Noname manuscript No.
(will be inserted by the editor)

Using Embodied Multimodal Fusion to Perform Supportive and Instructive Robot Roles in Human-Robot Interaction

Manuel Giuliani · **Alois Knoll**

Abstract We present a robot that is working with humans on a common construction task. In this kind of interaction, it is important that the robot can perform different roles in order to realise an efficient collaboration. For this, we introduce *embodied multimodal fusion*, a new approach for processing data from the robot's input modalities. Using this method, we implemented two different robot roles: the robot can take the *instructive role*, in which the robot mainly instructs the user how to proceed with the construction; or the robot can take the *supportive role*, in which the robot hands over assembly pieces to the human that fit to the current progress of the assembly plan. We present a user evaluation that researches how humans react to the different roles of the robot. The main findings of this evaluation are that the users do not prefer one of the two roles of the robot, but take the counterpart to the robot's role and adjust their own behaviour according to the robot's actions. The most influential factors for user satisfaction in this kind of interaction are the number of times the users picked up a building piece without getting an explicit instruction by the robot, which had a positive influence, and the number of utterances the users made themselves, which had a negative influence.

Keywords Human-Robot Interaction, Embodied Multimodal Fusion, Robot Roles

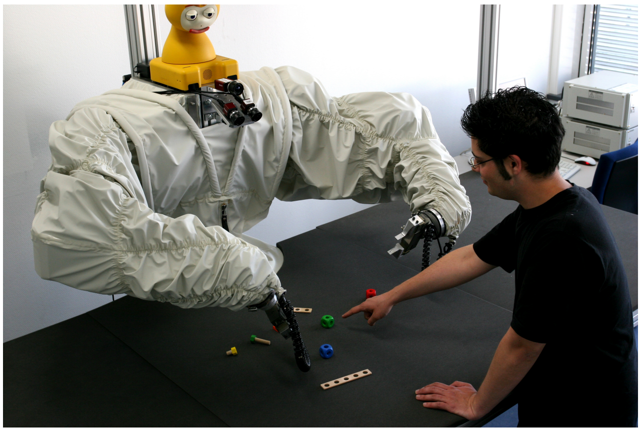


Figure 1: JAST/JAMES human-robot interaction system.

1 Introduction

When humans work together, they take different roles in the interaction. For example, when two persons assemble a shelf, usually one of them takes the lead and gives instructions on how to follow the assembly plan. The other person helps the instructor to build the shelf and to gather the right parts for the next building step. This aspect of role taking in social interactions, such as common construction tasks, is studied in social psychology as well as in robot-robot interaction. Previous work shows for example that the assignment of roles can bias the opinions of human interaction partners about each other [24] and that it leads to more efficient task

Manuel Giuliani is a researcher at fortiss GmbH, Guerickestraße 25, 80805 München, Germany, E-mail: giuliani@fortiss.org
Alois Knoll is a professor at the Technical University Munich, Robotics and Embedded Systems, Boltzmannstraße 3, 85748 Garching bei München, Germany, E-mail: knoll@in.tum.de

solving, both in human-human as well as in robot-robot cooperation [27,29].

We believe, that role assignment is also important for a successful human-robot interaction. Therefore, in this publication we are following two main research questions: first, which mechanisms do we have to implement on a robot in order to assign different roles to the robot when it is working together with a human on a common task? And second, in a common construction task, do humans prefer a robot that takes the role of an instructor or a supporter? Thus, we introduce a new methodology, the so-called *embodied multimodal fusion* (EMF), with which we realise different roles on a human-robot interaction system. The main idea behind EMF is a combination of a representation for the actions a robot can execute with an algorithm that computes the relevance of these actions based on the robot's input modalities.

To prove the applicability of embodied multimodal fusion, we conduct a human-robot interaction experiment in which naïve participants have to build target objects from a wooden toy construction set together with the robot shown in **Figure 1**. The robot takes either the role of an instructor or a supporter. We use the robot in both settings for a between participants experiment, in which we collect objective and subjective measurements to compare the changes in behaviour and opinion about the robot between the two experiment participant groups.

2 Related Work

In sociology, the so-called role theory [2] outlines how the different roles that humans take when interacting with each other influence human behaviour. Role theory discriminates different types of roles: cultural roles (e.g., priest), social differentiation (e.g., teacher), situation-specific roles (e.g., eye witness), bio-sociological roles (e.g., as human in a natural system), and gender roles (e.g, wife or husband). In this work, we are interested in situation-specific roles, because we perform different roles with our robot and measure how humans react to these robot roles.

The effects of role taking in human-human interaction are studied since the 1970ies. The well-known Stanford prison experiment by Zimbardo et al. [35] showed what effect the assignment of roles has on the behaviour of humans. In the experiment, a group of male students took on randomly assigned roles of prisoners and guards in a mock prison situated in the basement of the Stanford psychology building. The experiment showed that participants adapted to their assigned roles quickly: the guards put down a rebellion by the prisoners and ultimately subjected some of the prisoners to psychological torture. Many of the prisoners accepted their role and began to act submissively in order to avoid the guards' measures.

Ross et al. [24] show that the assignment of roles can change the judgement of interaction partners. In an experiment, they grouped participants into pairs and assigned a role to each of them: one participant, the so-called questioner, had to ask questions of general knowledge while the other participant, the answerer, had to answer the questions. When asked for their judgement, the majority of all experiment participants as well as independent experiment observers rated the questioner as being more superior than the answerer. The effectiveness of role assignment in human-human interaction was shown by Stewart and Strasser [27]. The authors showed in an experiment that people share more information in discussions when some of them were explicitly assigned the role of a knowledge expert before the interaction.

In contrast to this work, we are looking at a situation in which a human and a robot work together on a construction task. This means that human and robot interact in a case of joint action, which Sebanz and Knoblich define as "any form of social interaction whereby two or more individuals coordinate their actions in space and time to bring about a change in the environment" [25]. Research in joint action showed that for successful joint action, different aspects are important, such as joint attention, action observation, task sharing, action coordination, and agency. The effectiveness of joint action was researched by Clark and Krych [6], who used an experiment design which was similar to the experiment design that we are reporting in this work. In Clark's and Krych's experiment, two humans were asked to assemble Lego work pieces together. During the interaction, one of the humans was assigned the role of a director, who knew the building plan. The other human got the role of the builder, who had to follow the director's instructions. Clark and Krich showed that the two partners were much slower when the directors could not see the workspace of the builder, because they had to use more words and could not use gestures. However, they did not measure the effect of different roles on the experiment participant's behaviour.

In robotics research, there are two areas in which the role of a robot is of importance: on the one hand, there are robots that have to interact with humans. Here, the research focuses on different roles the robot can take in an interaction and how the human partners of the robot react to these roles. On the other hand, researchers are interested in the roles of robots in multi-robot teams.

Here, robots use different roles to solve a given task more effectively.

Breazeal et al. [4] were among the first authors who realised that for a social robot that is capable to interact with a human it is of importance, which social role the robot should take in this interaction. Hinds et al. [15] presented one of the earliest studies that researched how humans react to different robot roles. They conducted an experiment, in which a human and a robot had to work together. In the experiment, the authors varied the appearance of the robot as well as the behaviour (i.e. the role) of the robot. The results of the experiment show that humans rely more on human-like robots and feel more responsible for the task when the robot looks more machine-like. The experiment also showed that the participants felt less responsible for the task when they worked with a robot who took the role of a supervisor, which is also supported by our findings.

Tapus and Mataric [30] show a therapist robot that monitors and encourages humans in rehabilitation exercises. This robot shows either an introverted or an extroverted personality. Tapus and Mataric were able to show in an experiment that introverted patients interacted significantly longer with the introverted robot, while extroverted patients interacted longer with the extroverted robot, respectively. In contrast to our findings, it seems that in this type of interaction, humans prefer to have a partner with similar personality traits to their own.

Stone and Veloso [29] were among the first who realised that role assignment is also important for multi-robot teams. They presented an approach for dynamic assignment of roles in robot-robot teamwork. For that, they teamed up their robots in formations, in which every robot was given a certain role. The increased effectiveness for task solving was demonstrated by the authors by wining the RoboCup football league in 1999. Chaimowicz et al. [5] proposed a new methodology for coordinating multi-robot teams in the execution of cooperative tasks, which was based on dynamic role assignment. They used this mechanism to solve a cooperative transportation task that had to be fulfilled by a group of robots. Looije et al. [21] argument that for urban search & rescue robots (USAR) affective computing is important. Amongst other things, they present the theoretical basis for the implementation of social roles on a USAR, so that the robot can adapt its own behaviour on a rescue mission, corresponding to whether it is interacting with a fellow helper or a victim.

In summary, the related work from sociology shows that roles are an important element of human-human interaction, which affects human behaviour. The work from both, sociology and robotics, furthermore shows that in human-human, human-robot, and even robot-robot joint action, the assignment of roles to the interaction partners leads to a more efficient and successful interaction. On the following pages, we will now add a new aspect to this research and explore how humans react to different roles of a human-robot interaction system with which they are asked to work together.

3 Human-Robot Interaction System

For this work, we use a completely autonomous human-robot interaction system (**Figure 1**) which supports multimodal human-robot collaboration on a joint construction task. This system was developed for the EU-funded project JAST[1] (Joint Action Science and Technology) and is now used in the EU-funded project JAMES[2] (Joint Action for Multimodal Embodied Social Systems). The robot has a pair of manipulator arms with grippers, mounted in position to resemble human arms, and an animatronic talking head [31] capable of producing facial expressions, rigid head motion, and lip-synchronised synthesised speech. The robot uses *speech*, *gesture* [34], and *object recognition* [23] as input modalities. Additionally, the robot has a *task planner* [12] that provides building plans for the target objects that the robot builds together with a human partner.

The robot is able to work together with a human on a common construction task. In this task, both partners assemble wooden construction toys (**Figure 2**) on a common workspace, coordinating their actions through speech and gestures. The robot can pick up and move objects in the workspace and perform simple assembly tasks. In the scenario considered here, human and robot are both given an assembly plan that they jointly execute. The robot assists the human by explaining necessary assembly steps and by offering pieces as required. To ensure that the robot has to engage in conversation with the human and needs to hand over assembly pieces, the workspace is divided into two areas—one belonging to the robot and one to the human. That means that joint action is necessary for task success.

We programmed the robot so that it can take two different roles in the interaction: in the *instructive role*, the robot first gives instructions to the user how to assemble pieces according to the assembly plan before handing over construction pieces from its own work area. In the *supportive role*, the robot first hands over construction pieces from its own work area to the human and only gives instructions if the human does not pick up the right construction pieces from her/his work

[1] `http://www6.in.tum.de/Main/ResearchJast`

[2] `http://www.james-project.eu`

area or if the robot has no further pieces to handover from its work area. This means that in theory when the robot is set to the supportive role, it can be the case that it never gives instructions to the human, but only if the human picks up all necessary pieces, before the robot can give instructions. However, this never happened in the experiment. Although the construction pieces can be screwed and stuck together, the robot is not able to perform assembly actions itself. However, the robot supports its human co-worker with the following list of actions:

- *give*, the robot hands over a construction piece from its own work area to the human. While handing over the piece, the robot also says which piece it is currently giving to the human.
- *tellAbout*, the robot instructs the human to pick up a piece from the human's work area because it fits to a building step of the currently loaded plan.
- *askFor*, the robot asks the human to put a certain construction piece on the work area, because it is needed for a given plan and the robot cannot detect it with its object recognition.
- *tellBuild*, the robot asks the human to build one of the substeps of the currently loaded plan.
- *thankFor*, the robot thanks the human for a construction piece that the human has put on the table.

4 Embodied Multimodal Fusion

The different behaviours that the robot displays when it is set to either the instructive or the supportive role is generated by executing robot actions in varying orderings. In this section, we describe a new approach for selecting when to execute these actions, based on the input information that the robot gets from its input modalities: *embodied multimodal fusion* (EMF).

A robot that should interact with humans in a way that seems natural to the human, needs to have input modalities for understanding the human's utterances, for example speech and gesture recognition, and to observe the state of its environment, for example object recognition. Furthermore, the robot needs to integrate all of this information in order to successfully interact with humans, a process that is called *multimodal fusion*. The goal of multimodal fusion is to enable a robot to understand the information from various input modalities and to combine this information into an integrated representation so that the robot is able to perform its designated task.

Multimodal fusion has been applied mainly in multimodal dialogue systems. Here, the proposed approaches for multimodal fusion can be separated into two main classes: *late fusion* or *semantic fusion* methods process the data from their input modalities with separate recognition modules and fuse the interpretations of the modules in a central interpretation module. Some newer examples of multimodal dialogue systems that use late fusion are SmartKom [32] and COMIC [3]. In contrast to that, *early fusion* or *feature level fusion* systems integrate input channels that are closely bound to each other, for example speech and lip movements or speech and pointing gestures. These modalities can be fused by combining features from both modalities into one feature vector. The resulting feature vectors are used for training of statistical models, which are then used for recognition of multimodal events. Examples for multimodal systems that use early fusion are Quickset [7] and MATCH [17].

Late and early fusion have in common that they are centred around processing of human utterances, since they were developed for multimodal dialogue systems. In these systems, the fusion process is only started when a verbal utterance by the human is at hand. This spoken utterance is then enriched with information from other modalities. Hence, these approaches work well in scenarios in which robots are servants that get direct spoken commands by a human. Successful implementations of these classic approaches for multimodal fusion on human-robot interaction systems are for example the works by Holzapfel et al. [16] and Stiefelhagen et al. [28]. However, we argue that for an interaction that humans perceive as natural, especially in the context of task-based scenarios, the robot needs to be able to engage in a mixed-initiative interaction. For that, the robot has to execute actions not only if it gets a direct command by the human, it also has to take in consideration information about its environment and the progress of a given task and choose appropriate actions to respond to this information.

EMF processes information from all input modalities in parallel to make use of human utterances as well as knowledge about the robot's environment and task. The main idea behind EMF is that each robot is able to execute a defined set of actions. For example, the robot that we are using in this publication can execute the actions *askFor*, *give*, *tellAbout*, *tellBuild*, and *thankFor*, which we described in the previous section. If the robot should collaborate with a human in a meaningful way, it needs to produce the correct action at the right time. For this reason, rather than only focussing on the input by the human and how to represent it, as it is done in classical multimodal fusion for multimodal dialogue systems, the robot should also have a representation for its own actions and a way to determine which of these actions it should execute in a given context. This enables the robot to engage in a mixed-initiative interaction with humans, which leads to an interaction that is perceived by the humans as more natural.

Thus, the basic processing steps in EMF are: (1) the robot generates representations of objects in correlation to actions that it can execute with these objects. For representation generation, the robot uses information from all input modalities that provide information about objects, for example object recognition or task planning. (2) The robot uses information from all modalities, including modalities providing information about human utterances, to calculate the relevance of each of its possible actions that it finds in its generated representations. (3) The robot uses the relevance values to select one representation and executes the action that is represented by it.

In the following sections, we introduce the representations for objects and actions, which are used in the EMF approach. We show how EMF generates these representations using information from object recognition and task planning. Subsequently, we present an action selection mechanism that is used by EMF to select the right action to execute in any given situation.

4.1 Objects and Actions

The representations used in EMF are based on the idea that for a robot, objects and actions are inseparably intertwined, which is influenced by Gibson's Affordances [13]. The rational behind this is that objects are defined by the actions one can execute with them. For example, a glass can be used as a container for liquids, but the same glass can also be a supporting stand for other objects when it is turned upside down. Vice versa, some actions cannot be executed without the appropriate object, for example nailing cannot be done without a hammer. The European project Paco+[3] called this connection between objects and actions *object action complex* (OAC). Krüger et al. [19] presented a formal definition of OACs, which states that an OAC consists of a unique identifier, a prediction function that codes the systems belief on how the world will change through the OAC, and a statistical measure that represents the success of the OAC within a window of the past. In our work, we do not need the full capabilities of OACs, but a way to represent objects and actions in the EMF approach. Thus, from now on we will talk about *OAClets*, which are defined in **Definition 1**.

Definition 1 An OAClet is a triple $(\mathcal{O}, \mathcal{A}, \mathcal{R})$ containing an object $\mathcal{O}$ which is associated to an action $\mathcal{A}$ and vice versa, and a score $\mathcal{R}$ that denotes the relevance of the OAClet in a given situation.

[3] `http://www.paco-plus.org/`

EMF uses the data from the input modalities of a robot to maintain a set of OAClets. This set is a list of all possible actions that the robot can execute in a given situation. Using this list, EMF calculates the relevance of any of these actions for that situation. In order to add OAClets to the list and to rate their relevance, we formally define two classes of modalities: *OAClet-generating modalities* (OGM) are used to *generate* OAClets (**Definition 2**); *OAClet-evaluating modalities* (OEM) (**Definition 3**) are used to calculate the relevance scores of a set of OAClets.

Definition 2 An *OAClet-generating modality* is a robot input modality that provides information about an existing or abstract object $\mathcal{O}$. This information is used to add one or more OAClets containing $\mathcal{O}$ with appropriate actions $\mathcal{A}$ to a set of OAClets.

Definition 3 An *OAClet-evaluating modality* is a robot input modality that provides information about an object $\mathcal{O}$ and/or an action $\mathcal{A}$. This information is used to recalculate the relevance score $\mathcal{R}$ of every OAClet in a set of OAClets containing either $\mathcal{O}$ or $\mathcal{A}$. For relevance score calculation, every OEM has an individual function.

The robot that we are using in this publication has two OGMs, object recognition and task planning, and three OEMs, speech and gesture recognition, and task planning as well. The following example illustrates how EMF uses the information from the robot's input modalities. In the example, object recognition has recognised two objects in the working area of the robot, a red cube and a yellow bolt, and two objects in the human's working area, a blue cube and a green bolt. Furthermore, the task planner has loaded the plan to build a tower, which consists of a green bolt, a blue cube, and a red cube as shown in **Figure 2c**. Hence, EMF maintains the following list of OAClets:

cube(red)	*give*	0.33
bolt(yellow)	*give*	0.0
cube(blue)	*tellAbout*	0.33
bolt(green)	*tellAbout*	0.33

In this work, we are using relevance scores that range between 0 and 1. Furthermore, the relevance scores of all OAClets sum up to 1. At this point in the interaction, the robot can either choose to hand over pieces to the human by using action *give* or it can tell the humans to pick up pieces from their work area by using action *tellAbout*. Three OAClets have the same relevance score in this situation, because they contain objects that are needed to build a tower, an information that was provided by the task planner. When the human points to the red cube and says "give me this cube", the information from speech and gesture recognition is used to recalculate the relevance scores of all OAClets:

cube(red)	*give*	0.6
bolt(yellow)	*give*	0.1
cube(blue)	*tellAbout*	0.2
bolt(green)	*tellAbout*	0.1

Here, EMF increases the scores of all OAClets that contain either a cube or action *give*. Furthermore, it increases the score of the OAClet that contains the cube the human pointed at. This leads to a situation in which the robot can use the relevance score to decide which OAClet it should execute next. In the next section, we explain the full action selection mechanism the robot is using in our implementation.

4.2 Action Selection

Due to the continuous update of possible actions in its list of OAClets, the robot can now choose OAClets for execution independently from the human. For that, we implemented a simple action selection algorithm, which is sufficient for the experiment that we describe in **Section 5**. There are three events in which the robot will execute one of its OAClets:

- When the human gives a direct command containing an action and an object (e.g., "give me a red cube") and there is an OAClet in the list of OAClets that represents the according action and object. This replicates the typical behaviour of classical multimodal fusion algorithms.
- When the relevance score of an OAClet exceeds a predefined threshold.
- When the relevance score of several OAClets exceed the predefined threshold at the same time. In this case, an action hierarchy is used to determine which OAClet should be executed.

The action hierarchy is inspired by the task hierarchies of Sentis and Kathib [26]. An action hierarchy defines an order in which actions should be executed. For this work, we used two action hierarchies for the two robot roles that we implemented. In the instructive role we used an action hierarchy as follows:

$tellAbout > give > tellBuild > askFor > thankFor$

This hierarchy leads to a behaviour were the robot preferably tells the humans which building piece they need to pick up, and only hands over pieces to the human when that is not possible. In comparison to that, the action hierarchy for the supportive role stacks the robot actions in a different order:

$give > tellBuild > tellAbout > askFor > thankFor$

Here, the robot preferably hands over building pieces to the humans and tells them to assemble pieces. The robot only tells the humans which pieces to pick up when these other actions are not available.

5 Experiment

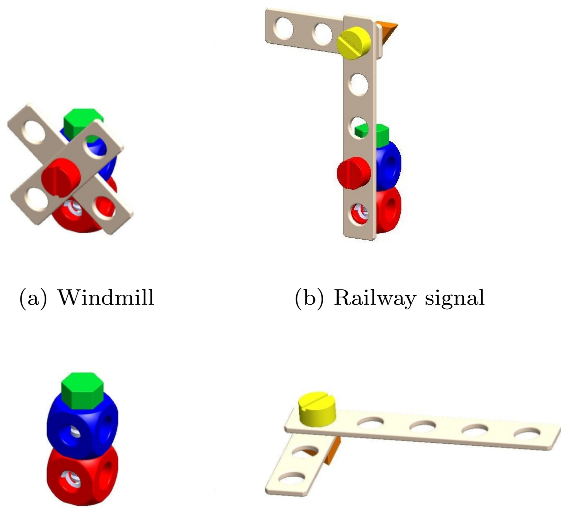
In the previous section, we have shown how we implement robot roles on our human-robot interaction system by using EMF. Now, we report the results of an evaluation that was targeted to research how humans rate these different robot roles. We describe the experiment set-up, demography of the experiment participants, data collection, data analysis, and results.

5.1 Experiment Design

This experiment used a between participants design with one independent variable: each participant interacted either with the robot that used the supportive role setting, or else with a system that used the instructive role. We will refer to these dimensions as *supportive* versus *instructive*. The 40 participants were assigned randomly to one of the two experiment conditions.

The robot was completely autonomous and did not get any other information than the data from its input modalities: speech, gesture, and object recognition, as well as task planning. The participants were instructed that they should work together with the robot on a common task, but they were not informed that the robot would take a role in the interaction. The common task of human and robot was to build target objects together. In each of the experiments, human and robot built two target objects together, always in the same order, first the *windmill* (**Figure 2a**), after that the *railway signal* (**Figure 2b**). For both target objects, the human was given an assembly plan on paper and the robot got the plan from its task planner.

The participants stood in front of the table facing the robot, equipped with a headset microphone for speech recognition. Participants got instructions that they could speak with the robot by using a set of predefined phrases: they could either ask the robot for one of the pieces in the robot's work area by giving a direct order, for example by saying "give me a blue cube", or they could ask the robot to repeat its last utterance by saying "pardon me?".



(a) Windmill (b) Railway signal

(c) Tower (d) L shape

Figure 2: Target objects for the experiment. Both target objects consist of a base that is named *tower*. A windmill is a tower combined with two small slats; a railway signal is a tower combined with an *l shape*.

The pieces required for the target object were placed on the table, using the same layout for every participant. The layout was chosen to ensure that there would be enough similar construction pieces on both sides of the table for every subplan of the target objects so that the robot could either perform action *give* and handover an object from its side of the table or action *tellAbout* and instruct the users to pick up an object from their work area. For example, for the tower of the windmill there was a red cube in both table areas, so that the robot could either hand over the cube from its own work area or instruct the participants to pick up the cube from their work area. Along with the assembly plan mentioned above, the participants were given a table with the names of the pieces they could build the objects with.

5.2 Participants

40 participants (27 male), who never worked with the robot before, took part in this experiment. The mean age of the participants was 27.20 (7.06), with a minimum of 17 and a maximum of 59. Of the participants who indicated an area of study, the two most common areas were mathematics (11 participants) and computer science (8 participants). On a scale of 1 ("I do not agree at all") to 5 ("I completely agree"), participants gave a mean assessment of their knowledge of computers at 3.68 (1.00), of speech recognition systems at 1.90 (1.03), and of human-robot interaction systems at 1.60 (1.01). For their participation in the experiment, the participants got the chance to win a voucher for an online shop.

5.3 Hypotheses

In this experiment, we compare how humans react to the instructive and supportive role of the robot when they build target objects with it. We are mainly interested if the experiment participants accept both roles of the robot or if there is a clear preference for one of the two roles. In particular, we have the following two hypotheses:

H1 Experiment participants who work with the robot in the supportive role, generally asses their interaction with the robot more positive.

H2 Experiment participants who work with the robot in the supportive role display a more proactive behaviour, while participants using the instructive robot will take a more passive role in the interaction.

We derive hypothesis **H1** from work by Dautenhahn et al. [9] who presented experiment results that indicate that humans prefer robots that take the role of an assistant or servant rather than the role of a friend. This notion was also recently reported in two industry-sponsored surveys [22,1]. Hypothesis **H2** is motivated by the work of Hinds et al. [15] who showed that humans take more responsibility for the assigned task when the robot does take the role of a supporter.

Since we gathered a wide range of subjective and objective measures in this study, we did not make predictions as to which specific measure the experimental manipulations will have an effect.

5.4 Data Acquisition

At the end of each experiment, the participants responded to a usability questionnaire consisting of 29 items, which fell into four main categories: *feelings of the user* (10 items), *intelligence of the robot* (7 items), *robot behaviour* (6 items), and *task success* (6 items). We present all questionnaire items in **Table 1**. The items were based on those used in two previous user evaluations [10] [14], but were adapted for the scenario and research questions of the current study. The questionnaire was presented using software that let the participants choose values between 1 ("I do not agree at all") and 100 ("I completely agree") with a slider. Due to the presentation of the questionnaire items on a computer display, we decided to use a slider instead of more widely used scales, such as for example a Likert scale. It has been shown by Cook et al. [8] that the usage of sliders in screen-based questionnaires produces the same reliability in the ratings as using Likert scales, but provides an increased usability for the users.

In addition to the questionnaire, we collected a set of objective measurements from the automatically generated system log files and from annotations of the videos we took during the experiments. All in all, we had five different objective measurements:

– the *number of verbal utterances* by the participants, which is the number of times the users asked the robot for a certain construction piece or to repeat its last utterance,
– the *number of times the participants picked up a construction piece* from their side of the table, where the robot did not instruct them to pick up the object, we will refer to these pickup actions as *anticipatory pick up actions*,
– the *number of instructions the robot gave to the participants*, i.e. the instructions in which the robot told the human which piece to pick up next from the work area,
– the *number of correct pick up actions after instruction*, i.e. the number of times a participant picked up the correct object when the robot instructed him/her to do so, we will refer to this as *instructed pick up actions*, and
– the *overall duration* the participants needed to build windmill and railway signal.

We took the first two measurements from the system log files and annotated the videos of the experiment participants with Anvil [18] to collect the remaining three measurements. Not all participants agreed that we videotaped them, thus we only have video data for 32 of the 40 participants, 17 videos of participants who used the instructive robot and 15 videos of participants who used the supportive robot.

5.5 Results

In this study, we analysed the collected data in several ways. First, we compared the subjective answers of the experiment participants to the user questionnaire to find out if there are any significant differences between the answers of the group that worked with the supportive robot and the group that worked with the instructive robot. Second, we compared the objective measurements that we took from the system logs and the videos to find differences between the two groups. Third, we calculated which of the objective measurements could potentially predict the subjective answers by the experiment participants.

5.5.1 Subjective Measurements

Table 1 shows the results from the user questionnaire. We computed Cronbach's alpha, which is a measurement of internal consistency. The consistency of the questionnaire has a value of $\alpha = 0.838$ which falls over the acceptable range of $\alpha > 0.700$. We applied a Mann-Whitney test on the answers to the user questionnaire to analyse if the different robot roles had a significant effect on the ratings by the two participant groups. Generally, participants gave a positive feedback of an average 82.84 (20.26) points on the questions of the *feelings of the user* category, in which they had to rate if their interaction with the robot was enjoyable. However, the participants rated the robot's intelligence with only 56.35 (26.16) points. There was no significant difference in these questions between the two groups.

We found significant differences (p-value < 0.05) in the ratings for 4 of the 29 statements of the user questionnaire. **Table 1** shows the results for all items of the user questionnaire, the significant results are marked with bold text.

5.5.2 Objective Measurements

We show the results of the objective measurements in **Table 2**. We computed if there is a significant difference between the two user groups, again via a Mann-Whitney test. We found a significant difference for the number of robot instructions, which is not surprising, but shows that the instructive robot gave significantly more instructions to the user. Participants who got more instructions by the robot, also did significantly more instructed pickups. Furthermore, participants who worked with the supportive robot significantly picked up more construction pieces without getting instructions from the robot to do so.

5.5.3 Predictive Measurements

To complete the result analysis of this study, we calculated a predictor function to compute if the objective measurements we collected in this evaluation could predict the subjective statements of the user questionnaire. Being able to predict subjective user satisfaction from more easily-measured objective properties can be very useful for developers of interactive systems: in addition to making it possible to evaluate systems based on automatically available data without the need for extensive experiments with users, such a performance function

Table 1: Statements of the user questionnaire with mean rating and standard deviation for both participant groups. Significant results are displayed in **bold fonts**. Please note that some of the statements are negatively stated, which means that a lower score is better in these cases.

Statement	Supportive mean (stdev)	Instructive mean (stdev)	Mann-Whitney p-value	U-value
Feelings of the user				
I felt confused when using the robot.	22.50 (27.97)	17.50 (23.66)	$p = 0.651$	$U = 183.0$
It was easy to work with the robot.	83.75 (19.22)	86.10 (17.19)	$p = 0.978$	$U = 198.5$
I found the robot easy to use.	**83.80 (12.81)**	**90.80 (13.03)**	$\mathbf{p = 0.043}$	$\mathbf{U = 126.5}$
I found the conversation engaging.	44.40 (33.56)	46.85 (27.19)	$p = 0.860$	$U = 193.0$
I found it exciting to interact with the robot.	76.40 (23.60)	79.35 (21.02)	$p = 0.693$	$U = 185.0$
I felt tense when using the robot.	32.30 (29.33)	42.85 (27.47)	$p = 0.137$	$U = 144.5$
I really had to concentrate to use the robot.	21.75 (24.09)	26.30 (28.83)	$p = 0.645$	$U = 182.5$
I found the conversation boring.	42.45 (29.60)	41.00 (25.19)	$p = 0.989$	$U = 199.0$
I liked using the robot.	84.75 (16.38)	87.65 (17.59)	$p = 0.378$	$U = 167.5$
The robot appeared to be intelligent.	56.55 (27.34)	56.15 (25.63)	$p = 0.946$	$U = 197.0$
Intelligence of robot				
The robot understood what I said to it.	62.95 (23.50)	75.55 (27.85)	$p = 0.108$	$U = 141.5$
The robot responded quickly.	55.60 (19.86)	68.20 (29.02)	$p = 0.071$	$U = 133.5$
I found the voice of the robot easy to understand.	95.60 (5.31)	95.85 (8.20)	$p = 0.415$	$U = 172.0$
I knew what I could say or do at each point in the conversation.	**71.05 (30.32)**	**90.10 (12.73)**	$\mathbf{p = 0.038}$	$\mathbf{U = 123.5}$
It was clear what to do when the robot did not understand me.	**70.65 (21.46)**	**57.35 (15.66)**	$\mathbf{p = 0.034}$	$\mathbf{U = 124.0}$
The robot worked the way I expected it to.	61.90 (26.56)	72.40 (24.01)	$p = 0.310$	$U = 162.0$
I found the robot to be knowledgeable.	88.60 (13.24)	88.05 (19.59)	$p = 0.514$	$U = 176.0$
Robot behaviour				
It seemed natural when the robot picked up objects from the table.	63.50 (27.61)	73.05 (25.19)	$p = 0.309$	$U = 162.0$
found it helpful when the robot picked up objects from the table.	69.80 (29.83)	76.15 (21.71)	$p = 0.635$	$U = 182.0$
The robot showed an active behaviour.	72.30 (23.14)	75.55 (17.09)	$p = 1.000$	$U = 199.5$
The robot showed an passive behaviour.	22.20 (23.42)	20.20 (21.97)	$p = 0.596$	$U = 180.0$
The robot gave too many instructions.	**33.95 (28.21)**	**16.70 (22.53)**	$\mathbf{p = 0.026}$	$\mathbf{U = 117.5}$
The instructions from the robot were sufficient.	87.95 (18.98)	95.00 (7.90)	$p = 0.114$	$U = 143.0$
Task success				
I was able to work with the robot successfully.	89.25 (10.51)	91.85 (16.63)	$p = 0.114$	$U = 143.0$
The robot helped me well to build the windmill.	83.80 (11.04)	77.40 (25.17)	$p = 0.725$	$U = 186.5$
The robot helped me well to build the railway signal.	74.75 (21.40)	79.00 (25.52)	$p = 0.409$	$U = 169.0$
The assembly tasks were too difficult.	7.20 (12.68)	1.55 (2.56)	$p = 0.321$	$U = 167.0$
It was easy to understand how to put the pieces together.	92.50 (8.59)	94.15 (8.41)	$p = 0.523$	$U = 177.0$
It was easy to understand which pieces to use.	94.20 (9.56)	91.30 (16.54)	$p = 0.862$	$U = 193.5$

can also be used in an online, incremental manner to adapt system behaviour to avoid entering a state that is likely to reduce user satisfaction, or can be used as a reward function in a reinforcement-learning scenario [33].

To compute the predictor function, we employed a procedure similar to that used in the PARADISE evaluation framework (PAradigm for DIalogue System Evaluation) [33]. The PARADISE model uses stepwise multiple linear regression to predict subjective user satisfaction based on measures representing the performance dimensions of task success, dialogue quality, and dialogue efficiency, resulting in a predictor function of the following form:

$$Satisfaction = \sum_{i=1}^{n} w_i * \mathcal{N}(m_i)$$

The m_i terms represent the value of each measure, while the $\mathcal{N}$ function transforms each measure into a normal distribution using z-score normalisation. Stepwise linear regression produces coefficients w_i describing the relative contribution of each predictor to the user satisfaction. If a predictor does not contribute significantly, its w_i value is zero after the stepwise process. **Table 3** shows the predictor functions that we calculated using stepwise multiple linear regression.

Table 2: Objective results. Significant results are displayed in **bold fonts**.

Measure	Supportive mean (stdev)	Instructive mean (stdev)	Mann-Whitney p-value	U-value
No. of user utterances	1.25 (1.94)	1.65 (1.69)	$p = 0.336$	$U = 165.5$
No. of user actions	**4.80 (1.97)**	**0.76 (0.90)**	$\mathbf{p < 0.001}$	$\mathbf{U = 64.0}$
No. of robot instructions	**4.60 (2.28)**	**10.3 (1.49)**	$\mathbf{p < 0.001}$	$\mathbf{U = 8.0}$
No. of correct pickups	**2.10 (2.36)**	**9.35 (1.93)**	$\mathbf{p < 0.001}$	$\mathbf{U = 5.0}$
Assembly duration (seconds)	256.43 (50.47)	265.86 (46.22)	$p = 0.683$	$U = 124.0$

Table 3: Calculated predictor functions using stepwise linear regression. For calculation, four objective measurements were used, which are abbreviated in the table with *Dur* (duration to build both target objects), *AntiPick* (number of anticipatory pick up actions by experiment participant), *InstrPick* (number of pick up actions by experiment participants when the robot instructed them to do so), and *Utt* (number of utterances by experiment participant) .

Measure	Function	R^2	Significance	
Feelings	$198.51 + 1.07 * \mathcal{N}(\text{Dur}) + 43.67 * \mathcal{N}(\text{AntiPick})$ $-36.86 * \mathcal{N}(\text{Utt}) + 26.16 * \mathcal{N}(\text{InstrPick})$	0.28	Dur: $p = 0.060$ Utt: $p = 0.009$	AntiPick: $p = 0.124$ InstrPick: $p = 0.117$
Intelligence	$141.70 + 0.72 * \mathcal{N}(\text{Dur}) + 32.55 * \mathcal{N}(\text{AntiPick})$ $-17.90 * \mathcal{N}(\text{Utt}) + 23.23 * \mathcal{N}(\text{InstrPick})$	0.32	Dur: $p = 0.0375$ Utt: $p = 0.0341$	AntiPick: $p = 0.0620$ InstrPick: $p = 0.026$
Behaviour	$487.33 - 10.96 * \mathcal{N}(\text{AntiPick})$	0.12	AntiPick: $p = 0.051$	
Task success	$447.74 + 0.40 * \mathcal{N}(\text{Dur}) - 17.54 * \mathcal{N}(\text{Utt})$	0.23	Dur: $p = 0.010$	Utt: $p < 0.001$

The calculated predictor functions show that all of the objective measurements influence user satisfaction in one way or the other:

- The number of user utterances has a strongly negative influence on the three categories *feelings of the user* (abbreviated with *Feelings* in table), *intelligence of the robot* (abbr. *Intelligence*), and *task success*. The duration to build both target objects had a slight positive effect in the same three categories.
- The number of anticipatory pick up actions by the participants had a positive influence on categories *feelings of the user* and *intelligence of the robot*, and a negative influence on category *robot behaviour*.
- The number of instructed pick up actions had a strong positive influence on the categories *feelings of the user* and *intelligence of the robot*, but not on the other categories.

The R^2 values of this study are in the same range as the values of our previous user evaluations [11,14]. However, the values are not as high as those reported in [33] and [20].

5.6 Discussion

The results of this study show an interesting correlation: we expected that the experiment participants will prefer the supportive robot over the instructive robot (see **H1**). However, the data suggests that the users accept both robot roles and simply take the counterpart in the interaction with the robot. This can be seen from the significant answers to the statements of the user questionnaire, where the users that worked with the supportive robot answered more positive to the statement "I knew what I could say or do at each point in the conversation". This indicates that the participants showed a more proactive behaviour themselves and followed the assembly plan more by themselves when the robot gave less instructions. This is in line with the work of Hinds et al. [15], who found that humans who work with a robot that takes the role of a supervisor, felt less responsible for the task.

In contrast to that, the users who worked with our instructive robot rated the statement "The robot gave too many instructions" lower than the users from the other group, which we interpret as confirmation for hypothesis **H2**: participants who worked with the instructive robot show a more passive behaviour. One of the objective measurements also supports this claim: users who worked with the supportive robot showed a proactive behaviour and executed anticipatory pick up actions significantly more often than users of the other group. These results are in line with research from cognitive psychology and cognitive neuroscience. For example, Sebanz et al. [25] review a set of studies from these fields, which show that humans attune their actions when working together.

The results of the calculated predictor functions are not very surprising. However, it is interesting to note that the number of anticipatory pick up actions had a positive influence on the statements in the category *feelings of the user* and a negative effect on the category *robot behaviour*. The positive influence on the feelings of the user is a confirmation for hypothesis **H1**: the participants prefer to be proactive, thus a supportive robot fits better to their preferences. The negative effect of these measurements on the assessment of the robot's behaviour can be explained with robot errors during the interaction: when the robot made an error and for example gave the wrong instructions to the user or stopped working (which could happen sometimes during the experiments because of wrongly recognised construction pieces), the users had to pick up the pieces to finish building the target objects without getting instructions by the robot.

The number of user utterances also had a negative influence on the user satisfaction. This can be easily explained: in this experiment, the system was configured so that the participants did not have to speak with the robot, as long as it performed well. The users only had to talk to the robot when they either did not understand the robot's utterances and had to ask for repetition or they needed to give a direct command to the robot to ask for a piece of the robot's work area, which almost only happened when the robot made an error. Thus, the number of user utterances is a clear indicator for problems during the experiment.

One shortcoming of our study is the fact that we did not add an experiment condition in which we told experiment participants which role the robot will take in the interaction. This way, we possibly would have been able to replicate the findings by Stewart and Strasser [27], who showed that the assignment of roles increases effectiveness of collaboration. Then again, the interesting result that shows that participants simply attune their behaviour to the robot role would have been weakened by such an experiment design.

6 Conclusion

In this work, we researched two aspects of role assignment for robots in human-robot interaction: on the one hand, we introduced a new approach for implementing roles on a robot, on the other hand, we showed in an evaluation how humans rate two different robot roles in a collaborative scenario.

Therefore, we first introduced *embodied multimodal fusion* (EMF), a multimodal fusion approach that is based on a representation format in which objects and actions are represented as combined units. This way, EMF maintains a list of all actions that a robot can execute in a given situation at all time points in an interaction. We showed how the EMF action selection mechanism can be used to implement different robot roles. With that, we realised two different roles on a collaborative human-robot interaction system. Secondly, we evaluated these roles with a user study. In this experiment, our robot assembled target objects from a wooden toy construction set together with 40 experiment participants. Here, we used EMF to program the robot to take different roles in the interaction: on the one hand, the robot took the role of an instructor and gave the humans instructions on how to build the target objects before helping them by handing over appropriate construction pieces; on the other hand, the robot took the role of a supporter that directly started handing over construction pieces to its human partner and only gave instructions when necessary. The analysis of the gathered experiment data showed that, in contrast to our expectations, the participants did not prefer one of the two robot roles but simply took the counterpart to the role of the robot and adjusted their own behaviour to the behaviour of the robot. This was shown in one of the objective measurements as well as in the subjective ratings of the users.

To our knowledge, there have been no similar experiments conducted yet to research the role of a robot in such collaborative construction tasks. We see two major findings that arise from our work: (1) in cooperative human-robot interaction, the robot can either take an instructive or a supportive role, the users will accept both of them; (2) however, once the robot role is set, the robot behaviour needs to be consistent to its role, otherwise user satisfaction will decline.

Acknowledgements

This research was supported by the European Commission through the projects JAST (FP6-003747-IP) and JAMES (FP7-270435-STREP). Thanks to Sören Jentzsch for help in annotating the video data.

References

1. Persuadable research survey shows many willing to borrow money to buy a domestic robot, Jan. 2012. http://www.prweb.com/releases/2012-market-research/robot-survey/prweb9140526.htm.
2. B. Biddle. *Role theory: Expectations, identities, and behaviors.* Academic Press New York, 1979.
3. L. Boves, A. Neumann, L. Vuurpijl, L. Bosch, S. Rossignol, R. Engel, and N. Pfleger. Multimodal interaction in architectural design applications. *Lecture Notes In Computer Science*, pages 384–390, 2004.
4. C. Breazeal. Social interactions in HRI: the robot view. *Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 34(2):181–186, 2004.
5. L. Chaimowicz, M. Campos, and V. Kumar. Dynamic role assignment for cooperative robots. In *Robotics and*

Automation, 2002. Proceedings. ICRA'02. IEEE International Conference on, volume 1, pages 293–298. IEEE, 2002.
6. H. Clark and M. Krych. Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50(1):62–81, 2004.
7. P. R. Cohen, M. Johnston, D. McGee, S. Oviatt, J. Pittman, I. Smith, L. Chen, and J. Clow. Quickset: multimodal interaction for distributed applications. In *MULTIMEDIA '97: Proceedings of the fifth ACM international conference on Multimedia*, pages 31–40, New York, NY, USA, 1997. ACM Press.
8. C. Cook, F. Heath, R. L. Thompson, and B. Thompson. Score reliability in web- or internet-based surveys: Unnumbered graphic rating scales versus likert-type scales. *Educational and Psychological Measurement*, 61(4):697–706, 2001.
9. K. Dautenhahn, S. Woods, C. Kaouri, M. Walters, K. Koay, and I. Werry. What is a robot companion-friend, assistant or butler? In *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on*, pages 1192–1197. IEEE, 2005.
10. M. Foster, M. Giuliani, A. Isard, C. Matheson, J. Oberlander, and A. Knoll. Evaluating description and reference strategies in a cooperative human-robot dialogue system. In *International Joint Conference on Artificial Intelligence (IJCAI 2009)*, Pasadena, California, July 2009.
11. M. E. Foster, M. Giuliani, and A. Knoll. Comparing objective and subjective measures of usability in a human-robot dialogue system. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2009)*, Singapore, Aug. 2009.
12. M. E. Foster and C. Matheson. Following assembly plans in cooperative, task-based human-robot dialogue. In *Proceedings of the 12th Workshop on the Semantics and Pragmatics of Dialogue (Londial 2008)*, London, June 2008.
13. J. Gibson. *The ecological approach to visual perception.* Lawrence Erlbaum, 1986.
14. M. Giuliani, M. Foster, A. Isard, C. Matheson, J. Oberlander, and A. Knoll. Situated reference in a hybrid human-robot interaction system. In *International Natural Language Generation Conference (INLG 2010)*, Dublin, Ireland, July 2010.
15. P. Hinds, T. Roberts, and H. Jones. Whose Job Is It Anyway? A Study of Human–Robot Interaction in a Collaborative Task. *Human-Computer Interaction*, 19:151–181, 2004.
16. H. Holzapfel, K. Nickel, and R. Stiefelhagen. Implementation and evaluation of a constraint-based multimodal fusion system for speech and 3d pointing gestures. In *ICMI '04: Proceedings of the 6th international conference on multimodal interfaces*, pages 175–182, New York, NY, USA, 2004. ACM Press.
17. M. Johnston, S. Bangalore, G. Vasireddy, A. Stent, P. Ehlen, M. Walker, S. Whittaker, and P. Maloor. Match: An architecture for multimodal dialogue systems. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 376–383, 2002.
18. M. Kipp. Multimedia annotation, querying and analysis in ANVIL. *Multimedia Information Extraction*, 2010.
19. N. Krüger, J. Piater, F. Wörgötter, C. Geib, R. Petrick, M. Steedman, A. Ude, T. Asfour, D. Kraft, D. Omrcen, et al. A Formal Definition of Object-Action Complexes and Examples at Different Levels of the Processing Hierarchy. *PACO+ Technical Report, available from http://www.paco-plus.org*, 2009.
20. D. Litman and S. Pan. Designing and evaluating an adaptive spoken dialogue system. *User Modeling and User-Adapted Interaction*, 12(2–3):111–137, 2002.
21. R. Looije, M. Neerincx, and G. Kruijff. Affective Collaborative Robots for Safety & Crisis Management in the Field. In *Intelligent Human Computer Systems for Crisis Response and Management (ISCRAM 2007)*, Delft, Netherlands, May 2007.
22. S. Meyer. *Mein Freund der Roboter: Servicerobotik für ältere Menschen - eine Antwort auf den demographischen Wandel?* VDE-Verlag, 2011.
23. T. Müller, P. Ziaie, and A. Knoll. A wait-free realtime system for optimal distribution of vision tasks on multicore architectures. In *Proc. 5th International Conference on Informatics in Control, Automation and Robotics*, May 2008.
24. L. Ross, T. Amabile, and J. Steinmetz. Social roles, social control, and biases in social-perception processes. *Journal of Personality and Social Psychology; Journal of Personality and Social Psychology*, 35(7):485–494, 1977.
25. N. Sebanz, H. Bekkering, and G. Knoblich. Joint action: bodies and minds moving together. *Trends in Cognitive Sciences*, 10(2):70–76, 2006.
26. L. Sentis and O. Khatib. Task-oriented control of humanoid robots through prioritization. In *Proceedings of the IEEE-RAS/RSJ International Conference on Humanoid Robots*, 2004.
27. D. Stewart and G. Stasser. Expert role assignment and information sampling during collective recall and decision making. *Journal of Personality and Social Psychology*, 69(4):619, 1995.
28. R. Stiefelhagen, H. Ekenel, C. Fugen, P. Gieselmann, H. Holzapfel, F. Kraft, K. Nickel, M. Voit, and A. Waibel. Enabling multimodal human–robot interaction for the karlsruhe humanoid robot. *Robotics, IEEE Transactions on*, 23(5):840–851, 2007.
29. P. Stone and M. Veloso. Task decomposition and dynamic role assignment for real-time strategic teamwork. *Intelligent Agents V: Agents Theories, Architectures, and Languages*, pages 293–308, 1999.
30. A. Tapus and M. Mataric. Socially assistive robots: The link between personality, empathy, physiological signals, and task performance. In *AAAI Spring*, volume 8, 2008.
31. A. J. N. van Breemen. iCat: Experimenting with animabotics. In *AISB 2005 Creative Robotics Symposium*, 2005.
32. W. Wahlster. *SmartKom: Foundations of Multimodal Dialogue Systems.* Springer, 2006.
33. M. Walker, C. Kamm, and D. Litman. Towards developing general models of usability with PARADISE. *Natural Language Engineering*, 6(3–4):363–377, 2000.
34. P. Ziaie, T. Müller, and A. Knoll. A novel approach to hand-gesture recognition in a human-robot dialog system. In *Proceedings of the First Intl. Workshop on Image Processing Theory, Tools & Applications*, Sousse, Tunesia, Nov. 2008.
35. P. Zimbardo and A. Cross. *Stanford prison experiment.* Stanford University, 1971.

Manuel Giuliani is a research associate at fortiss, an affiliated institute of the Technical University Munich. He received a Master of Arts in computational linguistics from the Ludwig-Maximilian-University in Munich, a Master of Science in computer science from

the Technical University Munich, and a PhD in computer science from the Technical University Munich. He worked on the European project JAST (Joint Action Science and Technology), the DFG-funded project AudiComm, which is part of the cluster of excellence "Cognition for Technical Systems" (CoTeSys), and is now part of the European project JAMES (Joint Action for Multimodal Embodied Social Systems). His research interests include social robotics, human-robot interaction, natural language processing, multimodal fusion, and robot architectures.

Alois Knoll is a professor at the Technical University Munich. He received a PhD in computer science from the Technical University of Berlin, and the qualification for teaching computer science at a university (habilitation) at the Technical University of Berlin. He is on the board of directors of the Central Institute of Medical Technology at TUM (IMETUM-Garching). Between April 2004 and March 2006 he was Executive Director of the Institute of Computer Science at TUM. He is an executive board member of the Graduate School of Information Science in Health (GSISH). Furthermore, he is one of the founders of fortiss, an innovation hotspot for software-intensive systems at the Technical University Munich. His research interests include cognitive, medical and sensor-based robotics, multi-agent systems, data fusion, adaptive systems, and multimedia information retrieval.