Human-Robot Dialogue for Joint Construction Tasks

Mary Ellen Foster Tomas By Markus Rickert Alois Knoll
Robotics and Embedded Systems Group
Department of Informatics, Technical University of Munich
Boltzmannstraße 3, 85478 Garching, Germany
{foster,by,rickert,knoll}@in.tum.de

ABSTRACT

We describe a human-robot dialogue system that allows a human to collaborate with a robot agent on assembling construction toys. The human and the robot are fully equal peers in the interaction, rather than simply partners, and joint action is supported at all stages of the interaction: the participants agree on a construction task, jointly decide how to proceed to proceed with the task, and also implement the selected plans jointly. The symmetry between participants provides novel challenges for a dialogue system, and also makes it possible for findings from human-human joint-action dialogues to be easily implemented and tested.

Categories and Subject Descriptors: H.5.2 [User Interfaces]: Natural language; I.2.9 [Robotics]: Operator interfaces

General Terms: Human Factors

Keywords: Human-robot interaction, multimodal dialogue systems

1. INTRODUCTION

When humans and robots work together, there are two extremes for initiative in the interaction. On the one hand, the robot can be seen as a tool for accomplishing specific tasks. When this view is taken, the obvious human-robot interface is *teleoperation*: the human gives commands, and the robot carries out the requested actions. On the other hand, a robot may work largely autonomously, with a human acting as a supervisor, intervening only when things go wrong. For many robot applications, interfaces such as these are sufficient; however, as the tasks to be performed by robots grow more complex, a more interactive interface, dividing the initiative between the robot and the human and incorporating communication techniques such as natural-language dialogue, can provide a more intuitive and flexible means of coordinating human-robot activity.

Several recent systems have been developed that permit natural-language human-robot interaction. These include both task-based systems like the NASA peer-to-peer human-robot system [5], as well as embodied, more social systems such as Mel [13] and Leonardo [2]. In many of these systems, the domain or conversational roles of the human and the robot system are predefined and distinct, and while there is the possibility for joint or collaborative activity, it is constrained to specific tasks—e.g., the user may help the robot's vision system to identify the correct target, or the robot may instruct the user in performing a task.

In this paper, we describe a human-robot dialogue system where the robot and human are fully equal peers in performing a task: either participant may drive the selection of the goals to address and the strategies to be used to address them, and either may also perform any part of the task. In this type of interaction, *joint action* is a central concept in two ways: the participants jointly address the domain task, and the dialogue itself is also a form of joint action, as pointed out by Clark [4].

The paper is structured as follows. First, in Section 2, we describe the goals of the JAST project and present the human-robot dialogue system being developed as part of that project. Next, in Section 3, we give technical details of the architecture and components of the current dialogue system. In Section 4, we then compare the JAST system to other similar systems. Finally, in Section 5, we outline the plans for the JAST system and draw some conclusions.

2. HUMAN-ROBOT DIALOGUE IN JAST

The overall goal of the JAST project ("**J**oint **A**ction **S**cience and **T**echnology") is to investigate the cognitive and communicative aspects of jointly-acting agents, both human and artificial. The human-robot dialogue system being built as part of the project is designed as a platform to integrate the project's empirical findings on cognition and dialogue with its work on autonomous robots, by supporting symmetrical, multimodal human-robot collaboration on a joint construction task.

The robot (Figure 1) consists of a pair of mechanical arms with grippers, mounted in a position to resemble human arms, and an animatronic talking head capable of producing facial expressions, rigid head motion, and lip-synchronised synthesised speech. The input channels are speech recognition, object recognition, robot sensors, and face tracking; the outputs include synthesised speech, head motions, and robot actions. The user and the robot work jointly to assemble a Baufix wooden construction toy (Figure 2) on a common work area, coordinating their actions through speech, gestures, and facial motions. Joint action can take several

ICMI'06, November 2–4, 2006, Banff, Alberta, Canada.

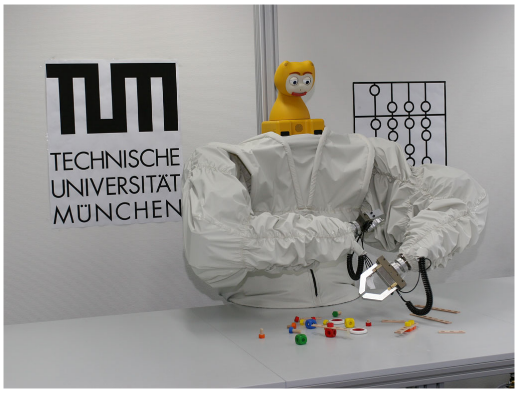

Figure 1: JAST construction robot

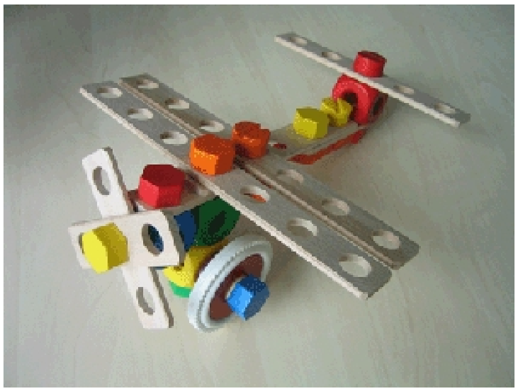

Figure 2: Assembled Baufix airplane

forms: for example, the robot may ask the user to provide assistance by holding one part of a larger assembly, or by assembling or disassembling components. In the current version of the system, the robot is able to manipulate objects in the workspace and to perform simple assembly tasks.

The sample interaction in Figure 3 shows the user and the system cooperating to build the tail-section component of the airplane shown in Figure 2, first combining a bolt and a slat and then securing the assembly with a cube. In this interaction, both agents know the target construction, and the majority of the interaction deals with determining the allocation of subtasks and coordinating the execution of those tasks. Interactions are also possible in which only one participant—either the robot or the user—knows how to build the target, and must guide the other; such interactions contain more direct commands such as *Pick up the red cube*, in addition to task-allocation dialogue as in the example.

JAST: Welcome to JAST. Would you like to build a tail section?

User: Okay.

JAST: Can you take care of the bolt and slat?

User: Sure. *[Picks up a red bolt and a five-hole slat]*

JAST: *[Picks up a red cube]* Tell me when you are done.

User: *[Puts bolt through slat]* I'm done.

JAST: *[Gives user the cube]* Here is the cube.

User: *[Takes the cube]* Thanks! *[Screws cube onto bolt]*

Figure 3: Sample interaction

3. DETAILS OF THE JAST SYSTEM

3.1 System architecture

Figure 4 shows the set of modules that make up the JAST system;[1] the highlighted modules are those responsible for maintaining the system state. Input comes from several channels: face tracking, object recognition, hand tracking, speech recognition and processing, and robot sensors. Input from all of these sources is analysed and processed by the central processing components, which choose appropriate responses to the input and maintain the system state. The high-level response specification is passed to the output-generation component, which sends module-specific commands to the talking-head and robot control modules. Communication among all components of the system is implemented using Ice, the Internet Communications Engine [6], which supports communication among a population of heterogenous agents implemented in multiple languages and running on multiple computers. In the remainder of this section, we give more details on the system components that are implemented in the current prototype.

3.2 Input processing

The speech-recognition module is based on the Dragon NaturallySpeaking speech recogniser [11]. The microphone is always enabled, allowing the user to give input to the system at any point. The recognised speech is parsed into logical forms by the syntactic processing module, which is based on the OpenCCG natural-language-processing library [17]. Object recognition and face recognition are both implemented using the OpenCV computer-vision library [8]. The object recogniser provides information about the location, type, size, colour, and orientation of all of the objects in the common workspace, using geometric and pattern invariants. The face recogniser uses the OpenCV implementation of the Viola-Jones face detector [16] to locate the user's face so that the system can look at them when appropriate.

The sensor buffer and history module records the physical activity of the system and passes on data to the appropriate processing modules. All physical events, such as speech or physical movements, are recorded, both those detected by the input modules and those performed by the robot itself. The events are indexed by their start and end times, as well as by additional information such as the specific objects involved. This enables the semantic interpretation and central decision-making modules to easily retrieve all activity at a specific time, or a complete history for a specific object.

3.3 System state and behaviour selection

The central components of the JAST system process messages from the input-processing modules and send command specifications to the output generator. This part of the system consists of several modules responsible for maintaining and updating various aspects of the system state, as well as two main processing modules: the semantic interpretation

[1] Note that not all of the components shown in the figure are fully implemented in the current version of the system.

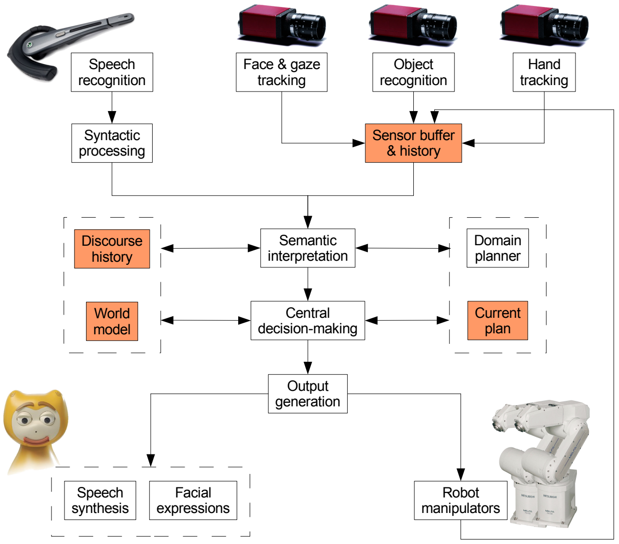

Figure 4: JAST system architecture

module, and the central decision-making module.

The discourse history keeps a record of the contributions made by both participants to the interaction, at a semantic level; specific physical details can be retrieved from the sensor buffer and history when necessary. The world model module contains a continually updated record of the physical objects (i.e., Baufix pieces) in the world, and any assembled components. For example, this module would record the fact that two pieces are screwed together, but not the angle between the pieces. For words such as "tail", "wing", and "airplane"—which denote not single Baufix objects, but rather collections of them in particular configurations—the world model also stores and may even update the definition of these words.

The domain-planner module uses the JSHOP2 planner [7] to create action sequences to realise Baufix construction tasks. A desired construction is specified as a set of subassemblies (i.e., bolts inserted through and screwed into specific other parts). When a new assembly target is selected, the planner create a plan to achieve that goal; the current plan is then used to guide the interaction, both by specifying the next steps that the system should take and by helping to recognise and interpret the user's inputs. The JAST system is also able to work in a purely reactive mode, where only the user has full knowledge of the intended assembly; in that case, the planner creates a sequence of plan fragments to respond to user requests as they are made.

The semantic interpretation module receives the parser output (logical forms) and makes calls to the sensor history, dialogue history, and world model as needed to find referents for noun phrases, including demonstrative and anaphoric pronouns, and to identify the domain state or action that verbs refer to. The output of this module consists of the type of the utterance and the disambiguated propositional content. The possible utterance types include requests for information or action, statements answering a question, proposal of a new goal, and dialogue feedback such as approval, confirmation, and denial.

Based on the output of the semantic interpretation, the central decision-making module selects the appropriate system response to user actions and utterances. Processing in this module is based on Blaylock and Allen's [1] collaborative problem-solving model of dialogue, which models dialogue as an iterative, interleaved process of jointly selecting the goals to address, choosing procedures for achieving the goals, and executing the selected procedures. In response to a user input, this module updates the system state as necessary—for example, by updating the current position in the plan or by marking a query or request as resolved—and sends specifications of appropriate feedback and robot actions to the output-generation module.

3.4 Output generation

There are two main forms of output in the JAST system: feedback from the talking head, and actions from the robot arms. The output-generation module converts high-level specifications from central components into concrete command sequences for both of these channels, and also monitors the command execution to ensure the output is coordinated. The output is generated incrementally and can be interrupted by the user; only when an output segment is completed is it used to update the system state.

The talking head is a Philips iCat animatronic head [15], which supports three parallel forms of output: gaze control, lip-synchronised synthesised speech, and facial expressions. As with Mel the penguin [13], gaze control is used for immediate feedback: the talking head looks at objects on the table as it manipulates them, and looks at the user when it addresses them directly. The content of the speech is generated by creating logical forms that are translated into text by the OpenCCG realiser, using the same basic grammar as the syntactic processing module. The facial expressions are determined based on the content of the speech—for example, the head might indicate agreement by smiling, nodding, and saying *Okay*.

The robot behaviours are implemented through action primitives [10]: parameterised, chainable motion specifications that divide actions into subtasks. For example, grasping an object on the table is implemented by the following sequence of primitives: move to the object location, open the gripper, move slowly downward until the table is reached, move up slightly, close the gripper, and move upwards. Feedback from multiple sensors, including cameras, force/torque sensors, and robot encoders, is used to control the execution of the primitives; these sensors also provide an additional input modality for the system.

4. RELATED WORK

The JAST system is partly based on previous work [9] carried out at the University of Bielefeld between 1994 and 1999 as part of a project on situated artificial communicators. That previous system also addressed the same scenario as the JAST system—joint human-robot construction of Baufix models. The current system updates and extends that system with more powerful input-processing, dialogue-management, and output-generation capabilities to create a robot that can truly collaborate with a human user.

Several more recent systems aim to allow humans and robots to collaborate on various tasks, communicating via natural-language dialogue. The system that is most similar to JAST is Leonardo [2], a fully-embodied humanoid robot with social skills that allow it to learn and collaborate effectively in human settings. Leonardo is able to respond to requests from a human, to learn both the names of objects and new procedures, and to execute learned procedures in collaboration with the user. There are two main differences between JAST and Leonardo. First, while Leonardo can understand spoken commands, it communicates only through behaviour and non-verbal signals such as body language and facial expressions; for JAST, verbal communication is more central. Also, Leonardo's tasks consist of pushing buttons in learned sequences, while the JAST tasks require much more dexterity from the robot.

In most other human-robot dialogue systems, the roles of the agents are less symmetrical. The the NASA peer-to-peer HRI system [5] treats robots and humans as peers in that they have balanced work roles, but they still have distinctive capabilities and roles. The main form of collaboration in this system occurs when one agent must ask another for help in dealing with a situation—e.g., confirming that a recognised object is of the intended type or requesting assistance with performing a task. Mel the robotic penguin [13] acts as a host, guiding the user as they explore a lab; Mel cannot itself carry out any physical. The main goal for Mel is to create social engagement in a conversation; collaboration takes place when the robot instructs the user on using the equipment in the lab. The emphasis in the Coyote system [14] is on a model of visual perspective-taking and spatial reasoning used to resolve ambiguous requests and to handle handle multiple spatial reference frames that are typical when agents collaborate on a physical task.

Dialogue with animated on-screen agents has also been extensively studied in recent years—see [3] for a survey. Although these systems generally deal with information-seeking rather than physical tasks, they provide relevant data on non-verbal communicative cues that can be used to inform the design of embodied robotic conversational agents.

5. FUTURE WORK AND CONCLUSIONS

The JAST project is ongoing, and the dialogue system described here is still under development; we intend to enhance its abilities and robustness in all areas. The speech-recognition and speech-processing modules will be extended to deal with more possible inputs from the user, and to process ungrammatical and badly-understood contributions. We aim to incorporate real-time object tracking [12] to follow both the objects and the user's hands; similarly, we intend to add gaze-direction tracking to the current face-tracking module. For the dialogue manager, we aim to support more complex interaction patterns, including imperfect understanding, failed actions, and revisiting previously-made decisions. The range of spoken outputs, facial expressions, and robot skill primitives will also be extended.

In addition to the technical improvements to the individual modules listed above, we will also extend the capabilities of the system as a whole, using findings from human studies. Other JAST partners are currently analysing human-human dialogues in the same joint-action construction domain as the system uses. We will take advantage of the fully symmetrical roles in the human-robot dialogue system to implement and test interaction strategies derived from these dialogues; like Trafton et al. [14], we also hope that heuristics derived from human-human interaction will provide good building blocks for human-robot interaction systems.

In conclusion, the JAST human-robot dialogue system provides an implementation of fully symmetric, collaborative, natural-language human-robot interaction. It differs from other human-robot dialogue systems in that the participants are fully equal peers, with completely interchangeable roles, and the system supports joint action at all levels: both dialogue *in support of* joint action and dialogue *as a form of* joint action. This symmetry provides novel challenges for a dialogue system, and also makes it possible to implement and test interaction models based directly on human-human joint-action dialogues.

6. ACKNOWLEDGEMENTS

This work was supported by the EU FP6 IST Cognitive Systems Integrated Project "JAST" (FP6-003747-IP), http://www.euprojects-jast.net/. Thanks to the ICMI reviewers for useful comments and suggestions.

7. REFERENCES

[1] N. Blaylock and J. Allen. A collaborative problem-solving model of dialogue. In L. Dybkjær and W. Minker, editors, *Proceedings, 6th SIGdial Workshop on Discourse and Dialogue*, pages 200–211, 2005.

[2] C. Breazeal, A. Brooks, J. Gray, G. Hoffman, C. Kidd, H. Lee, J. Lieberman, A. Lockerd, and D. Chilongo. Tutelage and collaboration for humanoid robots. *International Journal of Humanoid Robotics*, 1(2):315–348, 2004.

[3] J. Cassell, J. Sullivan, S. Prevost, and E. Churchill. *Embodied Conversational Agents*. MIT Press, 2000.

[4] H. H. Clark. *Using Language*. Cambridge University Press, 1996.

[5] T. Fong, C. Kunz, L. M. Hiatt, and M. Bugajska. The human-robot interaction operating system. In *Proceedings, 1st Annual Conference on Human-Robot Interaction (HRI 2006)*, 2006.

[6] M. Henning and M. Spruiell. Distributed Programming with Ice. http://www.zeroc.com/download/Ice/3.1/Ice-3.1.0.pdf. Revision 3.1.0, July 2006.

[7] O. Ilghami and D. S. Nau. A general approach to synthesize problem-specific planners. Technical Report CS-TR-4597, UMIACS-TR-2004-40, University of Maryland, October 2003.

[8] Intel Corporation. Open Source Computer Vision Laboratory. http://www.intel.com/technology/computing/opencv/.

[9] A. Knoll. A basic system for multimodal robot instruction. In P. Kühnlein, H. Rieser, and H. Zeevat, editors, *Perspectives on Dialogue in the New Millennium*, volume 114 of *Pragmatics & Beyond New Series*, pages 215–228. John Benjamins, 2003.

[10] J. D. Morrow and P. K. Khosla. Manipulation task primitives for composing robot skills. In *Proceedings, IEEE International Conference on Robotics and Automation*, pages 3354–3359, 1997.

[11] Nuance Communications. Dragon NaturallySpeaking. http://www.nuance.com/naturallyspeaking/.

[12] G. Panin, A. Ladikos, and A. Knoll. An efficient and robust real-time contour tracking system. In *Proceedings, IEEE International Conference on Computer Vision Systems*, 2006.

[13] C. L. Sidner and M. Dzikovska. A first experiment in engagement for human-robot interaction in hosting activities. In N. O. Bernsen, L. Dybkjær, and J. van Kuppevelt, editors, *Advances in Natural Multimodal Dialogue Systems*. Springer, 2005.

[14] J. G. Trafton, A. C. Schultz, N. L. Cassimatis, L. M. Hiatt, D. Perzanowski, D. P. Brock, M. D. Bugajska, and W. Adams. Communicating and collaborating with robotic agents. In R. Sun, editor, *Cognition and MultiAgent Interaction: From Cognitive Modeling to Social Simulation*. Cambridge University Press, 2006.

[15] A. J. N. van Breemen. iCat: Experimenting with animabotics. In *Proceedings, AISB 2005 Creative Robotics Symposium*, 2005.

[16] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, May 2004.

[17] M. White. Efficient realization of coordinate structures in combinatory categorial grammar. *Research on Language and Computation*, 4(1):39–75, June 2006.