A Dialogue Manager for Multimodal Human-Robot Interaction and Learning of a Humanoid Robot

Hartwig Holzapfel
Interactive Systems Labs
Universität Karlsruhe (TH)
hartwig@ira.uka.de

Purpose – This paper gives an overview of our dialogue manager and recent experiments with multimodal human-robot dialogues. It identifies requirements and solutions in the design of a human-robot interface. The paper presents essential techniques for a humanoid robot in a household environment, including interaction techniques to control a humanoid robot using speech, as well as more complex tasks with more profound dialogues. Along with these techniques the paper describes their application to representative interaction scenarios that are based on standard situations for a humanoid robot in a household environment.

Design/methodology/approach – The presented dialogue manager has been developed within the German collaborative research center SFB-588 on "Humanoid Robots – Learning and Cooperating Multimodal Robots". The dialogue system is embedded in a multimodal perceptual system of the humanoid robot developed within this project. The implementation of the dialogue manager is geared to requirements found in the explored scenarios. The algorithms include multimodal fusion, reinforcement learning, knowledge acquisition and tight coupling of dialogue manager and speech recognition.

Findings – Within the presented scenarios several algorithms have been implemented and show improvements of the interactions. Results are reported within scenarios that model typical household situations.

Research limitations/implications – Additional scenarios need to be explored especially in real-world (out of the lab) experiments.

Practical implications – The paper includes implications for the development of humanoid robots and human-robot interaction.

Originality/value – This paper explores human-robot interaction scenarios and describes solutions for dialogue systems.

Keywords – Human-robot interaction, Multimodal dialogue

Article Type – Research Paper

in Industrial Robots Journal, Emerald Publishing, 2008, to appear

1 Introduction

Recently there has been growing interest in the development of humanoid robots. With humanoid robots, new types of robots are built that are neither controlled by a simple remote control nor execute static pre-programmed activities. They participate in our every-day life and need to interact with humans in a style that seems natural to them.

For exactly this purpose it is important that the robot can understand and communicate with humans by natural means, i.e. speech, language, gestures and dialogue. Natural human-robot interaction includes speech as a main modality, but exceeds purely spoken communication and includes additional information channels such as visual processing. Vision for example allows the robot to understand who is talking to the system, recognize the state or mood a person is in, perceive objects a person is talking about, or recognize pointing gestures by the human. These multimodal perceptions are interpreted by a dialogue manager. It interprets given information in context and reacts by responding or executing specific actions. In this paper we present our dialogue manager Tapas which enables a robot to communicate with humans using multimodal information. It allows humans to talk to the robot, command the robot, and give orders for task execution. It furthermore enables the robot to acquire new information such as introducing unknown objects or persons.

Tapas is a dialogue tools collection implemented in Java. It provides algorithms for dialogue processing, including natural language understanding, discourse and context modelling and dialogue strategies, for development, evaluation and runtime systems. The dialogue system presented here has been implemented over the past few years – within the German collaborative research center SFB-588 on "Humanoid Robots – Learning and Cooperating Multimodal Robots" – with the goal to discover requirements for such a system, and to provide a tool to improve human-robot communication. Research within this project is based on the robot Armar 3 which is developed in Karlsruhe within SFB 588. A picture of the robot in the kitchen environment can be seen in figure 1. For a detailed description of the ARMAR platform, we refer to (Asfour et al. 2006).

Most traditional dialogue management approaches consider speech-only interactions. Dialogue management for a humanoid robot is a fairly complex task and requires interplay of many components. Many robots use command-based speech input or simple dialog control. Some dialogue systems for robots are based on finite-state automata e.g. the robots HERMES and BIRON (Bischoff and Graefe 2002, Toptsis et al. 2004). In many scenarios this is sufficient. The assumption is that in complex environments these models are not adequate, e.g. to deal with speech recognition errors, process multimodal information, and handle the manifold contextual states.

More advanced approaches are implemented e.g. for the robot Pearl (Montemerlo et al. 2002), which uses a probabilistic approach to cope with recognition errors, or in the dialogue system WITAS for unmanned vehicle control (Lemon et al. 2001), which adopts the information state update (ISU) approach.

So far, performance of different systems for human-robot interaction is hard to compare since most systems are designed for different tasks and evaluation tools are missing to compare complete systems. Therefore we compare key technologies of our system within the respective field.

To understand which technologies are relevant for human-robot interaction, we have explored different situations and requirements from humans, during the course of our research project. These requirements span from very simple control interactions to complex dialogues where the robot needs to identify and grasp objects or acquire new information. To accomplish these complex tasks, detailed information needs to be provided through dialogue.

In the following the paper describes different scenarios, associated dialogue implementations and experiments, which demonstrate representative situations for human-robot interaction and exhibit main properties of human-robot interaction.

in Industrial Robots Journal, Emerald Publishing, 2008, to appear

2 Control and Interaction

The most basic functionality for successful interaction is controlling the robot with multimodal commands. These commands provide a basis for more complex scenarios. We describe the architecture of the dialogue manager in the context of this control scenario and extend the description with the following sections.
The control and interaction functionality required by a humanoid robot exceeds the standard command-and-control metaphor. The idea of the basic command-and-control is to utter one command which is then executed by the robot. This idea however, is idealistic, since it doesn't consider recognition errors, uncertainties and ambiguities, which need to be handled by the system. Here, a dialogue manager is required, which includes context tracking, clarification and confirmation questions to conduct successful control of the robot.
Natural communication is an additional requirement for humanoid robots. To be more natural, other modalities than speech need to be considered. Multimodal recognition can be used to understand information that isn't expressed verbally or to provide more robust recognition technologies by processing redundant information.

2.1 Dialogue Manager Architecture

Figure 2 gives an overview over the components in the dialogue manager's architecture. The dialogue manager is embedded in a perceptual system with speech recognition and visual processing for person tracking, identification and gesture recognition.
It receives inputs from several recognition components. These are interpreted and converted to semantic representations. Semantics are then interpreted again in the system's context and update the discourse model. A dialogue state abstraction represents the progress made in the dialogue and provides a decision basis for the dialogue strategy, which selects the next dialogue move. A dialogue move includes actions such as spoken or multimodal output or sending commands to the robotic platform. It also updates the context model with expectations about next user actions. Most parts of the dialogue design is language independent. By clear separation of speech grammars and dialogue management, the system can be designed language independently and be employed for different languages (Holzapfel 2005).
Recognition and processing of dialogue goals, multimodal fusion and dialogue strategies is described in the next sections. A system overview including details about the speech recognizer, person tracking and gesture recognition, can be found in (Stiefelhagen et al. 2004).

2.2 Robot Control by Spoken Interaction

Over the past few years we have explored the kitchen domain for human-robot interaction. Within this domain, dialogue scenarios for simple tasks include information handling such as providing recipe information for the user, requests to bring or manipulate objects, and execute tasks which require parameters which are clarified in dialogue. Basic functionality of the robot can also be controlled by speech commands, such as open/close hands, look into a specified direction, move to a given position.
A basic principle is that communication is task and goal-oriented. To achieve a goal information must be given which is required by the system to execute this goal. The implementation of dialogue goals and dialogue strategies follow the Ariadne architecture (Denecke 2002) which uses information-based preconditions and executes actions bound to a goal when all information for a goal is accumulated. The dialogue strategy decides which goal to follow and clarifies missing information to achieve a goal.
Our current system uses 98 dialogue goals in the kitchen environment. Out of these 98 goals, 55 goals control the robotic platform and 43 goals cover kitchen tasks and social interaction. The following code shows an example how required information is defined. It represents the dialogue goal to initialize the left or right arm.

```
InitializeArmConfirm
OBJ  [ obj_arm ]
```

LEFTRIGHT [prp_side]
CONFIRM [prp_conf]

One parameter defines which arm to initialize, a second parameter defines whether to initialize the right or the left arm, and the third parameter defines a confirmation slot to execute the associated action. These parameters can be given within one utterance, e.g. "initialize the right arm", or be collected over multiple utterances. If parameters are missing, they are requested by the system, e.g. "which arm do you want me to initialize?".
The confirmation parameter represents the confidence of the system in executing the right action, and is used by the strategy to prompt for explicit confirmation before executing an action that might be dangerous for the environment or harmful for the system itself.
The implementation of the robot control metaphor is not always trivial. Interaction with the system often is error-prone because of insufficient capabilities of the robot but also because of the human due to limited understanding of the robot's capabilities (Gieselmann and Stenneken 2006). Error sensitive strategies significantly improve dialogue performance (Gieselmann and Ostendorf 2007).

2.3 Multimodal Fusion

A multimodal command in its classic form follows Bolt's scheme of "put that there" (Bolt 1980). This example addresses multimodal fusion of speech with two pointing gestures, also called deictic gestures. Different approaches have since been presented that address multimodal fusion and interpretation of speech and gestures (Oviatt et al. 1997, Johnston 1998). Natural communication with a humanoid robot includes 3D pointing gestures, e.g. (Corradini et al. 2002, Nickel and Stiefelhagen 2003), which have different characteristics compared to pen input, e.g. much lower detection accuracy of gestures.
In (Holzapfel et al. 2004) we present robust multimodal fusion for speech and 3D pointing (deictic) gestures. It extends (Johnston 1998) and provides a fusion which is robust against false detections and exploits n-best lists of pointing references.
Deictic gestures can be used to point at an object (nonverbal interaction) and thus giving either redundant information which helps the system to improve its recognition accuracy or delivers additional information not given by speech. One example is "bring me the blue cup (over there)". For correct understanding, the system needs to recognize the spoken utterance, recognize the gesture, resolve the object the person points to and merge the information with information from speech recognition. If there is more than one blue cup, the object can be identified from gesture information. Speech and deixis can be considered to be more or less synchronous or are at least tightly coupled with the speech signal. Fusion of these two modalities can thus be handled as input fusion in the dialogue manager, i.e. before discourse updates are performed.
The fusion algorithm operates on a pool of input events and checks for matching events by applying a set of constrained-based fusion rules with n-best list processing. N-best resolution is crucial in our approach since the pointing gesture is often not specific enough to resolve the object correctly only by pointing. Events remain in the pool if they aren't covered by fusion rules. After a predefined timeout they are abandoned. Figure 3 illustrates the fusion process and shows an example for merging information from one speech and one gesture event.
In contrast to related work referenced above, the presented approach is robust against false detections of gestures. In our setup we have observed 87% recall for detecting gestures. However out of all reported gestures, only 47% were gestures actually performed by the user. The remaining falsely detected gestures could be sorted out almost completely by the fusion algorithm, due to statistical correlation in time. Out of a total of 102 multimodal user inputs 74% were correct. About half of the failed attempts were due to missing gesture detection (which was 87% on all gestures), 22% of the errors were caused by bad recognition of pointing direction, 17% of the errors were caused by speech recognition errors, 2% incorrect fusion, and 7% other errors were

observed. The fusion algorithm thus shows nearly optimal behaviour. Remaining errors, e.g. speech recognition errors, are left for correction in dialogue.

2.4 Tight coupling with recognition components

The approach presented above could suggest independent sequential processing of speech recognition, natural language understanding and dialogue management. However, valuable information is given by the dialogue state that can be used to improve speech recognition in dialogue context. This is achieved by tight coupling between speech recognizer and dialogue manager in sharing as much information between these two components as possible (Fügen et al. 2004, Holzapfel and Waibel 2006). Sharing linguistic knowledge sources, i.e. recognition and understanding grammars, improves processing speed and robustness, less knowledge sources need to be maintained. Improvements in recognition accuracy over loose coupling could be observed especially for contextual utterances and distant speech recognition. Here, contextual information plays a more important role to distinguish acoustic signals. Contextual utterances are for example typical question-reply pairs. In contrast to existing work, the approach is generic in a way that for a new system no training is required to enable contextual weighting. Rather, grammar weights are determined by automatically mapping expected information types to grammar rules using ontological information.
Experiments in the barkeeper scenario (described in the following) with generic contextual weighting show improvements of 33% (relative) on close-speech and 21% (relative) on distant-speech recordings. Recognition rates have been measured on semantic concepts at 5.2% error rate for close-speech and 15.7% error rate for distant-speech. The gain of employing this technology is significant and has been employed for the following complex scenarios.

3 Complex Interaction Scenarios

Besides simple control and interaction techniques, complex scenarios show additional requirements for dialogue abilities and more complex dialogue technologies. Here we present such technology in two scenarios, which are first steps to create real-world scenarios. The first scenario is a barkeeper scenario. It models the task to have the robot serve a selected object as requested by a human user. This scenario shows application of multimodal integration and application of optimization techniques to dialogue strategies for robust processing including clarification questions and recognition errors. The second scenario describes a robot receptionist. Here the dialogue system is used to engage in a conversation with persons and conduct a receptionist dialogue. It exhibits another important aspect of an autonomous system, which is its ability to learn autonomously, here, by acquiring information about previously unknown persons.

3.1 Barkeeper Scenario

The barkeeper scenario originates from a task which a household robot fulfils, namely to bring a selected object to a human. To do so, the system needs to clarify which objects the user refers to, which is done in a dialogue with the user.
We simulate this task in a setup where both the robot and the user can look at a table, which contains a number of objects. The user can now tell the robot which object he would like to get using speech and pointing gestures. The task of the robot is to find the right object and use speech and pointing himself to confirm his current belief. The problem that needs to be solved by the dialogue manger is to find an optimal strategy to confirm recognized user input, correct (repair) information and confirm the object selection with as few turns as possible. In the future we plan to extend this scenario with detection and learning of unknown objects.
To develop a dialogue strategy for this scenario we have first conducted a Wizard-of-Oz study. A Wizard-of-Oz study is frequently applied in the design of dialogue systems. It includes a data collection with real users, where the system's decision making is taken

over by a human, the "wizard". Based on this data we have created a dialogue strategy with handwritten dialogue moves. This rule based approach is the standard way to define a dialogue strategy. A dialogue move uses preconditions based on the abstract dialogue state. The strategy selects the best matching move and executes its actions. One shortcoming of handcrafting dialogue strategies is that it is a time-consuming and non-trivial task, especially with increasing complexity of the dialogue. Additional problems are robustness on unseen data. One promising approach to avoid these problems is to use collected dialogues for automatic training of dialogue strategies. For this task especially reinforcement learning has become popular. So far, reinforcement learning has successfully been applied in a couple of dialogue scenarios (Singh et al. 1999, Levin et al. 1998, Levin et al. 2000, Walker 2000). More recently, there have also been approaches to training dialogue strategies with a user simulation, which allows to generate a vast number of dialogues (Scheffler and Young 2002, Schatzmann et al. 2005, Williams and Young 2003). For training and evaluation of such optimization algorithms, we have chosen a setup with 20 objects (plates, cups, bottles) which are located on the table between the robot and the human. The objects have three different colours, red, blue and yellow. The maximum number of identical objects was 5, which differ only in their location on the table.
An example of a dialogue in one turn is "give me the red cup over there (+ pointing)" which is correctly understood and the gesture can be resolved correctly. Other interactions start with the user only asking for a cup and then redefining their selection. The system also needs to take into account speech or gesture detection errors and inaccuracy of pointing resolution. Thus, the strategy needs to confirm the selected object, e.g. "did you mean this object", or confirm and repair specific information, e.g. "did you say you want a red cup?".
Which actions are selected in each situation should be learned by the system and depends on the reward function, which defines the evaluation criteria for each dialogue. The reward function was created from feedback from the Wizard-of-Oz experiments taking into account objective measures such as dialogue length and dialogue success and subjective measures, e.g. people did not like if the system repeats a question.

For strategy training we have created a user simulation from the Wizard-of-Oz data using bigram-statistics of specialized speech act types plus a stochastic error model based on Word-Error-Rates for speech and pointing accuracy for gestures.
The 6-step process of training the dialogue strategy is illustrated in figure 4 and described in the following. After the Wizard-of-Oz experiment (step 1) the user simulation and error models are trained (step 2) from collected data. Then the reinforcement learning setup is defined which includes a reward function and the Markov Decision Process (MDP) which models the state transition and observations of the learner (step 3). Different training runs can then be started with different reinforcement learning configurations (step 4). The best strategy is then deployed to the runtime system and evaluated with real users (step 5). During runtime of the system new data is collected, from which the data corpus and models are updated (step 6). Figure 5 illustrates operation of the reinforcement learning agent in the simulation environment, which corresponds to step 4. It shows that the actions of the agent influence the user model and that the agent receives only reward and state information (implicitly including user actions) from the environment.
In order to empirically validate the benefit of our overall approach, we compared the performance of the handcrafted baseline strategy with our learned strategy within the simulation and real user experiment (Prommer et al. 2006). In the real user experiment 18 subjects were engaged and a total number of 94 dialogs (576 utterances) was collected in sequential runs of four to six dialogs for each test subject. Hereby, in order to fairly balance a potential learning effect of the user, we evenly switched between use of the two strategies. It shows a significant improvement for the reinforcement strategy with a task success of 86.9% versus 80.4% in the real user experiment and 91.3% versus 83.3% in simulation and an average dialogue length of 4.9 utterances versus 5.9 utterances in the real user experiment and 5.0 utterances versus 5.9 utterances in simulation.

in Industrial Robots Journal, Emerald Publishing, 2008, to appear

The superiority of the reinforcement learning strategy is due to its higher number of features and more fine grained rules than could be created with handcrafting.

3.2 Receptionist Scenario

Besides the ability to locate, identify and talk about objects, social communication is a key feature for communication with humans. Robots must be able to get in touch with persons and initiate a conversation. They should have knowledge about who they are talking to and maybe what roles can be associated to this person. When meeting unknown persons, the robot should be able to learn the name of a person and store face or voice samples to be able to recognize the person later on.
A receptionist task is a good scenario to practice exactly these capabilities. Firstly, the robot can proactively offer help and try to engage in a conversation. A reception is also a place where people register to visit someone else, so the robot needs to obtain the name to announce the person.
We have addressed the receptionist scenario with a series of experiments. The experiments cover Wizard-of-Oz experiments to analyze the interaction, experiments with handcrafted models and later experiments with optimized models, which were trained with reinforcement learning.
The Wizard-of-Oz experiment was conducted with a fully integrated system, where the wizard only replaced the dialogue decisions. Fully integrated means automatic person tracking and face ID, speech recognition, natural language understanding and user model. The scenario employs a robot in a corridor. Subjects were given the task to deliver a parcel to a predefined person. In the corridor they obtained further information from the robot. The wizard's task was to greet an arriving person and offer help, ask who the parcel should be delivered to, ask for the person's name for registration, give directions where to deliver the parcel, re-engage in a dialog after delivery of the parcel. The person and the name of the person was unknown in the beginning until the name has been learned through spelling. Further experiments were conducted with name learning using large background vocabularies. Figure 6 shows snapshots from the robot's camera taken during the experiment.
To obtain clean results, the subjects hadn't been instructed how to behave beforehand, and they didn't know that the robot was operated by a human. Team members and the wizard were hidden in rooms and didn't interfere with the scene physically. Observation and recording for later analysis was possible through two cameras in the scene. The Wizard-of-Oz experiment was conducted on three consecutive days with 16 persons, where each person had to do one interaction per day and went through an interview after each interaction. The data, close and distant speech input, robot vision and system logging, were recorded for a data corpus.
The technologies required in this scenario are to engage in a dialogue and attracting attention when the person arrives, learning new names, learning person ids and optimizing dialogues. Prestudies were conducted with experiments to initiate a conversation and for person identification and ID learning (Holzapfel et al. 2007). The receptionist scenario led to longer, more complex dialogues than the barkeeper scenario. Here, several subtasks had to be conducted sequentially. Using a modular dialogue design and sequentially processing dialogue goals (greeting, receptionist task, person registration, etc.) provided the basis for successful training of dialogue strategies using reinforcement learning (Holzapfel and Waibel 2008). Each dialogue goal is associated with one module. This way different modules can contain completely different implementations, for example the first module contains hand-written dialogue strategies, while the second module contains dialogue strategies trained with reinforcement learning. For reinforcement learning (i.e. strategy optimization) the same approach has been taken as in the barkeeper scenario. With the Wizard-of-Oz experiment, data is collected which is analyzed and used for training of a user simulation. In the simulation setup, the dialogue strategy is trained with reinforcement learning. Additional experiments are then conducted to evaluate the dialogue strategy with real users.
The strategies have so far been compared in the simulation setting, where a large number of dialogues can be conducted with different strategy configurations. The results

show that the reinforcement learning strategy produces less errors on the training set and on a held out evaluation set in identifying the persons' names (90% / 91.5% accuracy) than the handcrafted strategy (87.5% / 86.5% accuracy on the same sets) at comparable dialogue length (3.9 / 4.1 turns on avg. versus 4.0 / 3.9 turns on avg.).

4 Conclusion and Outlook

We have presented a dialogue manager for multimodal human-robot interaction and its application to scenarios for which we have selected realistic situations for a humanoid robot in a household environment. The paper tries to motivate the need of and requirements for a dialogue component for robot control and presents several approaches to advance the state of the art with practically applicable algorithms. In the future we think that it is important to continue working on realistic scenarios and to extend the list of current interaction situations, to extend the robot's capability to acquire new information with evaluation over longer time periods, and to address social interaction, multiple speakers and distant speech recognition.

5 References

T. Asfour, K. Regenstein, P. Azad, J. Schröder, A. Bierbaum, N. Vahrenkamp, and R. Dillmann (2006) "ARMAR-III: An integrated humanoid platform for sensory-motor control," *in IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 2006

R. Bischoff, V. Graefe (2002) "Dependable multimodal communication and interaction with robotic assistants" In *Proceedings of the International* Workshop on Robot-Human Interactive Communication (ROMAN), 2002

R. A. Bolt (1980) ""put-that-there": Voice and gesture at the graphics interface," in *Proceedings of the 7th annual conference on Computer Graphics and Interactive Techniques*. ACM Press, 1980, pp. 262–270.

A. Corradini, R.M. Wesson and P.R. Cohen (2002) "Map-Based System Using Speech and 3D Gestures for Pervasive Computing" in *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces (ICMI)*, 2002

M. Denecke (2002) "Rapid Prototyping for Spoken Dialogue Systems" *in Proceedings of the 19th International Conference on Computational Linguistics*, 2002

C. Fügen, H. Holzapfel and A. Waibel (2004) "Tight Coupling of Speech Recognition and Dialog Management -- Dialog-Context Dependent Grammar Weighting for Speech Recognition" in *Proc. of the Intl. Conf. on Speech and Language Processing (ICSLP)*, 2004

P. Gieselmann, P. Stenneken (2006): "Communication with Robots: Evidence from a Web-based Experiment on Human-Computer Interaction" *In Proceedings of the IEEE/ACL Workshop on Spoken Language Technology, 2006*

P. Gieselmann and M. Ostendorf (2007) "Problem-Sensitive Response Generation in Human-Robot Dialogs" *in Proceedings of the 8th SIGDial workshop on discourse and dialogue*, 2007

H. Holzapfel, K. Nickel and R. Stiefelhagen (2004) "Implementation and Evaluation of a Constraint-Based Multimodal Fusion System for Speech and 3D Pointing Gestures" in *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*, 2004

H. Holzapfel (2005) " Building multilingual spoken dialogue systems" in *Archives of Control Sciences - Special Issue on Human Language Technologies as a Challenge for Computer Science and Linguistics (Part II)*, 2005, vol. 15, 555-566

H. Holzapfel and A. Waibel (2006) "A Multilingual Expectations Model for Contextual Utterances in Mixed-Initiative Spoken Dialogue" in *Proceedings of INTERSPEECH*, 2006

H. Holzapfel, T. Schaaf H.K. Ekenel, C. Schaa and A. Waibel (2007) "A Robot learns to know people - First Contacts of a Robot" in *Lecture Notes in Computer Science - KI 2006: Advances in Artificial Intelligence, Springer*, vol. 4314 , 2007

H. Holzapfel and A. Waibel (2008) "Learning and Verification of Names with Multimodal User ID in Dialog" in *Proceedings of the International Conference on Cognitive Systems (CogSys)*, 2008

M. Johnston (1998) "Unification-based Multimodal Parsing", in Proceedings of *COLING-ACL*, 1998, 624-630

O. Lemon, A. Bracy, A. Gruenstein, and S. Peters (2001) "The WITAS Multi-Modal Dialogue System I" in Proceedings of Eurospeech, 2001

E. Levin, R. Pieraccini and W. Eckert (1998) "Using Markov Decision Process for Learning Dialogue Strategies" in *Proceedings of the IEEE, Transactions on Speech and Audio Processing*, 1998

E. Levin, R. Pieraccini and W. Eckert (2000) "A Stochastic Model of Human-Machine Interaction for Learning Dialog Strategies" in *IEEE Transaction On Speech and Audio Processing*, January 2000

M. Montemerlo, J. Pineau, N. Roy, S. Thrun and V. Verma (2002) "Experiences with a mobile robotic guide for the elderly" in Proceedings of the National Conference on Artificial Intelligence (AAAI), 2002

K. Nickel and R. Stiefelhagen, "Pointing gesture recognition based on 3d-tracking of face, hands and head orientation", *in Proceedings of the Fifth International Conference on Multimodal Interfaces (ICMI)*, 2003.

S.L. Oviatt, A. DeAngeli and K. Kuhn (1997) "Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction" in *CHI '97: Proceedings of the SIGCHI conference on Human factors in computing systems,* 1997, 415-422

T. Prommer, H. Holzapfel and A. Waibel (2006) "Rapid Simulation-Driven Reinforcement Learning of Multimodal Dialog Strategies in Human-Robot Interaction" *in Proceedings of Interspeech - ICSLP*, 2006

J. Schatzmann, K. Georgila and S. Young (2005) "Quantitative Evaluation of User Simulation Techniques for Spoken Dialogue Systems" in Proceedings of the *6th SIGdial Workshop on Discourse and Dialogue,* 2005

K. Scheffler and S. Young (2002) "Automatic Learning of Dialogue Strategy Using Dialogue Simulation and Reinforcement Learning" in *Proceedings of the Human Language Technology Conference,* 2002

S. Singh, M. Kearns, D. Litman and M. Walker (1999) "Reinforcement Learning for Spoken Dialogue Systems" in *Proceedings of the Conference on Neural Information Processing Systems,* 1999

R. Stiefelhagen, C. Fügen, P. Gieselmann, H. Holzapfel, K. Nickel, and A. Waibel (2004) "Natural human-robot interaction using speech, gaze and gestures," in *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, 2004

I. Toptsis, S. Li, B. Wrede and G. Fink (2004) "A multi-modal dialog system for a mobile robot" in Proceedings of the International Conference on Spoken Language Processing (ICSLP), 2004

M. Walker (2000) "An Application of Reinforcement Learning to Dialogue Strategy Selection in a Spoken Dialogue System for Email" in *Journal of Artificial Intelligence,* 2000

J. Williams and S. Young, (2003) "Using Wizard-of-Oz Simulations to Bootstrap Reinforcement-Learning-Based Dialog Management Systems" in *Proceedings of the 4th SigDial Workshop on Discourse and Dialogue,* 2003

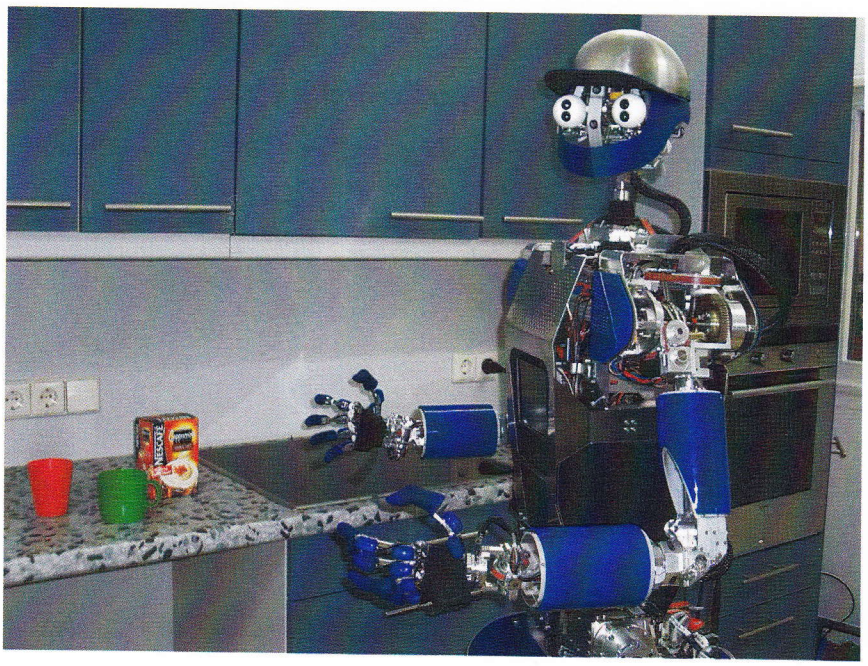

Figure 1: The humanoid robot Armar 3 in the kitchen environment.

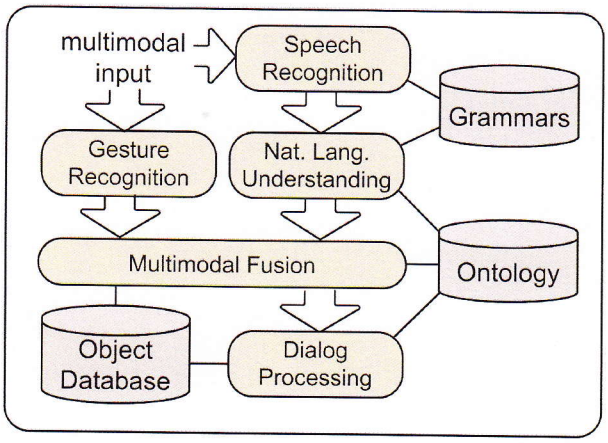

Figure 2: Dialogue architecture with speech and gesture recognition, fusion and knowledge bases.

in Industrial Robots Journal, Emerald Publishing, 2008, to appear

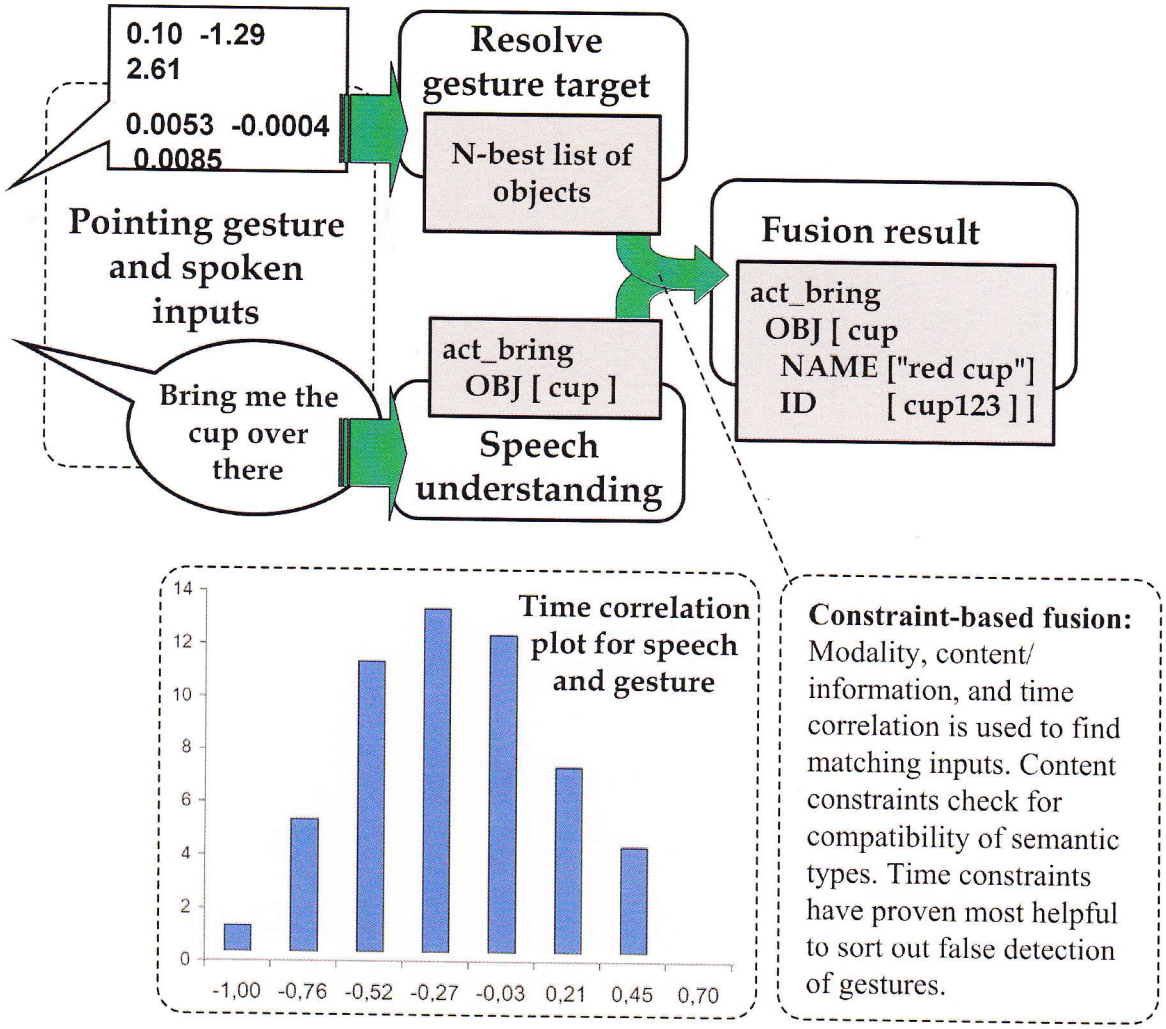

Figure 3: Multimodal fusion scheme. The upper picture shows the schematic processing chain for speech and gesture input. The lower picture shows time deltas (x-axis: seconds, y-axis: occurrences) between referring words in speech and gesture events. They were used to formulate time constraints.

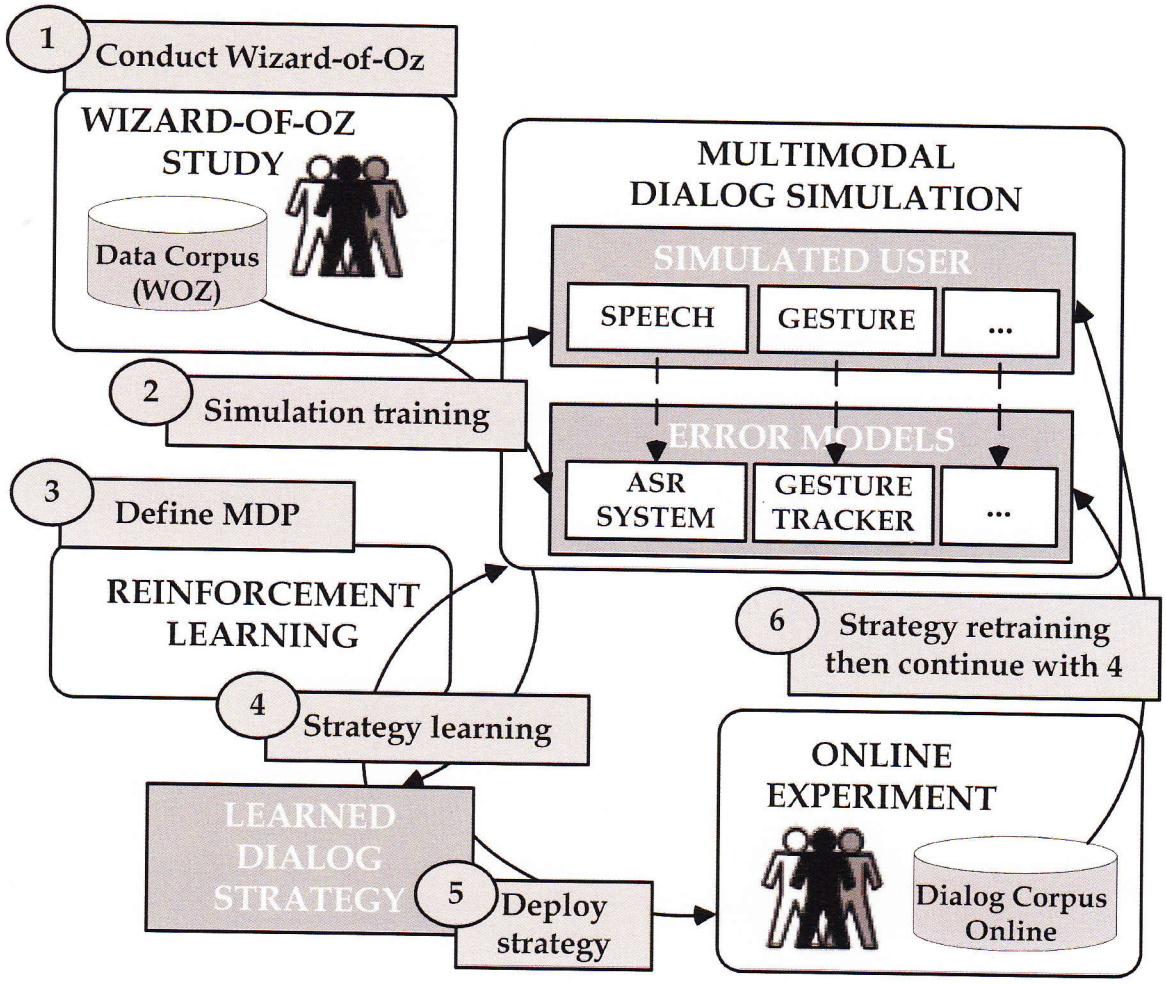

Figure 4: Design process for training robust dialogue strategies using reinforcement learning with a simulated user.

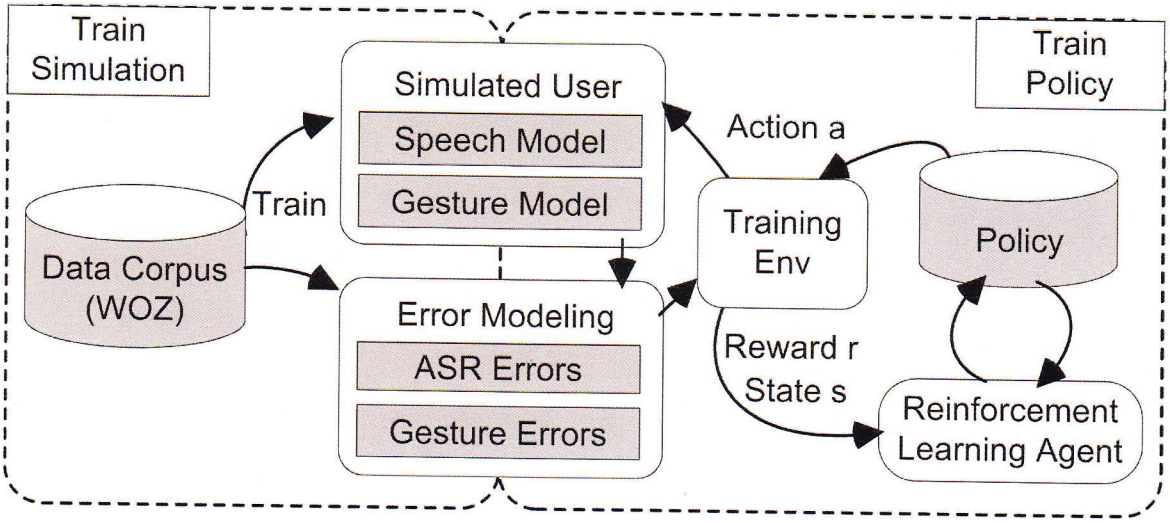

Figure 5: Simulation environment for reinforcement learning of the dialogue strategy.

Figure 6: Snapshots from the robot's camera with corridor view in a robot receptionist experiment. The subject's nametag was helpful for video annotation, it is obfuscated here.

in Industrial Robots Journal, Emerald Publishing, 2008, to appear