FP6-IST-002020

COGNIRON

The Cognitive Robot Companion

Integrated Project

Information Society Technologies Priority

D1.1.1
Declarative dialogue model and strategy

Due date of deliverable: 31/12/2004
Actual submission date: 31/01/2005

Start date of project: January 1st, 2004 **Duration:** 48 months

Organisation name of lead contractor for this deliverable:
Bielefeld University

Revision: Final
Dissemination Level: PU

Executive Summary

The objective of the WP 1.1 in the first phase is to develop a dialogue model that is able to handle basic dialogic operations including initiation of clarifying questions and explanation of robot capabilities. It also should facilitate the processing of multi-modal inputs and the representation of semantic content by perception-based semantics.
In this first phase of the project we successfully employed a state-based strategy to model the dialogue which enables basic human robot interaction as specified. The dialogue component integrates input from speech and gesture recognition for the processing of multi-modal deictic phrases. Based on results from user studies carried out in WP 1.3 we developed a more advanced dialogue model which will be integrated in the current system in phase 2. A basic concept for multi-modal representation based on perception-based semantics has been developed in WP 1.2. A first implementation will be finished at the end of phase 1.
Results from work in WP 1.1 have been presented at several international conferences. The current dialogue manager was presented at the 8th International Conference on Spoken Language Processing (ICSLP) [2]. First tentative results of system tests from user interactions with the implemented dialogue on the robot platform BIRON were presented at the International Conference on Systems, Man and Cybernetics in 2004 [1] .
As a proof of concept the dialogue system was demonstrated on the mobile robot platform BIRON at the IST-Event 2004 in The Hague where several users interacted with the robot via the multi-modal dialogue component in a demanding situation outside the laboratory.

Role of dialogue in Cogniron

Using language is one of the most intuitive ways of humans to communicate with each other and the ability to process symbolic information conveyed by language is one of the most important cognitive abilities of humans. A robot companion that can communicate with humans this way has to exhibit extensive cognitive ability in order to increase its acceptability to users.
While traditional dialogue modelling focuses on the modelling of pure verbal language performance we are extending our field of research to multi-modal dialogue that makes human-robot-interaction more natural. This point is discussed in the deliverable WP 1.2 in more detail. Furthermore, we are interested in investigating and modelling the involvement of other cognitive modules in language performance.

Relation to the Key Experiments

The main scenario for the multi-modal dialogue is the Robot Home-Tour scenario (KE 1). The current implementation has been developed within the Home-Tour scenario and further research will mainly focus on this KE.
In the key-experiment 1 ”Robot Home-Tour”, the interaction with the human user is mainly carried out via the dialogue system. The robot acquires information about the home environment via speech and gesture and forwards it to other system modules for further processing. These requirements are addressed in the dialogue model. Also, user studies of the dialogue module have been carried out within this scenario in WP 1.3.

1 Dialogue Model

In the first year of the project a first version of the dialogue module has been implemented with a substantial part of the work being dedicated to the integration of the dialogue manager to the mobile platform BIRON. Since the running system was not ready from the beginning of the project, the evaluation in WP 1.3 had to be carried out in a Wizard-of-Oz scenario according to the dialogue specifications. In the next phase, it will be possible to carry out evaluations with the real system running on the BIRON platform. As a proof of concept we ran first system tests with untrained users. The integrated system and its tests have been presented at the International Conference on Systems, Man and Cybernetics in 2004 [1].

A detailed description of the underlying mechanisms and implementation of the dialogue module is given in [2]. In short, it is implemented on the mobile platform BIRON in a modular communication system architecture which clearly separates the speech understanding, dialogue management, and modality integration components. The dialogue manager receives semantic analysis results of the speech input from the speech understanding component. In case that the semantic structure indicates the involvement of other modalities, the dialogue manager will consult a modality integration component for further information. After successful interpretation of the user commands they will be sent to the robot control component for execution. The current system status and the execution results will be communicated to the dialogue manager via periodical events sent by the robot control component. The dialogue manager is based on a finite state machine that is extended with the ability of recursive activation of other finite state machines and the execution of an action in each state. Actions that can be taken in certain states are specified in the policy of the dialogue manager. By dividing the dialogue into sub-dialogs a modular structure could be achieved. Each sub-dialogue is associated with a task and is modelled by a Finite State Machine. The dialogue strategy is based on a slot-filling mechanism that has proved to be a robust technique in dialogue modelling. A slot is an information item for which a value is required. Users' utterances contain information that can be quantised into such information items. The task of the dialogue manager is to pick out the required information items from the utterances and fill the pre-specified slots to meet the dialogue goal, which is defined as the goal state in the corresponding sub-dialog. The processing of each sub-dialogue can be interrupted by another sub-dialogue and then resumed later.

This dialogue model has proved to be a simple and robust solution for human robot interaction via speech. In the first year of the project we were able to fully integrate the dialogue module on our robot platform BIRON and test and demonstrate it on several occasions. In a controlled situation 21 users went through a simple test procedure and their feedback was collected via questionnaires. Results from these system tests are described in detail in [1]. During the 3 days at the IST Event 2004 the system was running without any major problems and many exhibition visitors interacted with BIRON or actively observed the interaction.
From these experiences and from observations from the Wizard-of-Oz studies carried out in WP 1.3 important conclusions can be drawn that are relevant for the further design of the human robot interface. When interacting with a robot users tend to make extensive use of multi-modal interaction cues especially when showing things to the robot. The data collected in WP 1.3 suggest that by making use of such multi-modal cues users tend to decrease their verbal communication. The dialogue has to take this into account. Another important multi-modal event is the start of an interaction when the user wants to get the attention of the robot by waving or even whistling. All these multi-modal cues have not yet been foreseen by the dialogue model and need to be integrated in the dialogue module.

The data further indicate that there are highly personalised interaction styles by the different users which indicates the need for adaptation strategies of the dialogue module. As for the verbal part of the dialogue, first user's reactions indicate that there is a desire for less restrictions in the sequence of instructions and the wording in general. Also, the system is currently lacking a systematic handling of cases of mis- or non-understanding. It would be desirable to introduce an internal monitoring of the dialogue history in order to estimate the success rate of the interaction.

2 Future Work

Based on our experiences from system tests and observations from user studies performed in WP 1.3 we defined three areas of research in the second phase of the project:

- *Adaptation to the user:* The user studies have shown that there are highly individual ways of users interacting with a robot. In phase 2 we want to develop both verbal and non-verbal strategies for (1) determining the interaction style of the user and (2) adapting the robot's behaviour to the user. Therefore, a strong collaboration with UH is planned, who will join WP 1.1 as a new partner in phase 2 in order to facilitate a transfer of results from user studies in RA 3 on non-verbal interaction styles into the dialogue system.

- *Integrating multi-modal cues in a more intuitive way:* The integration of speech, deictic gestures, and basic visual features is discussed in deliverable WP 1.2 in detail. This integration process will be carried out according to the shared cognitive basis of different modalities in communication. This approach will heavily rely on the multi-modal representation of objects, actions, situations, etc. In phase 2 the representation of objects for resolving object references and locations for navigational tasks will be another major issue in RA 1. For this, we will strengthen our cooperations with RA 2 (recognition of objects) and RA 5 (map building) in the second phase.

- *Improving the naturalness of the dialogue:* As observed in first system tests the human users would prefer more freedom in their wording and topic transition. A promising way to realise this is to change from a system-oriented sub-dialogue predefinition to user-oriented discourse modelling by common ground building. Dialogue can be viewed as a collaborative act that is built on mutual understanding of the current task or intentions, the so-called common ground. Based on this psycholinguistic law of dialogue we will be able to model a more flexible and more powerful dialogue.

3 References

3.1 Applicable documents

3.2 Reference documents

[1] S. Li, M. Kleinehagenbrock, J. Fritsch, B. Wrede, and G. Sagerer. "BIRON, let me show you something": Evaluating the interaction with a robot companion. In W. Thissen, P. Wieringa, M. Pantic, and M. Ludema, editors, *Proc. IEEE Int. Conf. on Systems, Man, and Cybernetics, Special Session on Human-Robot Interaction*, pages 2827–2834, The Hague, The Netherlands, October 2004. IEEE.

[2] Ioannis Toptsis, Shuyin Li, Britta Wrede, and Gernot A. Fink. A multi-modal dialog system for a mobile robot. In *Intern. Conf. Spoken Language Processing*, volume 1, pages 273–276, Jeju, Korea, 2004.

Annexes

The papers cited above were all published in the first phase of the project and are attached to this report.

"BIRON, let me show you something": Evaluating the Interaction with a Robot Companion*

Shuyin Li, Marcus Kleinehagenbrock, Jannik Fritsch, Britta Wrede, and Gerhard Sagerer
Faculty of Technology, Bielefeld University, 33594 Bielefeld, Germany
Email: shuyinli@TechFak.Uni-Bielefeld.DE

Abstract – *Current research on the interaction with a robot is driven by the desire to build intuitive and natural interaction schemes. In order for our robot BIRON to behave naturally we integrated an attention system that enables the robot to search for and eventually focus on human communication partners by detecting and tracking persons. Via a natural language interface the user can then interact with BIRON and teach him new objects or ask him to follow her. First evaluation results from 21 users interacting with the robot indicate that users appreciate the natural language capabilities of BIRON. However, users are very sensitive to speech recognition failures even though all of our subjects had prior experience with speech recognition systems. The results also indicate that feedback on the internal status of the robot is extremely helpful for users.*

Keywords: human-robot interaction, robot companion, system evaluation.

1 Introduction

The development of cognitive robots serving humans as assistants or companions is currently an active research field. In order to be accepted as a communication partner by non-expert users such robot companions must exhibit a human-like communicative behavior. In the literature a trend to develop more human-like behaving robots can be observed. The increasing interaction capabilites of such integrated systems now start to reach a level where they can (and have to) be tested by more naive users. This raises the issue of which aspects should actually be evaluated.

A robot companion is a highly complex system consisting of many different components that have to interact with each other in a meaningful way. For example, our robot BIRON – the Bielefeld Robot Companion – uses different sensors to increase reliability in case of occlusions (e.g. in vision) and robustness against processing errors within a single modality. At the cognitive level a robot companion needs to be able to detect humans and has to be aware when a person wants to interact with the robot. Therefore, a multi-modal attention system and a multi-modal dialog manager have been developed for our robot system.

*0-7803-8566-7/04/$20.00

Figure 1: Several users interacting with BIRON.

This complex robot system poses problems with respect to its evaluation. On the one hand it would be desirable to compare the performance of single components by objective measures of correctness, e.g., as known from speech recognition in terms of word error rate. Thus, we could measure the word error rate of the recognizer, the speech understanding performance, or the correctness of the dialog manager and attention system. However, there are several problems with this approach. Firstly, it is not always possible to determine what a "correct" system response would be. For example, a system may react correctly in that it performs the action explicitly requested by the user. But the user might expect the system to also give adequate human-like feedback such as nodding or other verbal and non-verbal positive signals. Secondly, it is not necessarily the case that a "correct" system is the most user-friendly one. This is a broadly made experience with telephone information services which guide users through very lengthy and inflexible dialogs to ensure a very high accuracy of the speech recognition system. Thirdly, evaluations of single components neglect the fact that crucial functionalities might be missing. For example, users might ask the robot to perform actions that it can not execute. In this case the robot should not only be able to answer that it can not perform the required action, but also indicate which actions it can do. Finally, the integrated sys-

tem is more than the sum of its parts. In other words, the interplay of different system components can lead to more (or less) "correct" behavior of the system. A simple example is a speech understanding component that discards function words, such as articles or prepositions, for further processing because they are highly susceptible to speech recognition errors. This way errors from the speech recognizer can be discarded and partially recognized utterances may still be transformed into correct actions.

In this paper we will present qualitative results from interactions of 21 users with our robot companion BIRON. Our goal is to use BIRON in the so-called *home-tour* scenario. Here, the basic idea is that a human introduces to a newly purchased robot all the objects and places in a private home relevant for later interaction. Figure 1 shows typical scenes where users interact with the robot. In our evaluation scheme we asked the users to rate different aspects concerning the interaction with the robot and to point out the most interesting as well as the most annoying features. We also assessed the users' attitudes and preferences with regard to the envisioned scenario of a robot as a companion for the home. These results allow us to draw conclusions on where to guide the further development of our system.

2 Related Work

The most advanced examples of robots realizing complex multi-modal human-robot interfaces are *SIG* [15] and *ROBITA* [14]. While only ROBITA is a truly mobile system both robots have a humanoid torso with cameras and microphones embedded in the robot's "head". SIG's focus of attention is directed towards the person currently speaking that is either approaching the robot or standing close to it. In addition to the detection of talking people, ROBITA is able to determine the addressee of spoken utterances.

There are also several complete service robot systems that integrate capabilities for human-robot interaction. For example, Care-O-bot II [7] is a multi-functional robot assistant for housekeeping and home care, designed to be used by elderly people. It receives user input via speech and touch screen. Although the system also produces speech output, it can not carry out natural dialogs. Lino [10] serves as user interface to intelligent homes. It perceives persons by processing visual and auditory information. Since the robot operates in an intelligent environment it makes use of external information sources. The humanoid service robot *HERMES* [2] can be instructed for fetch-and-carry tasks, and it was also adopted as museum tour guide. It integrates visual, tactile, and auditory data to carry out dialogs in a natural and intuitive way, but can only interact with single persons. Jijo-2 [1] is intended to perform tasks in an office environment, such as guiding visitors or delivering messages. It uses data coming from a microphone array and a pan-tilt camera to perceive persons, but a person is only focused after it says *"Hello"* to the robot.

Although different robotic systems have been developed in the last years, relatively little evaluation work was done on robots. The performance of Jijo-2 was roughly evaluated during some demonstrations [13]. They mainly concentrated on the basic performance of the speech processing component and the bridging function of the dialog system. The evaluation of the interaction capabilities of ROBITA, especially the impact of the combination of facial and verbal expressions, was documented in [17]. Ten subjects were asked to interact with ROBITA with and without facial expressions and to fill out questionnaires afterwards. Despite these efforts, there still seems to be a lack of qualitative evaluations of human-robot interactions that are of key importance for developing robot companions with natural interaction capabilities.

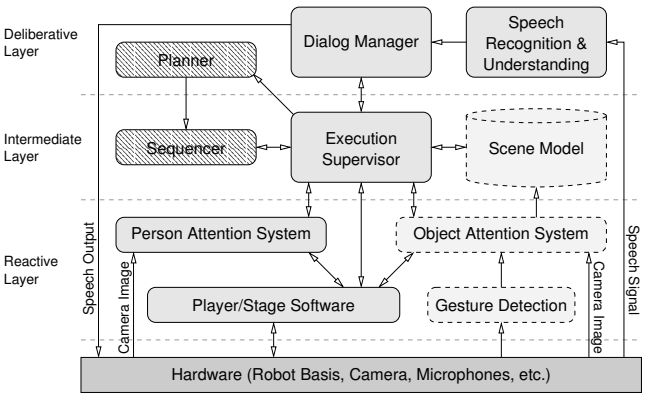

Figure 2: Overview of the BIRON architecture (implemented modules are drawn with solid lines, modules under development with dashed lines).

3 Overall System Architecture

Since interaction with the user is the basic functionality of a robot companion, the integration of interaction components into the architecture is a crucial factor. We propose to use a special control component, the so-called *execution supervisor*, which is located centrally in the robot's architecture. The data flow between all modules is event-based and every message is coded in XML. The modules interact through a specialized communication framework [20]. The robot control system (see Fig. 2) is based on a three-layer architecture [5] which consists of three components: a reactive feedback control mechanism, a reactive plan execution mechanism, and a mechanism for performing deliberative computations.

The execution supervisor, the most important architecture component, represents the reactive plan execution mechanism. It controls the operations of the modules responsible for deliberative computations rather than vice versa. This is contrary to most hybrid architectures where a deliberator continuously generates plans and the reactive plan execution mechanism just has to assure that a plan is executed until a new plan is received. To continuously control the overall system the execution supervisor performs only computations that take a short time relative to the rate of environmental change perceived by the reactive control mechanism.

While the execution supervisor is located in the intermediate layer of the architecture, the dialog manager is part of the deliberative layer. It is responsible for carrying out dialogs to receive instructions given by a human interaction partner. The dialog manager is capable of managing interaction problems and resolving ambiguities by consulting the user (see Sect. 7). It receives input from speech processing which is also located on the topmost layer (see Sect. 6) and sends valid instructions to the execution supervisor.

The person attention system represents the reactive feedback control mechanism and is therefore located on the reactive layer (see Sect. 5). However, the person attention system does not directly control the robot's hardware. This is done by the *Player/Stage* software [6]. *Player* provides a clean and simple interface to the robot's sensors and actuators. Even though we currently use this software to control the hardware directly, the controller can easily be replaced by a more complex component which may be based on, e.g., behaviors.

In addition to the person attention system we are currently developing an object attention system for the reactive layer. The execution supervisor can shift control of the robot from the person attention system to the object attention system in order to focus objects referred to by the user. The object attention will be supported by a gesture detection module which recognizes deictic gestures. Combining spoken instructions and a deictic gesture allows the object attention system to control the robot and the camera in order to acquire visual information of a referenced object. This information will be sent to the scene model in the intermediate layer.

The scene model will store information about objects introduced to the robot for later interactions. This information includes attributes like position, size, and visual information of objects provided by the object attention module. Additional information given by the user is stored in the scene model as well, e.g., a phrase like *"This is my coffee cup"* indicates owner and use of a learned object.

The deliberative layer can be complemented by a component which integrates planning capabilites. This planner is responsible for generating plans for navigation tasks, but can be extended to provide additional planning capabilities which could be necessary for autonomous actions without the human. As the execution supervisor can only handle single commands, a sequencer on the intermediate layer is responsible for decomposing plans provided by the planner. However, in this paper we will focus on the interaction capabilities of the robot.

4 Robot Hardware

Our system architecture is implemented on our mobile robot BIRON (see Fig. 3). Its hardware platform is a Pioneer PeopleBot from ActivMedia with an on-board PC (Pentium III, 850 MHz) for controlling the motors and the on-board sensors and for sound processing. An additional PC (Pentium III, 500 MHz) inside the robot is used for image processing and for data association.

The two PCs running Linux are linked by an 100 Mbit Ethernet LAN and the controller PC is equipped with wireless LAN to enable remote control of the robot. As additional interactive device a 12" touch screen display is provided on the front side.

A pan-tilt color camera (Sony EVI-D31) is mounted on top of the robot at a height of 141 cm for acquiring images of the upper body part of humans interacting with the robot. Two AKG far-field microphones which are usually used for hands free telephony are located at the front of the upper platform at a height of 106 cm, right below the touch screen display. The distance between the microphones is 28.1 cm. A SICK laser range finder is mounted at the front at a height of approximately 30 cm.

Figure 3: BIRON.

5 Person Attention System

A robot companion should enable users to engage in an interaction as easily as possible. For this reason the robot has to continuously keep track of all persons in its vicinity and must be able to recognize when a person starts talking to it. Therefore, both acoustic and visual data provided by the on-board sensors have to be taken into account: At first the robot needs to know which person is speaking, then it has to recognize whether the speaker is addressing the robot, i.e., looking at it. On BIRON the necessary data is acquired from a multi-modal person tracking framework which is based on *multi-modal anchoring* [4].

5.1 Multi-Modal Person Tracking

Multi-modal anchoring allows to simultaneously track multiple persons. The framework efficiently integrates data coming from different types of sensors and copes with different spatio-temporal properties of the individual modalities. Person tracking on BIRON is realized using three types of sensors. First, the laser range finder is used to detect humans' legs. Pairs of legs result in a characteristic pattern in range readings and can be easily detected [4]. Second, the camera is used to recognize faces and torsos. Currently, the face detection works for faces in frontal view only [11]. The clothing of the upper body part of a person is observed by tracking the color of the person's torso [3]. Third, the stereo microphones are applied to locate sound sources in front of the robot. By incorporating information from the other cues robust speaker localization is possible [11]. Altogether, the combination of depth, visual, and auditory cues allows the robot to robustly track persons in its vicinity.

In a natural situation, persons are usually moving around. Since also the robot itself is mobile, users can not be expected to be located at a predetermined position. In addition,

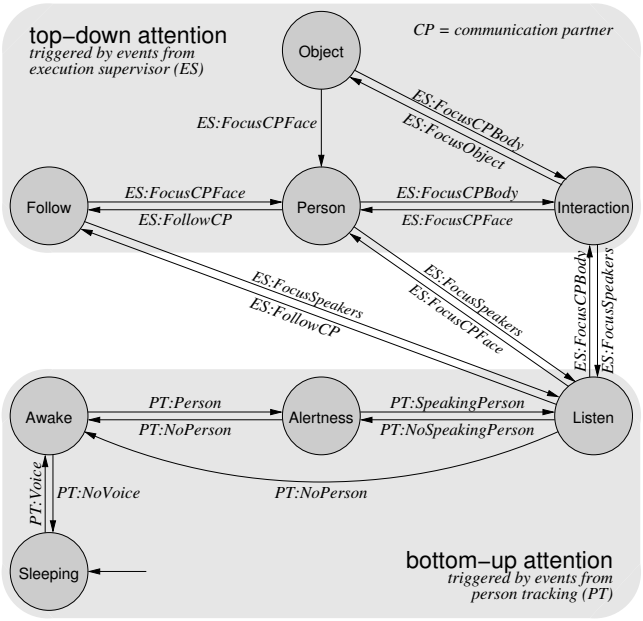

Figure 4: Finite state machine realizing the different behaviors of the person attention mechanism.

as the sensing capabilities of the robot are limited, not all persons in the vicinity of the robot can be observed with all sensors at the same time. To solve these problems an attention mechanism is required.

5.2 Attention Mechanism

The attention mechanism has to fulfill two tasks: On the one hand it has to select the person of interest from the set of observed persons. On the other hand it has to control the alignment of the sensors in order to obtain relevant information from the persons in the robot's vicinity.

The attention mechanism is realized by a finite state machine (see Fig. 4). It consists of several states of attention, which differ in the way the robot behaves, i.e., how the pan-tilt unit of the camera or the robot itself is controlled. The states can be divided into two groups representing *bottom-up attention* while searching for a communication partner and *top-down attention* during interaction.

When bottom-up attention is active, no particular person is selected as the robot's communication partner. The selection of the person of interest as well as transitions between different states of attention solely depend on information provided by the person tracking component. For selecting a person of interest, the observed persons are divided into three categories with increasing degree of relevance. The first category consists of persons that are not speaking. The second category comprises all persons that are speaking, but at the same time are either not looking at the robot or the corresponding decision is not possible, since the person is not in the field of view of the camera. Persons assigned to the third category are of most interest to the robot. These persons are speaking and at the same time are looking at the robot. In this case the robot assumes to be addressed and considers the corresponding person to be a potential communication partner. If a person is assigned to this category it is instantly selected and remains selected until the person changes to one of the other categories, e.g., by stopping to talk or looking in another direction. If no person has the status of a potential communication partner, the attention mechanism always selects the person that is of most interest, e.g., persons of the second category are selected prior to persons of the first category. If the mechanism has to decide between multiple persons of the same category, it selects the one that for the longest time was not selected. In addition, the mechanism will also switch between persons in order to obtain additional information, e.g., the identity of persons present. For this purpose, a person remains selected only for a limited amount of time, after which it is temporarily blocked for selection, realizing an effect known as inhibition of return.

Top-down attention is activated as soon as the robot starts to interact with a particular person. During interaction the robot's focus of attention remains on this person even if it is not speaking. Here, in contrast to bottom-up attention, transitions between different states of attention are solely triggered by the execution supervisor. The corresponding events sent by the execution supervisor depend on the current state of the dialog. The behavior of the robot concerning the states of the attention mechanism differs in the way the pan-tilt unit of the camera and the robot itself is controlled. For detailed information concerning the control of the hardware see [8].

6 Speech Processing

As speech is the most important modality for a multimodal dialog, speech processing has to be done thoroughly. On BIRON there are two major challenges: Speech recognition has to be performed on distant speech data recorded by the two on-board microphones and speech understanding has to deal with spontaneous speech.

While the recognition of distant speech with our two microphones is achieved by beam-forming [12], the activation of speech recognition is controlled by the attention mechanism presented in the previous section. Only if a tracked person is speaking and looking at the robot at the same time, speech recognition and understanding takes place. Since the position of the speaker relative to the robot is known from the person tracking component, the time delay can be estimated and taken into account for the beam-forming process. However, since noise and speech from interfering talkers standing at different positions can only be suppressed to some extent by beam-forming, the recognition quality will never reach the one obtained with a close-talking microphone.

The speech understanding component processes recognized speech and has to deal with spontaneous speech phenomena. For example, large pauses and incomplete utterances can occur in such task oriented and embodied communication. However, missing information in an utterance can often be acquired from the scene. For example the utterance *"Look at this"* and a pointing gesture to the table can be

combined to form the meaning *"Look at the table"*. Moreover, fast extraction of semantic information is important for achieving adequate response times.

We obtain fast and robust speech processing by combining the speech understanding component with the speech recognition system. For this purpose, we integrate a robust LR(1)-parser into the speech recognizer as proposed in [19]. Besides, we use a semantic-based grammar which is used to extract instructions and corresponding information from the speech input. A semantic interpreter forms the results of the parser into frame-based XML-structures and transfers them to the dialog manager (see Sect. 7). Hints in the utterances about gestures are also incorporated. For our purpose, we consider co-verbal gestures only. For the object attention system it is intended to use this information in order to detect a specified object. Thus, this approach supports the object attention system and helps to resolve potential ambiguities.

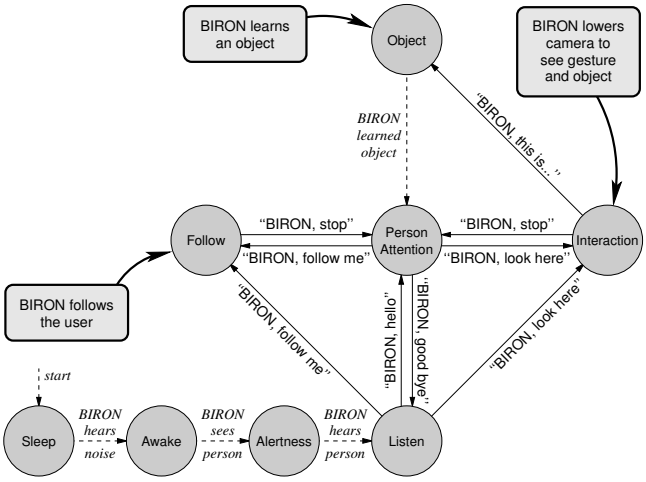

Figure 5: Speech commands and internal states of BIRON.

7 Dialog Manager

The model of the dialog manager is based on a set of *finite state machines* (FSM), where each FSM represents a specific dialog [18]. The FSMs are extended with the ability of recursive activation of other FSMs and the execution of an action in each state. Actions that can be taken in certain states are specified in the *policy* of the dialog manager. These actions include the generation of speech output and sending events like orders and requests to the execution supervisor. The dialog *strategy* is based on the so-called *slot-filling* method [16]. The task of the dialog manager is to fill enough slots to meet the current dialog goal, which is defined as a goal state in the corresponding FSM. The slots are filled with information coming from the user and other components of the robot system. This procedure can be viewed as a quantization of a user utterance into required information items. After executing an action, which is determined by a lookup in the dialog policy, the dialog manager waits for new input from the execution supervisor or the speech understanding system.

As users interacting with a robot companion often switch between different context, the slot-filling technique alone is not sufficient for adequate dialog management. Therefore, the processing of a certain dialog can be interrupted by another one, which makes alternating instruction processing possible. Dialogs are specified using a declarative definition language and encoded in XML in a modular way. This increases the portability of the dialog manager and allows an easier configuration and extension of the defined dialogs.

8 Interaction Capabilities

In the following we describe the interaction capabilities BIRON offers to the user in our current implementation. Initially, the robot observes its environment. If persons are present in the robot's vicinity, it focuses on the most interesting one (see section 5). A user can start an interaction by greeting the robot with, e.g., *"Hello BIRON"* (see Fig. 5). Then, the robot keeps this user in its focus and can not be distracted by other persons talking. Next, the user can ask the robot to follow him to another place in order to introduce it to new objects. While the robot follows a person it tries to maintain a constant distance to the user and informs the person if it moves too fast. When the robot reaches a desired position the user can instruct it to stop. Then, the user can ask the robot to learn new objects. In this case the camera is lowered to also get the hands of the user in the field of view. When the user points to a position and gives spoken information like *"This is my favorite cup"*, the object attention system is activated in order to center the referred object. If the user says *"Good-bye"* to the robot or simply leaves while the robot is not following the user, the robot assumes that the current interaction is completed and looks around for new potential communication partners.

9 Evaluation

We designed an experiment to evaluate the performance of the integrated system and the general interaction capabilities of BIRON. We were also interested in questions that related to general acceptance of robot companions among users.

The experiments were carried out in a large room so that the robot could move around without colliding with obstacles (see Fig. 1). We recruited 21 subjects at the age between 22 and 54 ranging from farmer to computer scientists. But most of them turned out to have a rather technical background. They were instructed to go through the following procedure using verbal commands (cf. Fig. 5):

1. "Awaking" BIRON to start interaction with it.
2. Asking BIRON to follow him/her and later to stop the following action.
3. Showing BIRON some objects by referencing them using speech and gesture.
4. Saying *"Good-bye"* to BIRON.

Each subject interacted with BIRON for about 3 to 5 minutes. Then they were asked to fill out our questionnaires.

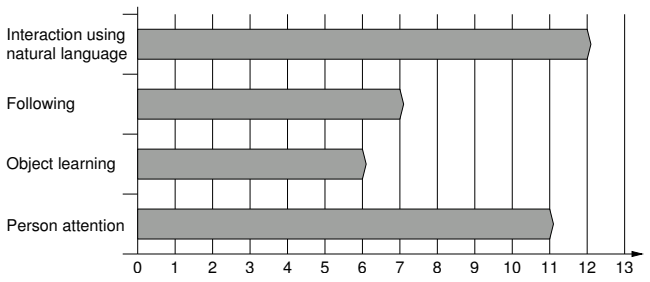

Figure 6: Histogram of answers to "What are the most interesting capabilities of the robot?" (multiple answers were possible.)

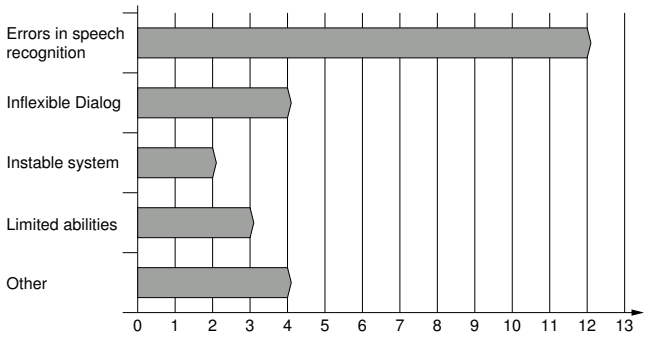

Figure 7: Histogram of answers to "What in the system didn't you like?" (multiple answers were possible.)

9.1 Evaluating individual system components

We qualitatively addressed the system performance issue of the individual components introduced above with two main questions: "What are the most interesting capabilities of the robot?" (multiple choice question) and "What in the system needs to be improved for a better human robot interaction?" (open question). We are going to analyze the results (see Fig. 6, 7) in the following paragraphs. In the last paragraph of this subsection we will discuss the system feedback issue that was also evaluated.

Person Attention System This component received the most positive feedback: Of the 21 subjects, 11 considered our robot's capability of focusing on its communication partners (Person attention) as interesting and 7 of them were impressed by its following action (see Fig. 6). None of the subjects signaled dissatisfaction with the performance of these components in the second question which also indicates that they are perceived as quite human-like.

Speech Processing It was interesting to see that the answers to the first question do not necessarily correlate with those to the second one. The interaction with the robot via natural language was interesting for 11 of the subjects (see Fig. 6), but 12 subjects also agreed that the performance of the speech recognition and understanding component needed to be improved (see Fig. 7). This result indicates that natural language interaction in a robot companion does improve the naturalness of the user interface and attractiveness of the overall system. However, this means at the same time that its performance is crucial for user acceptance of the system.

Dialog Manager The dialog manager plays a central role in the interaction with the user. Our current implementation enables basic communication with users, but it does not seem to be flexible enough, as it limits the user's freedom in wording. This is reflected in the improvements suggestions (see Fig. 7) where 4 subjects explictly suggested a more flexible dialog scheme. We are planning to extend our current dialog system with two additional mechanisms: First, correcting speech recognition errors by connecting speech recognition confidence scores with the dialog manager so that the robot can ask clarifying questions in case of ambiguous speech recognition results. Second, we are planning to add a linguistic component that builds up a discourse structure of the dialog exchanges to enable more flexible dialogs. We expect that the new implementation can tolerate more incorrect speech recognition results and enable a more sophisticated communication with users.

System Feedback Additionally to the speech output that the robot generated, we presumed that users might be also interested to know the robot's internal status. Therefore, we presented the results of the speech recognition system as well as the states of the person attention system (see Fig. 4) to the subjects on a display during the experiment. We divided the subjects into two groups: Only the 11 subjects in the first group were presented the speech recognition results. We tested our hypothesis by directly asking them if this information was helpful, 6 of them answered yes. In the second group where all the 10 group members did not get any speech recognition results during the experiment, 3 of them commented explicitly that they missed this information. Of the 21 subjects, 14 felt that the knowledge about the robot's internal states was very useful. Given the fact that the input processing in our robot takes about 2 to 3 seconds longer than it would in human-human interaction, feedback during these processes is extremely important for users. To directly present users the speech recognition results and robot internal states may not be the most sophisticated way due to its technical nature. Some subjects confirmed this point verbally after the experiment. An alternative is to associate speech recognition results with facial expressions of an animated face that is displayed on the robot's display. Users probably perceive it as the face of the robot that demonstrates its "mental" status such as thinking about things or having found answers. We already have a basic implementation of such a face, but it does not yet support rich facial expressions.

9.2 Studying User Attitudes and Preferences

When studying human-robot interaction it is important to know the attitudes and preferences of potential users concerning robots. This is neglected in many other studies. We

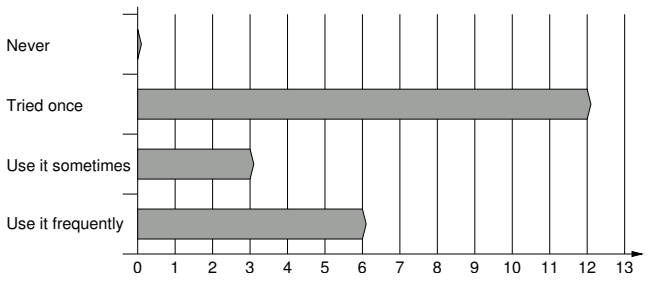

Figure 8: Histogram of answers to "How much experience do you have with speech recognition systems?"

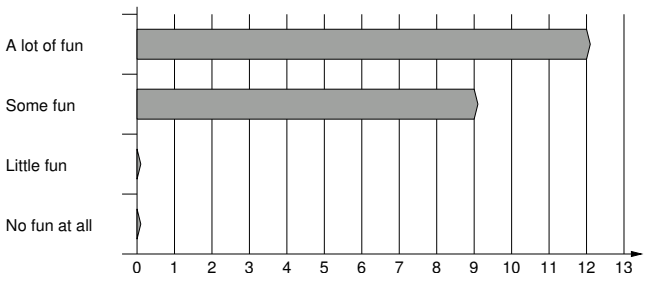

Figure 9: Histogram of answers to "How much fun did you have during the interaction with BIRON?"

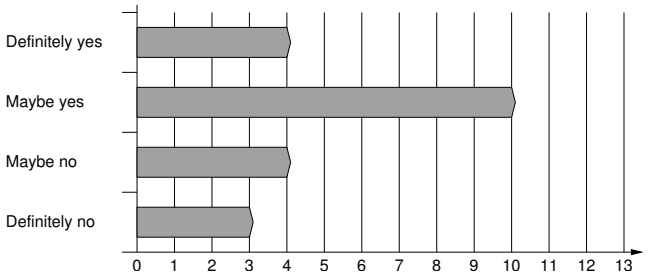

Figure 10: Histogram of answers to "Do you wish to have a robot at home?"

therefore asked some questions about the general attitude of people towards robots and the related technologies and their preferences (see Fig. 8, 9, 10).

Experience with Speech Technology We were quite surprised that all of the subjects reported that they had prior experience with speech recognition systems (see Fig. 8). In spite of the possible effect of social desirability that could arise in such experiments we can assume that most people, at least those who are interested in technics, are more or less prepared for interactions with machines via natural speech. This may be the result of the increasingly popular applications of speech technology in our everyday life, e.g., telephone information services. This result gives us more confidence in speech technology as a main user interface of a robot because of its naturalness and its increasing acceptance among potential users.

User Curiosity In spite of the performance limitations of our system at some points, all subjects had fun with our system (see Fig. 9). While this result implies a general openness of our subjects towards this kind of new technologies we can not ignore the effect of curiosity either. Most subjects had never interacted with a real robot before and it was interesting for them to try it out. It is likely that they will get bored after some time if the robot can only handle a small number of interactions and fulfill limited tasks. To address this problem we can try to develop all-round robots with a great deal of capabilities, but we can also try to enable robots to learn new skills. The latter solution, though more complex, will have more impact on human robot interaction.

Need for a Robot Companion Do people wish to have a robot companion at home? Of 21 subjects, 14 answered this questions with "definitely yes" or "maybe yes" (see Fig. 10). We asked further what kind of abilities and features the robot should have. Of the 14 subjects who answered this question, 9 preferred the robot to do household work, 4 wished to have an entertainment robot and one favored robots with a memory function. On the basis of these data we can conclude that most people still have a relatively "traditional" understanding of a robot's role and expect it to be able to support them in daily life rather than to behave as a real human companion. This result poses interesting questions concerning the direction of the development of robot companions. Should our final goal be robots that are able to accompany people like friends or should the natural interaction capabilities only serve as means to enable the robots to perform complex tasks better to support humans. To answer this question, further studies of user demands are needed.

10 Summary

We presented an overview of the robot companion BIRON and results from a qualitative evaluation scheme based on user judgments rather than objective correctness measures. The results show that while users appreciate a natural speech interface they are highly sensitive to speech recognition failures. In general, users appear to feel a strong need for more information about the internal status of the robot. We also found that users liked the human-like attention behavior, i.e., the movements of the camera in order to detect human communication partners.

From the assessment of the general user attitude we conclude that the idea of having a robot at home is attractive to most users. However, in order for such a robot companion to be accepted it has to work reliably and should be predictable. However, we were only able to assess users' opinions based on single interactions. For more valid statements that rule out effects of curiosity and social desirability it would be necessary to carry out long-term studies. In a long-term study with a service robot it has been shown that every day experience with a robot used for fetch and carry tasks in an office environment [9] lead to unexpected insights with respect to

the influence of contextual variables such as bypassing persons or ergonomic features. Similar studies are necessary for a robot companion that is supposed to "live" in a home environment to assess how attitude and opinion of the user towards the robot change over time.

11 Acknowledgments

The work described in this paper was supported by the EU Integrated Project COGNIRON ("The Cognitive Companion") and funded by the European Commission Division FP6-IST Future and Emerging Technologies under Contract FP6-002020 and by the German Research Foundation within the Research Training Groups "Strategies and Optimization of Behavior" and "Task Oriented Communication".

References

[1] H. Asoh, Y. Motomura, F. Asano, I. Hara, S. Hayamizu, K. Itou, T. Kurita, T. Matsui, N. Vlassis, R. Bunschoten, and B. Kröse. Jijo-2: An office robot that communicates and learns. *IEEE Intelligent Systems*, 16(5):46–55, 2001.

[2] R. Bischoff and V. Graefe. Demonstrating the humanoid robot *HERMES* at an exhibition: A long-term dependability test. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems; Workshop on Robots at Exhibitions*, Lausanne, Switzerland, 2002.

[3] J. Fritsch, M. Kleinehagenbrock, S. Lang, G. A. Fink, and G. Sagerer. Audiovisual person tracking with a mobile robot. In *Proc. Int. Conf. on Intelligent Autonomous Systems*, pages 898–906. IOS Press, 2004.

[4] J. Fritsch, M. Kleinehagenbrock, S. Lang, T. Plötz, G. A. Fink, and G. Sagerer. Multi-modal anchoring for human-robot-interaction. *Robotics and Autonomous Systems, Special issue on Anchoring Symbols to Sensor Data in Single and Multiple Robot Systems*, 43(2–3):133–147, 2003.

[5] E. Gat. On three-layer architectures. In D. Kortenkamp, R. P. Bonasso, and R. Murphy, editors, *Artificial Intelligence and Mobile Robots: Case Studies of Successful Robot Systems*, pages 195–210. MIT Press, 1998.

[6] B. P. Gerkey, R. T. Vaughan, and A. Howard. The player/stage project: Tools for multi-robot and distributed sensor systems. In *Proc. Int. Conf. on Advanced Robotics*, pages 317–323, 2003.

[7] B. Graf, M. Hans, and R. D. Schraft. Care-O-bot II—Development of a next generation robotic home assistant. *Autonomous Robots*, 16(2):193–205, 2004.

[8] A. Haasch, S. Hohenner, S. Hüwel, M. Kleinehagenbrock, S. Lang, I. Toptsis, G. A. Fink, J. Fritsch, B. Wrede, and G. Sagerer. BIRON – The Bielefeld Robot Companion. In *Proc. Int. Workshop on Advances in Service Robotics*, pages 27–32, 2004.

[9] H. Hüttenrauch and K. Severinson Eklundh. Fetch-and-carry with CERO: Observations from a long-term user study with a service robot. In *Proc. IEEE Int. Workshop on Robot-Human Interactive Communication (ROMAN)*, pages 158–163. IEEE Press, 2002.

[10] B. J. A. Kröse, J. M. Porta, A. J. N. van Breemen, K. Crucq, M. Nuttin, and E. Demeester. Lino, the user-interface robot. In *European Symposium on Ambient Intelligence (EUSAI)*, pages 264–274, 2003.

[11] S. Lang, M. Kleinehagenbrock, S. Hohenner, J. Fritsch, G. A. Fink, and G. Sagerer. Providing the basis for human-robot-interaction: A multi-modal attention system for a mobile robot. In *Proc. Int. Conf. on Multimodal Interfaces*, pages 28–35. ACM Press, 2003.

[12] S. J. Leese. Microphone arrays. In G. M. Davis, editor, *Noise Reduction in Speech Applications*, pages 179–197. CRC Press, 2002.

[13] T. Matsui, H. Asoh, J. Fry, Y. Motomura, F. Asano, T. Kurita, I. Hara, and N. Otsu. Integrated natural spoken dialogue system of Jijo-2 mobile robot for office services. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 1278–1283, 1998.

[14] Y. Matsusaka, T. Tojo, and T. Kobayashi. Conversation robot participating in group conversation. *IEICE Trans. on Information and System*, E86-D(1):26–36, 2003.

[15] H. G. Okuno, K. Nakadai, and H. Kitano. Social interaction of humanoid robot based on audio-visual tracking. In *Proc. Int. Conf. on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE)*, pages 725–734, 2002.

[16] B. Souvignier, A.Kellner, B. Rueber, H. Schramm, and F. Seide. The thoughtful elephant: Strategies for spoken dialog systems. *IEEE Trans. on Speech and Audio Processing*, 8(1):51–62, 2000.

[17] T. Tojo, Y. Matsusaka, and T. Ishii. A conversational robot utilizing facial and body expressions. In *Proc. IEEE Int. Conf. on Systems, Man, and Cybernetics*, pages 858–863, Nashville, TN, 2000.

[18] I. Toptsis, S. Li, B. Wrede, and G. A. Fink. A multi-modal dialog system for a mobile robot. In *Proc. Int. Conf. on Spoken Language Processing*, 2004. to appear.

[19] S. Wachsmuth, G. A. Fink, and G. Sagerer. Integration of parsing and incremental speech recognition. In *Proc. European Conf. on Signal Processing*, volume 1, pages 371–375, Rhodes, Greece, 1998.

[20] S. Wrede, J. Fritsch, C. Bauckhage, and G. Sagerer. An XML based framework for cognitive vision architectures. In *Proc. Int. Conf. on Pattern Recognition*, Cambridge, UK, 2004. to appear.

A Multi-modal Dialog System for a Mobile Robot

Ioannis Toptsis, Shuyin Li, Britta Wrede,
Gernot A. Fink

Faculty of Technology, Bielefeld University
33594 Bielefeld, Germany
{itoptsis, shuyinli, bwrede, gernot}@techfak.uni-bielefeld.de

Abstract

A challenging domain for dialog systems is their use for the communication with robotic assistants. In contrast to the classical use of spoken language for information retrieval, on a mobile robot multi-modal dialogs and the dynamic interaction of the robot system with its environment have to be considered. In this paper we will present the dialog system developed for BIRON – the Bielefeld Robot Companion. The system is able to handle multi-modal dialogs by augmenting semantic interpretation structures derived from speech with hypotheses for additional modalities as e.g. speech-accompanying gestures. The architecture of the system is modular with the dialog manager being the central component. In order to be aware of the dynamic behavior of the robot itself, the possible states of the robot control system are integrated into the dialog model. For flexible use and easy configuration the communication between the individual modules as well as the declarative specification of the dialog model are encoded in XML. We will present example interactions with BIRON from the "home-tour" scenario defined within the COGNIRON project.

1. Introduction

In Human-Computer Interaction (HCI) the ultimate goal of research is to make the interaction with intelligent devices more "natural", i.e. intuitive and easy to use for humans. In human-human communication spoken language can be considered the most natural and effective means of communication, though it frequently is complemented by other modalities, e.g. mimic or gesture. Therefore, spoken language dialog systems are applied in many areas of HCI to achieve a natural communication.

The classical domain of dialogue systems are telephony-based services. Such systems mainly enable human users to access information stored in some database by using spoken language only. During the interaction the dialog system is in complete control of the information appliance.

A radically different and extremely challenging new domain for dialog systems is their use in so-called *robot companions* – mobile robots serving humans as assistants in private homes and eventually even as companions during everyday life. The communication with such complex devices can not be limited to spoken language only but has to take into account all modalities used in human-human dialogs, such as gesture or the expression of emotions. Furthermore, the robot's behavior is not only dependent on the communication with the user but also on the rather complex interaction of the mobile platform and its environment. Therefore, the dialog system can not be the central control unit of the robot companion. It will, however, be the central interfacing component between human users and the robot control system.

In this paper we will present the design of the dialog management system of BIRON – the Bielefeld Robot Companion [1]. It uses speech as the main modality for communication but is also able to augment information presented by spoken language with hypotheses derived from additional modalities, as e.g. in the case of speech accompanied by deictic gestures. As the dialog manager is not the central control unit of BIRON the internal state of the robot control system is periodically communicated with the dialog manager. Commands to the robot are derived from multi-modal semantic interpretation structures for dialog acts. Depending on the current state of the robot control unit the dialog manager can decide early about the possibility to perform actions required by the user or inform him about the internal state of the robot in case of communication problems.

The development of BIRON is currently focused on the scenarios defined within the COGNIRON project. One of the key experiments there is the so-called *home-tour*, where a robot companion is shown around a user's private home in order to familiarize it with this new environment.

In the following sections we will first review some related work on dialog systems with emphasis on systems used for the interaction with mobile robots. Then we will in detail describe the design of the dialog manager developed for BIRON covering the general architecture, the dialog model used, and the integration with the robot control system. In section 4 we will outline the capabilities of the current dialog model and present an example dialog with BIRON.

2. Related Work

The first generation of dialog systems, and also the majority of dialog systems today, only handle speech input since spoken language is the most important modality in human-human interaction. The dialog-system presented in [2] is applied to information retrieval tasks and employs a *slot-filling* strategy. A slot is an information item for which a value is required. The dialog system collects information from the user by filling slots to reach the dialog goal. This way, the system is able to support implicit verification of application responses, which reduces the duration of the dialog. The dialog model developed at AT&T [3] defines states and actions which is similar to our approach. However it employs a stochastic dialog strategy which can automatically be adapted by reinforcement learning. Also, the slot-filling technique is used to collect information for database inquiries as in [2]. The PHILIPS dialog system [4] is designed, among others, for portability. Therefore, it is application independent and based on a modular architecture like our

*The work described in this paper was partially conducted within the EU Integrated Project COGNIRON ("The Cognitive Companion") and was funded by the European Commission Division FP6-IST Future and Emerging Technologies under Contract FP6-002020. It was also partly funded by the German Research Foundation (DFG) within the Graduate Program 'Task Oriented Communication'.

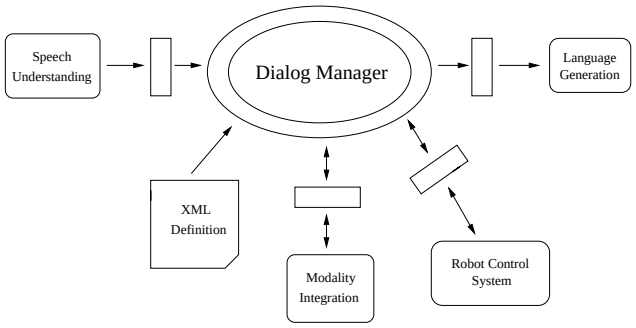

Figure 1: The dialog system of BIRON with the interface to the robot control system

system. For dialog control and speech understanding a definition language called *high-level dialogue description language* (HDDL) is used. With HDDL it is possible to divide the whole dialog into sub-dialogs, so called HDDL modules. This system is mainly used in automatic inquiry applications, where only spoken language is supported.

In the recent years the research of intelligent interfaces has focused on multi-modal dialog systems. The support of additional modalities enhances the robustness and the naturalness of the HCI. A representative of such interfaces is the MASK-kiosk [5]. It can handle multi-modal travel inquiries in form of spoken language and pointing on a touch screen. However, this kind of gesture can only approximate natural gesture used in human-human communication to a certain degree. In this system both modalities are fused on a semantic level inside the dialog-manager while in our system the fusion is achieved by a separate component. A multi-modal user registration system is presented in [6]. The dialog-manager contains states and actions and is similar to ours. But in this system the action to be taken does not depend on the state as in our system, but on the transition. Furthermore, the integration of the speech understanding and the modality fusion into the dialog-manager differs from our system. Information collected by different modalities is fused via a Bayesian network.

At present, only a small number of dialog systems supports intelligent human-machine interaction for mobile robots because of the higher complexity and dynamics of the task and the underlying system. The dialog system developed for an autonomous robot helicopter within the WITAS project [7] applies a combination of spoken language and pointing on a map. Its goal directed dialog strategy is not based on the slot-filing method and dialogs are open ended. The Hygeiorobot [8] is a mobile robotic assistant for hospital use. It can fulfill tasks like delivery of medicine or message and replying of inquiries of information about patients. Its uni-modal dialog system is state-based and designed to perform relatively short dialogs only. CARL is a mobile service robot [9], that is able to process input in form of spoken language and pointing gestures on a touch screen. Its system differs from ours in two points: First, their state-based and event-driven dialog-manager interprets user input via high-level reasoning. Second, the human-robot communication is modeled as an exchange of messages.

3. Dialog Manager

In the following, we first present the architecture of the dialog system developed for BIRON and then describe the dialog model in detail. We will close this section by emphasizing our system's capability of handling the internal robot states directly.

3.1. System Architecture

In many dialog systems the dialog manager is merged with other components, e.g. with speech understanding. This can lead to heavy dependencies of the dialog system on the application. We developed a modular architecture that separates the dialog management from speech processing as shown in Figure 1. The dialog manager is the main component of the dialog system and is also the focus of this paper. It communicates with other components over well defined interfaces, which use XML-structures for data exchange. This modular architecture of the dialog system enhances its portability.

The dialog manager receives the result of the semantic analysis of the speech input from the speech understanding component. In case that the semantic structure indicates the involvement of other modalities, the dialog manager will consult the modality integration component for further information. Consider the following example: The user says "This green cup" while pointing to it. The semantic structure delivered by the speech understanding contains anaphora "this" which indicates a possible involvement of gesture. The dialog manager then sends a request to the modality integration component to ask for integration of the semantic structure and the possible gestural information that can specify which object, in this case, which green cup, the user meant. Feedback to the user can be presented by the language generation module.

The dialog manager interprets the user's commands and sends them to the robot control system for execution. The robot control system is an independent component and can only process commands if the current status of the overall system allows it. Therefore, we implemented the control flow in a bidirectional way: The dialog manager sends user commands to the robot control and periodically receives messages from the robot control reporting its current status. Thus, the robot control system is not under control of the dialog system, but an equal "partner" of it.

3.2. Dialog Model

The model of the dialog manager is based on a *Finite State Machine* (FSM) that is extended with the ability of recursive activation of other FSMs and the execution of an action in each state. Actions that can be taken in certain states are specified in the *policy* of the dialog manager.

The implementation of the dialog manager is based on the so-called *slot-filling* strategy [2]. A slot is an information item for which a value is required. The task of the dialog manager is to fill enough slots to meet the dialog goal, which is defined as a goal state in the FSM. This can be viewed as a quantization of the semantic content of user's utterance into the required information items. Every state of the model is determined by the status of its slots. The slots can be empty, be filled with an attribute, or have logical values *true* or *false*. The incoming information from the user and the robot control system fills the slots, which are categorized into three sections and collected in a so-called dialog frames as shown in Figure 2. The USER section contains information provided by the user, the SYSTEM section represents the internal status of the robot control (see subsection 3.3 for details) and the CONTROL section contains items for internal use of the dialog manager.

The slot-filling technique alone is not powerful enough to support the complex interaction scenarios in robot domain [10]. To overcome this limitation we modeled the dialog in a modular way by dividing the dialog into sub-dialogs. Each sub-dialog is associated with a task and is modeled as a separate FSM. Each FSM has a goal state which indicates the completion of the current task. The processing of each sub-dialog can be interrupted by another sub-dialog, which enables alternated instruction processing. The interrupted sub-dialog can be resumed later.

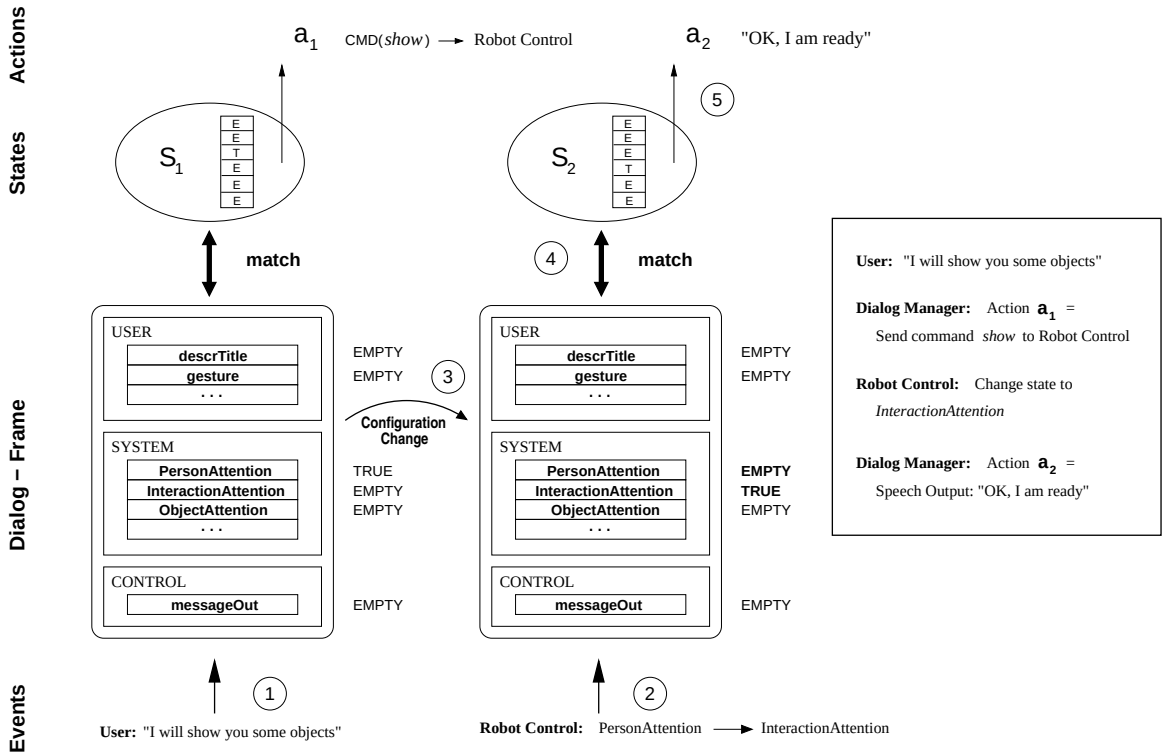

Figure 2: Dialog part with internal actions of the dialog manager and the structure of the dialog frame

The dialog management is event-based. Switching between the dialog states depends on the status composition of all slots in the dialog frame. In an ongoing dialog, the dialog manager compares slots in the newly updated dialog frame with those in the FSM to find out in which state the current dialog is. According to the specification of the state-action-association in the *policy* the appropriate action is executed, e.g., generation of speech output or sending events to the robot control system. We will illustrate this process with an example in subsection 3.3.

The dialog model is defined in a declarative definition language which is encoded in XML. This increases the portability of the dialog manager and allows an easier configuration and extension of the defined dialogs.

3.3. Integration of Internal Robot States

In robot applications it is important for the dialog manager to be informed about the current status of the sensory-motor system of the robot. This is often realized via message exchange in a multi-agent system in other applications [9]. Our approach is to integrate internal states of the robot control into the dialog model by representing the states of the robot control as an FSM. In an ongoing dialog, the current status of the robot control is represented as slots in the SYSTEM section of the corresponding dialog frame. Therefore, the dialog manager is permanently ”aware” of the status of the robot control.

In Figure 2 we demonstrate our approach with an example. Suppose the robot control system is in *PersonAttention* status, this means, that the robot is ready to start communication with the user. This status is represented as the slot PersonAttention in the SYSTEM section in the left dialog frame, its value is set to TRUE. The user speech input ”I will show you some objects” activates the sub-dialog ”show” and the result of its comparison with the current dialog frame is the state S_1. The associated action a_1 ”send command ’show’ to robot control” is then triggered as specified in the *policy*. After receiving this message the robot control system changes its status from *PersonAttention* to *InteractionAttention* which results in a change in the corresponding slots in the dialog frame’s SYSTEM section. After the match between this updated dialog frame with the sub-dialog ”show” the action a_2 is triggered. The robot generates the utterance ”OK, I’m ready!”.

The integration of the robot control states into the dialog model has several advantages. The dialog manager has dynamic knowledge about the abilities of the robot control system and can immediately make the decision if a certain user request can be processed or not without a transmission to the robot control. This reduces the reaction time of the robot. Another advantage is that the information about the task currently processed by the robot control system are available for the dialog manager. In case that the user tries to interrupt the current task the robot can give detailed information about the robot’s current status. This information can also be used to maintain the communication during long-term actions, e.g. by informing the user periodically about the current status of the task.

4. Scenario and Dialogs

Within the COGNIRON project we are currently implementing the home tour experiment. The central idea of this scenario is that a robot is delivered at home where the user familiarizes it with the environment by showing it different rooms and objects. During the home tour the robot should build internal representations of the environment and objects.

We have implemented five sub-dialogs for this scenario: (1) Greeting: the user logs into the system with common greeting phrases like “Hello”. The dialog manager sends the command “register” to the robot control system that changes its status from *Per-*

sonAlertness to *PersonAttention*. The robot then registers the user as an active communication partner and centers its focus on the user. (2) Parting: the user logs out of the system with common parting phrases like "Goodbye". The corresponding dialog manager command is "checkout" and the status of the robot control system is set back from *PersonAttention* to *PersonAlertness*. The robot returns to its standby mode. (3) Person following: the user can activate this function by saying "Please follow me". The dialog manager's command "follow" results in a status transition of the robot control system from *PersonAttention* to *PersonFollow* and the robot starts to follow the user. (4) Initiating gesture detection[1]: gesture detection can be triggered by user commands like "Look" or "I will show you some objects", which activate the dialog manager's command "show". This command changes the status of the robot control system from *PersonAttention* to *InteractionAttention* and the robot turns its camera to the direction of the user's hand. (5) Initiating object detection: The robot looks for the corresponding object in its current camera view if the user says, e.g., "This is a TV set". This process is initiated by the dialog manager's command "describe" and the following status transition of the robot control system from *InteractionAttention* to *ObjectAttention*.

In the following we illustrate the described procedures with a dialog example. (U: User; R: Robot, DM: dialog manager, RC: robot control)

U: Hello BIRON!
(DM: register, RC: PersonAlertness ⇒ PersonAttention)
R: Hello, what can I do for you?
U: Please follow me.
(DM: follow, RC: PersonAttention ⇒ PersonFollow)
R: OK, I'm following.
U: I will show you some objects.
(DM: show, RC: PersonAttention ⇒ InteractionAttention)
R: OK, I'm ready.
U: This is my TV set.
(DM: describe, RC: InteractionAttention ⇒ ObjectAttention)
R: OK, I can see it.
U: Thank you, BIRON, Good-bye.
(DM: checkout, RC: ObjectAttention ⇒ PersonAlertness)
R: Bye-bye.

As shown above, our system design can help to ensure smooth cooperation between the dialog manager and the robot control system and thus improve the robot's performance as a whole.

5. Conclusion

In this paper we presented the dialog system developed for the mobile robot BIRON. It assumes that speech is the main modality used for communication. However, the system is able to augment the semantic representations derived from user utterances by hypotheses for additional modalities as e.g. speech-accompanying gestures. The central component of the system is the dialog manager which communicates with its supporting modules via well defined interfaces using XML-encoded data structures. XML is also used for the declarative definition of the dialog model. As the dialog manager is not the central control unit of BIRON the internal states of the robot control system are periodically communicated and integrated into the current configuration of the dialog. In the current implementation a dialog model for the so-called "home-tour" scenario is defined[2].

[1] Currently, the gesture detection is not yet integrated in our system.

6. References

[1] S. Lang, M. Kleinehagenbrock, S. Hohenner, J. Fritsch, G. A. Fink, and G. Sagerer, "Providing the basis for human-robot-interaction: A multi-modal attention system for a mobile robot," in *Proceedings International Conference on Multimodal Interfaces*. Vancouver, Canada: ACM, November 2003, pp. 28–35.

[2] B. Souvignier, A.Kellner, B. Rueber, H. Schramm, and F. Seide, "The thoughtful elephant - strategies for spoken dialog systems," in *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 1, 2000, pp. 51–62.

[3] E. Levin, R. Pieraccini, and W. Eckert, "A stochastic model of human machine interaction for learning dialog strategies," in *IEEE Transactions on Speech and Audio Processing*, 2000.

[4] H. Aust and O. Schröer, "An overview of the PHILIPS dialog system," in *Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, Virginia, USA, 1998.

[5] J. Gauvain, S. Bennacef, L. Devillers, L. Lamel, and S. Rosset, "Spoken language component of the MASK kiosk," in *Human Comfort & Security of Information Systems*, K. Varghese and S. Pfleger, Eds. Springer, 1997, pp. 93–103.

[6] F. Huang, J. Yang, and A. Waibel, "Dialogue management for multimodal user registration," in *Proceedings International Conference on Spoken Language Processing*, Beijing, China, October 2000.

[7] O. Lemon, A. Bracy, A. Gruenstein, and S. Peters, "A multimodal dialogue system for human robot conversation," in *Proceedings North American Chapter of the Association for Computational Linguistics*, Pittsburgh, USA, June 2001.

[8] D. Spiliotopoulos, I. Androutsopoulos, and C. D. Spyropoulos, "Human-robot interaction based on spoken natural language dialogue," in *Proceedings of the European Workshop on Service and Humanoid Robots (ServiceRob '2001)*, Santorini, Greece, 25-27 June 2001.

[9] L. S. Lopes, A. Teixeira, M. Rodrigues, D. Gomes, C. Teixeira, L. Ferreira, P. Soares, J. Giro, and N. Snica, "Towards a personal robot with language interface," in *Proceedings European Conference on Speech Communication and Technology*, Geneva, Switzerland, September 2003.

[10] O. Lemon, A. Bracy, A. Gruenstein, and S. Peters, "The WITAS multi-modal dialogue system I," in *Proceedings European Conference on Speech Communication and Technology*, Aalborg, Denmark, 2001, pp. 1559–1562.

[2] A video of an example interaction with BIRON using German language can be found on our web site http://www.techfak.uni-bielefeld.de/ags/ai/projects/BIRON/.