observed differences in interaction styles among the subjects which result in their different reactions to this feature of the system. Our scenario of home tour has an explorative nature for subjects because they do not have much pre-knowledge about the robot's capabilities, including its language capabilities. Therefore, most subjects started the interaction by saying things to see the reaction of BIRON which often caused out-of-vocabulary errors in speech recognition at the beginning of the interaction. The interaction system of BIRON then asked the user to rephrase

or try again and also, from time to time, comfort her with utterances like "I know it is sometimes difficult with me, but don't feel discouraged!". This behavior turned out to be an effective means to reduce frustration level for the majority of subjects (as revealed by the questionnaires (Li, Wrede, & Sagerer 2006b)). However, as we observed, it is also this behavior that caused more frustration in some other users. These were "meta-commentators" who tended to reply to such apologies by saying, e.g., "I'm already discouraged." or "Apology has no use." etc. These comments were most times out of vocabulary of the robot which caused even more speech recognition errors and apologies from the robot. The upshot of this spiral of robot speech was, from the perspective of the meta-commentators, it apologized for its performance far too often and the whole interaction entered into a dead-lock around the apologies.

Now we seem to be in a dilemma: the robot's comments on its own performance are basically a good idea but for some users they cause more trouble than generate benefits. Since we have to account for different users we need to find out what of the robot should be improved to solve this problem. We can identify two features of the speech modality, which we used to convey the social awareness, as possible reasons for the problem: reciprocity and obtrusiveness. Verbal conversation is a social act and is subject to social conventions. If one utters something the conversation partner usually has the obligation to reply (Clark 1992), often using the same modality. Since the robot issued the apologies verbally it is natural for the subjects to also reply verbally. For us, however, it is difficult to predict what should be included into the robot's vocabulary because the content of the reply is highly person-dependent. Additionally, speech is obtrusive in general because the listener is more or less "forced" to listen to it given the conversational convention of turn-taking. In human-machine interaction it is also often the case that the machine is not able to execute other tasks before the current issue is resolved. Thus, the meta-commentators had to repeatedly listen to the apologies and barging in caused even more speech recognition errors. The solution to the problem, therefore, lies in replacing the verbal apologies with non-verbal ones.

The strength of non-verbal feedback is its unobtrusiveness because users can decide themselves whether or not they pay attention to these behaviors. Besides, non-verbal behaviors generally do not impose so strong obligations to reply on the interaction partner as speech so that the potential for unexpected replies is reduced. Additionally, non-verbal feedback can be demonstrated in parallel to speech which enables a fluent task-oriented conversation without losing social capabilities of the robot. In our case what can help is, e.g., a robot face that is able to demonstrate apologizing facial expressions and grows more and more sorrowful while the interaction quality gets worse. At a certain point speech can also be involved to make the concern of the robot more explicit. Thus, the robot is able to regulate its social-aware behaviors according to the severity of the situation.

Case two: quiet speakers

As a *situated* computer device (Brooks 1989) a robot can not take it for granted that the user is permanently present and willing to interact with it. We therefore developed a Person Tracking and Attention system (PTA) for BIRON (Lang *et al.* 2003) that can detect humans and interpret their interaction intention by searching and analyzing the (combination of) perceived human legs, voice location and face direction. The speech recognizer is activated only if the PTA identifies the interaction intention of a human user, i.e., if the status of the the robot system is changed from "Person detected" to "Communication Partner (CP) detected". This policy makes sense because it prevents the robot from erroneously attempting to interact with the user when the user is talking to another person or it is just the radio on. For the interaction system of BIRON, however, this means, several requirements have to be met before the processing of user speech begins: (1) the person has a reasonable distance to the robot; (2) the person stands straight; (3) the person is looking at the robot's camera; (4) the person speaks clearly and reasonably loud and (5) the robot correctly receives and analyzes all the incoming signals. In the real system operations these requirements sometimes may not be fulfilled which has serious consequences on the interaction.

Our speech recognizer is trained with predominately male voices and, therefore, does not work particularly well with female voices. One of our female subjects spoke with very low voice in high pitch so that the speech recognizer most of the time interpreted her voice as noise and did not forward it for further processing. Puzzled by no reaction from the robot, the subject looked frequently in the direction of the experimentator and asked why the robot did not react. Knowing that her own voice could also influence the robot's perception of the environment the experimentator tried to use gesture to make clear that she could not intervene. The subject seemed not to be able to interpret the meaning of experimentator's gesture and, therefore, went a few steps towards the experimentator so that she was out of the range where the robot could perceive her as a human. Then the subject came back and tried again, in vain. In this whole process, although the robot went through different internal states (Person detected, CP detected, CP lost, Person lost, Person detected, CP detected), there was no visible reaction from the robot for the subject which was a very frustrating interaction experience for her.

The problem of this case is that we had no means to convey the internal states to the subject before the active interaction starts. With our only available modality speech we only had two possibilities. The first one is to make the robot report its internal state every time when it is changed. However, this can mislead users to believe that a connection to the robot is already established which is not true because these states only represent the robot's perception and a connection is only established when the robot receives explicit speech from the user. The second possibility can avoid this problem by giving users the impression that the speech output on the state changes is some routine output from the robot: it can repeatedly generate output like "I saw you, I saw you, I saw you...". This is obviously not really an option

with the speech modality, but a good solution if the robot had non-verbal feedback capabilities. Non-verbal behaviors can be effectively used to represent static information which is occasionally updated. In our case, these behaviors can help users to understand how the robot perceives a human and, if needed, to adjust her own behaviors, e.g., not walking out of the view of the robot.

Speech can convey complex semantic information and is therefore indispensable for both human users and robots to communicate in a task-related context. However, as our case studies show this modality reaches its limit when the robot also has to simultaneously demonstrate its internal awareness of human user's possible affective status and physical presence. Thus, for a robot companion that should be able to carry out useful tasks and behave socially, both verbal and non-verbal modalities have to be included into its interaction framework.

A multi-modal interaction framework

In this section, we propose an interaction framework that is able to handle multi-modal input and output in a flexible way. This framework is based on a computational model of multi-modal grounding (Li, Wrede, & Sagerer 2006a) and is designed for face-to-face human interaction with robot companions. The central idea of our approach is to represent individual interaction contributions as Interaction Unit and organize them based on the principle of grounding. In this section, we first explain why the principle of grounding, which is traditionally used to model spoken dialog can be used to model multi-modal interaction. Then we extend the original concept of grounding in two aspects: extending the definition of "common ground" and representing the most basic interactional contributions as Interaction Units. Afterwards, we depict the essentials of how to organize these Interaction Units based on the grounding principle and illustrate this mechanism with an interaction example. Finally, we will describe our implemented systems and discuss the benefits and deficiencies of this interaction framework.

Why grounding?

As mentioned in the Introduction, it is not always easy to relate information involved in a multi-modal interaction to certain types or layers as those proposed by Cassell (Cassell 2000) and Traum (Traum & Rickel 2002) , so that one single contribution from interaction participants often has to be analyzed several times with respect to their relevance as different types or in different layers. For example, in Cassell's system, the interactional processing mainly involves the processing of non-verbal signals and the propositional processing verbal signals. This devision of task may fail to handle situations when verbal and non-verbal signals co-carry a propositional meaning, e.g., if the user says "Show me *that* room." and point to it with a deictic gestures. To avoid this problem the system has to analyze each non-verbal signal in terms of whether it changes the interactional state or it contributes to a proposition. And this is similar in Traum's system, e.g., the utterance "Look up!" changes the state of the attention layer because it performs an action of "Direct-attention" on this layer. At the same time, it is also relevant on the conversation layer because it is clearly a part of the conversation and contributes to the state change of the propositional discourse.

Rather than addressing differences between functions of different interaction contributions, we look at them from another perspective: by looking for their common feature, which is, as we have identified, the *reciprocity*. Most interactional behaviors, whether they contribute to the regulation of the interaction itself or to the propositional discourse, have *evocative function*. This means, these behaviors result in some reactions from other interaction participants. If a listener A raises a hand indicating that she wants to speak, then the speaker B will usually stop speaking and release the turn. In this example, interactional information, as defined by Cassell, is sent by A and it has the function that B reacts to it by releasing turn. Similarly, if C asks a question, her interaction partner D will usually reply to it or at least indicate her hearing. Here, C generates some propositional information which has the function that D addresses it. Whenever the evocative function of these behaviors can not be realized, e.g., in the above examples, if B does not release the turn and D does not answer the question, then the initiator of these behaviors will probably perform other interactional behaviors so that these functions can be fulfilled. For example, in case of the over-active speaker, A can generate some stronger signals such as waving hands to attract more attention of B; in case of the silent listener, C can initiate a question "Are you listening to me?". Interaction participants do this probably because they feel that their interactional and propositional information is not perceived or understood.

This pattern of behavior sequence is similar to that of spoken dialog in its traditional sense, i.e., feedback is expected most of the time. A promising concept that attempts to explain this phenomenon is *grounding*. The notion of grounding states that during a dialog, agents strive to reach a certain degree of common ground, i.e., mutual understanding about the current intentions, goals and tasks (Clark 1992). For this purpose, the initiator of an account, the so-called Presentation, observes the reactions of her dialog partner who is supposed to provide the so-called Acceptance, i.e., evidence of understanding for this Presentation. The agents can only be sure that the common ground concerning a Presentation is established if its Acceptance is available. If the Acceptance can not be provided by the other agent, the creator of the Presentation will make more effort to help her, e.g., by providing more information or strengthening her signal. We have extended this concept to cover multi-modality of social interaction. More specifically, our extensions include (1) extending the definition of "common ground" and (2) introducing Interaction Unit as the most basic units of social interaction.

Extension (I): common ground

In the original concept of grounding, the common ground that is shared between dialog participants usually refers to mutual understanding of what has been said propositionally up to a certain point of the dialog. However, if we

take into account the entire process of multi-modal social interaction, more than just the propositional mutual understanding is shared between interaction participants. First of all, the pre-condition of an interaction has to be available. e.g., the (potential) participants have visual access to each another and both have the motivation to talk. At this stage (the contact layer in Traum's model), what they share is the physical possibility of contact and their willingness to interact. After these pre-conditions are fulfilled, the interaction is established and what the participants share now is their mutual understanding. In our term, mutual understanding does not only refer to the understanding of propositional information that is exchanged, but also interactional information, e.g., A's raising hand requests the turn. As a summary, in our model, the definition of common ground is extended to "what interaction participants share during an interaction" and it includes (1) physical possibility of contact and the participants' willingness to interact and (2) mutual understanding of the interactional and propositional information that is generated during the interaction.

The extended definition of common ground implies that interaction participants can not only provide Acceptance to what has been said propositionally, but also to other parts of the common ground. Taking the previous example, if A raises a hand, she generates a Presentation which is intended to establish the common ground with B that the turn is requested. If B releases the turn to A, she signals her Acceptance to A's Presentation and the common ground is established. However, if B does not release the turn, she does not provide Acceptance to the Presentation and A will make more effort so that B can help establish the common ground. This process is parallel to the case where C asks a question (creation of a Presentation) with the intention to establish the common ground concerning the question. If D replies in a satisfying way (Acceptance is available), the speaker will take the common ground as established and no more effort needs to be made concerning this question. If D does not reply or her answer is not satisfying, C will not take the reply as an Acceptance to her Presentation and the common ground can not be established.

Including the physical possibility of interaction into the definition of common ground implies that whether an interaction participant can provide Acceptance or not does not only depend on her ability to *understand* the Presentation, but also to *perceive* it. For example, if A can see B but B can not see A, the common ground, i.e., the physical possibility of an interaction, can not be established. Following the principle of grounding, if A intends to talk with B she would strengthening the signal of her presence, e.g., by shouting to B.

Extension (II): Interaction Unit

Presentation and Acceptance are both contributions of interaction participants and can be demonstrated in multi-modal way. We model these contributions as Interaction Unit (IU). An IU is a two-layered structure (Fig. 2) consisting of a *Motivation Layer* and a *Behavior Layer*. On the Motivation Layer (MLayer), a motivation is conceived (by an agent) which drives the generation of some behaviors on the Behavior Layer (BLayer). Note, a motivation can be intentional or unintentional. For example, if A looks sad and B asks her "Are you OK?", then A's sadness is also an interaction-related motivation. Of course, A may not intend to communicate her sadness originally, but the act of interaction between A and B is already established. Thus in this case, we also view A's sadness as a motivation. This is similar at the stage of creating pre-condition for interaction: if A sits in the classroom and B walks into it so that A and B have visual access to each other, then the pre-condition of interaction is established, although B may not walk into the room on the purpose of interacting with A, originally. In an interaction, agents' intentional and unintentional motivations are manifested by some behaviors that are generated on the BLayer. Two generators (verbal and non-verbal generator) on this layer are responsible for generating spoken language and various non-verbal behaviors according to the motivation conceived, respectively.

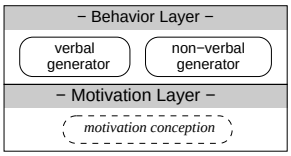

Figure 2: The structure of an Interaction Unit (IU)

The two generators do not need to be instantiated at the same time, this is to say, an agent may express her motivations using one or more modalities. For example, if one smiles upon the Presentation of her interaction partner, her non-verbal generator on the BLayer of her IU is instantiated while the verbal generator is not. However, if she smiles and says something at the same time, then both generators on the BLayer are instantiated. Note, the relationship between the two generators is variable. For example, Scherer&Wallbott (1979) stated that non-verbal behavior can *substitute*, *amplify*, *contradict* and *modify* the meaning of the verbal message. Iverson et al. (1999) studied human gestures and identified three types of *informational* relationship between speech and gesture: reinforcement (gesture reinforces the message conveyed in speech, e.g., emphatic gesture), disambiguation (gesture serves as the precise referent of the speech, e.g., deictic gesture accompanying the utterance "this cup"), and adding-information (e.g., saying "The ball is so big." and shaping the size with hands). In case of instantiation of both generators, we currently focus on the disambiguation function of non-verbal behaviors for user input processing and the amplifying function for behavior generation for the robot.

The interaction contributions involved in the example in Fig. 3 can be modeled using IUs, as illustrated in 4. Here, one or more IUs can be identified for each turn: turn A_1 is represented by an IU with an unintentional motivation on the MLayer and an instantiated non-verbal generator on the BLayer because no speech is involved. Turn B_1 consists of two IUs, the first one, B's turning to face A, has an intentional motivation on the MLayer (because B intends to talk with A) and only the non-verbal generator on its BLayer is

A_1: *(walks into the classroom)*
B_1: *(turns to face A)* Hi, How are you doing?
A_2: *(smiles)* Very well!
B_2: So you passed the exam?
A_3: *(shows B her certificate.)*

Figure 3: An interaction example between A and B (non-verbal behaviors are italic)

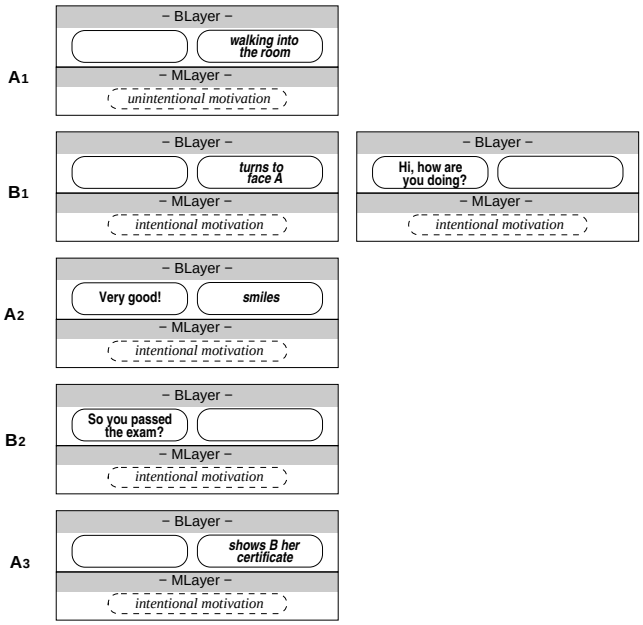

Figure 4: Representing interaction contributions involved in the interaction example in Fig. 3 using IUs

instantiated. The second IU in B_1, the greeting phrase, has an instantiated verbal generator on its BLayer. The rest of contributions from A and B in this example are structured following the similar principle. IUs are the central construct of our model to handle multi-modality. During an interaction, they are assigned to roles of either Presentation or Acceptance and then organized based on the grounding principle. The next subsection briefly discusses our grounding mechanism.

Organizing Interaction Units

As mentioned earlier, the discourse of the interaction does not only concern propositional information of the interaction, as it is the case in traditional spoken dialog discourse management, but also physical possibilities of contact and interactional information involved in the multi-modal interaction. Though, the grounding mechanism, i.e., how the IUs are organized in the discourse, follows the same grounding principle. Several grounding models have been proposed in the spoken dialog research, e.g., the finite state-based grounding model of Traum (1994) and the exchange model of Cahn&Brennan (1999). We combined their advantages and adopted a push-down automaton-based model. This grounding model was discussed in (Li, Wrede, & Sagerer 2006a) in detail. For the reason of relevance, we only briefly summarize its essentials here.

The grounding unit of the discourse, i.e., the unit of the discourse at which the grounding takes place, is the so-called Exchange and it consists of a pair of IUs. One IU plays the role of Presentation and the other one Acceptance. An Exchange reaches the state "grounded" only if the Acceptance of the Presentation is available Acceptance can be implicit or explicit which is a modification of Clark's *strength of evidence principle* and addresses the phenomena that not all the Presentations have to be explicitly accepted in verbal or non-verbal way. These Exchanges are organized in a stack which represents the ungrounded discourse up to the current state. The grounding status of the whole stack is dependent on the status of the individual Exchanges and the relations between them. We introduce four types of such relations (*default*, *support*, *correct* and *delete*). These grounding relations have local effects on the grounding status of their previous Exchanges. During an interaction, participants initiate IUs of either the role of Presentation or Acceptance that cause either a new Exchange to be pushed onto the stack or an existing Exchange to be popped from the stack, respectively. An empty stack means that there is nothing left to be grounded between interaction participants, and it is reasonable to assume that the participants strive to reach this state. All the popped exchanges are collected into a vector which records the complete interaction history.

Following the above grounding model, the interaction example in Fig. 3 is modeled as the structure in Fig. 5. In A_1, A walks into the classroom and it is viewed as a Presentation that initiates an Exchange Ex_1. Upon receiving this signal from A. B turns to face A as the direct reaction to this Presentation. Since the physical possibility of contact is thus established, B's turning to face A is the Acceptance of Ex_1, which is then popped from the stack (not illustrated). Then B proposes a propositional account (the greeting phrase) which initiates a new Exchange Ex_2. A addresses this Exchange by answering A's question using speech and facial expression so that this Exchange is grounded and popped from the stack. Subsequently, B initiates a new Exchange Ex_3 by asking the question "So you passed the exam?" and this Exchange is grounded by A_3 and popped. Now the discourse is empty indicating that there is nothing ungrounded left between A and B.

Implementation

This framework was already implemented on our mobile robot BIRON and humanoid robot BARTHOC (Fig. 6). In these robots, the interaction system creates IUs for each user input and message that is sent by the robot control system. Further, the interaction system assigns roles (either Presentation or Acceptance) to these IUs and organizes them in the discourse. Information that is used by the system to make the decision concerning the roles of IUs are the communication success (e.g., if the speech input is clearly understood) and the robot task execution status. BIRON and BARTHOC use this mechanism to interpret user conversational gestures (Li *et al.* 2005), which are viewed as generated by the non-

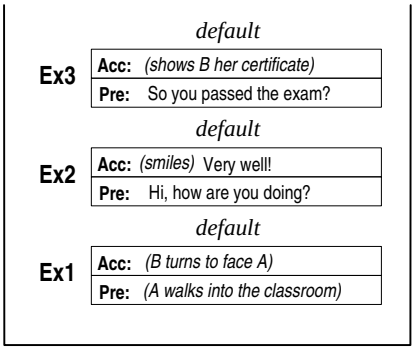

Figure 5: The structure of the interaction discourse for the example in Fig. 3 (Ex: exchange; Pre: Presentation; Acc: Acceptance)

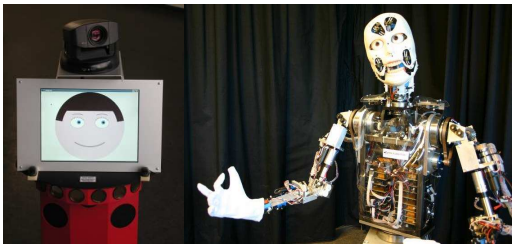

Figure 6: Robots BIRON and BARTHOC

verbal generator on the BLayer of the user's IU and having the disambiguation relation to the simultaneously generated speech from the verbal generator. Our user study (Li, Wrede, & Sagerer 2006b) provides strong evidence for the flexible interaction style enabled by this framework.

In the current system, the two interaction cases discussed in section Observation are handled as following: For the case of meta-commentators, the system generates an IU with only the non-verbal generator on the BLayer being instantiated to express its regret for doing a bad job. What the user sees is thus only a sorrowful face from the robot. This IU initiates an Exchange that only requires an implicit Acceptance which means that this Exchange is still grounded even if no Acceptance from the user is perceived. For the case of quiet speakers, the user's physical presence is viewed as a non-verbal Presentation of the user, for which the system should provide an Acceptance. The system does this by generating an IU that performs a certain facial expression on its BLayer representing its internal state. When the status of the user is changed from "Person" to "Communication Partner", i.e., when the user proposes a new Presentation, the system generates a new IU as Acceptance, on the CLayer of which a new facial expression is then presented.

Benefits and deficiencies

Our grounding-based multi-modal interaction framework attempts to model the exchange of both interactional and propositional information during an interaction using one grounding principle. It does not require categorization of this information and thus simplifies the interaction mechanism and increases the implementability of the model. Although it is not the focus of our work, it is still possible to extend the BLayer with a "Modality Manager", which accomplishes sophisticated modality fusion and selection. This local extension will not affect the overall interaction management. Another way to extend this model is to add synchronization mechanism into the BLayer to enable a synchronized behavior generation. Last but not least, the IUs can also be used to extend interaction systems with a simple discourse management mechanism (e.g., finite state-based) since it only affects the representation of local interaction contributions. Since the majority of the current conversational interaction systems running on robots are finite state-based (Li, Wrede, & Sagerer 2006a) this concept of IUs would enable many other robots to handle multi-modal input and output without changing the underlying interaction management mechanism.

As mentioned earlier, our interaction framework is based on the reciprocal nature of behaviors involved in an interaction and its strength lies in the modeling of information that is *exchanged* between interaction participants. Some applications focus on the development of subtle human-like behaviors from which no clear motivations relating to other participants can be derived, e.g., looking away when thinking about something. Although the current model can still be used (e.g., by instantiating non-verbal generator on the BLayer with "looking away"), it may not be the best suitable choice for such applications.

Conclusion

In this paper we discussed the importance of taking into account both verbal and non-verbal behaviors for a robot by drawing upon our observations of two interaction cases in a user study. These observations confirmed the complementary benefits of these behaviors for HRI. We also proposed a powerful interaction framework that makes use of the reciprocal nature of interaction behaviors and models them based on the grounding principle. This framework is able to flexibly handle complex, multi-modal human-robot interaction. We are currently conducting a new user study with our robots, that can generate facial expressions now, to further verify our framework.

Acknowledgment

This work is supported by the European Union within the Cognitive Robot Companion (COGNIRON) project (FP6-002020)

References

Billard, A. 2005. Challenges in designing the body and the mind of an interactive robot. In *AISB05, Symposium on Robot Companions: Hard Problems and Open Challenges*.

Breazeal, C. L. 2002. *Designing Sociable Robots*. MIT Press.

Brooks, A. R. 1989. A robot that walks; emergent behaviors from a carefully evolved network. MIT AI Lab Memo 1091.

Cahn, J. E., and Brennan, S. E. 1999. A psychological model of grounding and repair in dialog. In *Proc. Fall 1999 AAAI Symposium on Psychological Models of Communication in Collaborative Systems*.

Cassell, J.; Bickmore, T.; Campbell, L.; and Vilhjalmsson, H. 2000. Human conversation as a system framework: Designing embodied conversational agents. In Cassell, J.; Sullivan, J.; Prevost, S.; and Churchill, E., eds., *Embodied conversational agents*. MIT Press.

Cassell, J. 2000. More than just another pretty face: Embodied conversational interface agents. *Communications of the ACM* 43(4).

Clark, H. H., ed. 1992. *Arenas of Language Use*. University of Chicago Press.

Dautenhahn, K.; Woods, S.; Kaouri, C.; Walters, M.; Koay, K. L.; and Werry, I. 2005. What is a robot companion - friend, assistant or butler?

Iverson, J. M.; Capirci, O.; Longobardi, E.; and Caselli, M. C. 1999. Gesturing in mother-child interactions. *Cognitive Develpment* 14(1):57–75.

Kanda, T.; Ishiguro, H.; Ono, T.; Imai, M.; and Nakatsu, R. 2002. Development and evaluation of an interactive humanoid robot robovie. In *Proc. Int. Conference on Robotics and Automation*.

Lang, S.; Kleinehagenbrock, M.; Hohenner, S.; Fritsch, J.; Fink, G. A.; and Sagerer, G. 2003. Providing the basis for human-robot-interaction: A multi-modal attention system for a mobile robot. In *Proc. Int. Conf. on Multimodal Interfaces*.

Larsson, S., and Traum, D. 2000. Information state and dialogue management in the trindi dialogue move engine toolkit. *Natrual Language Engineering* 6(3–4).

Lemon, O.; Bracy, A.; Gruenstein, A.; and Peters, S. 2001. Information states in a multi-modal dialogue system for human-robot conversation. In *Proceedings Bi-Dialog, 5th Workshop on Formal Semantics and Pragmatics of Dialogue*.

Li, S.; Haasch, A.; Wrede, B.; Fritsch, J.; and Sagerer, G. 2005. Human-style interaction with a robot for cooperative learning of scene objects. In *Proc. Int. Conf. on Multimodal Interfaces*. ACM Press.

Li, S.; Wrede, B.; and Sagerer, G. 2006a. A computational model of multi-modal grounding. In *Proc. SIGdial workshop on discourse and dialog, in conjunction with COLING/ACL*. ACL Press.

Li, S.; Wrede, B.; and Sagerer, G. 2006b. A dialog system for comparative user studies on robot verbal behavior. In *Proc. 15th Int. Symposium on Robot and Human Interactive Communication*. IEEE Press.

Matsui, T.; Asoh, H.; Fry, J.; Motomura, Y.; Asano, F.; Kurita, T.; Hara, I.; and Otsu, N. 1999. Integrated natural spoken dialogue system of jijo-2 mobile robot for office services,.

Scherer, K. R., and Wallbott, H. G., eds. 1979. *Die Funktionen des nonverbalen Verhaltens im Gespräch, in Nonverbale Kommunikation: Forschungsberichte zum Interaktionsverhalten*. Beltz Verlag.

Traum, D., and Rickel, J. 2002. Embodied agents for multiparty dialogue in immersive virtual world. In *Proc. 1st Int. Conf on Autonomous Agents and Multi-agent Systems*.

Traum, D. 1994. *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. Dissertation, University of Rochester.