

Policy Adaptation through Tactile Correction

Brenna D. Argall and Eric L. Sauser and Aude G. Billard¹

Abstract. Behavior adaptation based on execution experience can be a practical tool to increase the robustness of a robot behavior learned from demonstration. While demonstration learning is a powerful technique for the development of robot behaviors, in general development remains a challenge. This work presents an approach for policy improvement through a *tactile* interface located on the body of the robot. We introduce the *Tactile Policy Correction (TPC)* algorithm, that employs tactile feedback for the adaptation of a policy learned from demonstration. We provide an initial validation of refinement under the TPC algorithm on humanoid robot performing a grasp positioning task, and policy performance is found to improve with tactile corrections. We additionally show different modalities, namely teleoperation and tactile corrections, to provide information about allowable variability in the target behavior in different areas of the state space.

1 Introduction

Motion control is fundamental to many robotics applications, yet the development of paradigms for control remains a challenge. Difficulties such as noisy sensors and inaccurate world models, for example, often confound the development of control *policies*, that map world state to robot actions. Moreover, many of the challenges associated with policy development only grow with domain and robot complexity, for example high degree of freedom humanoids.

Policy development therefore typically involves a significant measure of expertise and effort, prompting the advancement of techniques to reduce the requirements placed on a developer. One option is to increase policy robustness through adaptation in response to execution experience. Some development effort also might be mitigated if a new policy is built by bootstrapping from an existing policy. The approach to policy development taken in this work is founded on both ideas, of policy refinement and reuse. The policy modifications furthermore are guided by human teacher feedback, a fundamental characteristic of this work.

We present *Tactile Policy Correction (TPC)* as an algorithm that incorporates human feedback in the form of tactile corrections for the purposes of policy adaptation (Fig. 1). Policy corrections are indicated through the touch of a human teacher, and the teacher provides corrections in order to accomplish both policy refinement and reuse. More specifically, our approach initially derives a policy via *Learning from Demonstration (LfD)* techniques. Under LfD, the robot generalizes a policy from data recorded during demonstration executions by a task expert. Our generalization formulation additionally produces policies that reflect the inherent *variability* seen within the teacher demonstrations, allowing for greater flexibility in the result-

ing behavior during learner reproduction. Policy *corrections* from a human teacher then are provided through a tactile interface. The policy predictions provide target poses for the robot body, and when a correction is indicated, the robot immediately modifies its pose to accommodate the adjustment. The resulting execution then is treated as new demonstration data for the policy.

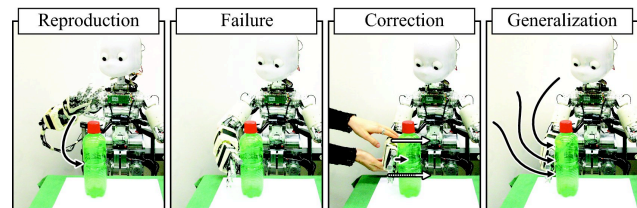


Figure 1. Our approach of policy adaptation via corrections provided through a tactile interface on the body of the robot. Black arrows indicate robot motion and white arrows human hand movement.

We provide results from a preliminary validation of policy refinement under TPC, with a humanoid robot performing a grasp-positioning task. We show the policies produced under our approach to both be flexible with respect to teacher demonstration variability, as well as to improve in performance when provided with tactile corrections. We additionally find that different teaching modalities (i.e. task demonstration, tactile correction) provide information about the allowable variability in task execution within different areas of the execution state space.

The following section provides related literature and motivation for our approach. The TPC algorithm is detailed in Section 3, along with our approach for preserving demonstration variability in the learned policy. Initial validation results on humanoid robot and discussion are presented in Section 4, and in Section 5 we conclude.

2 Background and Motivation

This section begins with a discussion of policy development and refinement under *Learning from Demonstration (LfD)*, proposes the use of tactile corrections for refinement and then delineates the motivating factors for our approach.

2.1 Learning from Demonstration

Under LfD, teacher executions of a desired behavior are recorded and a policy is derived from the resultant dataset. LfD has seen success on a variety of robotics applications, and has the attractive characteristics of being an intuitive means for human teacher to robot

¹ Learning Algorithms and Systems Laboratory (LASA), École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, email: [brennadee.argall, eric.sauser, aude.billard]@epfl.ch

learner knowledge transfer, as well as being an accessible policy development technique for those who are not robotics-experts. There are many design decisions to consider when building a LfD system. These range from who executes the demonstrations and how they are recorded, to the technique used for policy derivation. Here we overview only those decisions specific to our particular system, and refer the reader to [2] and [5] for a full review of robot LfD.

When recording and executing demonstrations the issue of *correspondence* is key, where teacher demonstrations do not directly map to the robot learner due to differences in sensing or motion [13]. Correspondence issues are minimized when the learner records directly from its own sensors while under the control of the teacher. For example, under *teleoperation* the teacher remotely controls the robot platform (e.g. [18]), while under *kinesthetic demonstration* the teacher touches the robot for control (e.g. [7]). Teleoperation requires an interface for direct robot control, and possibly also a translation by the teacher to the body of the robot; kinesthetic teaching requires a (passive or active) responsiveness to human touch. Both techniques are employed in our work.

Policy derivation then amounts to building a predictor that reproduces the demonstrated behavior. Many approaches exist within LfD to derive a policy from the demonstration data [2], the most popular of which either directly approximate the underlying function mapping observations to actions, or approximate a state transition model and then derive a policy using techniques such as *Reinforcement Learning (RL)*. Our work derives a policy under a variant of the first approach, where probabilistic regression techniques are used to predict a target robot pose based on world state, and a controller external to the algorithm selects an action able to accomplish this target pose. Our reason for splitting policy prediction into these two steps is tied to the mechanism by which the algorithm responds to teacher feedback, discussed in Section 3.1.

2.2 Policy Refinement and Reuse

Even with the advantages secured through demonstration, policy development typically is still a non-trivial task. Here we consider two techniques to assist the policy development process, or equivalently to reduce the strain on the policy developer: *policy refinement* and *policy reuse*. To have a robot learn from its execution performance, or *experience*, is a valuable policy improvement tool for any development technique, and execution experience may be used to refine a policy. The ability to reuse existing policies, designed to address related tasks, is also a practical feature for any policy learning system.

Within the context of LfD specifically, execution experience can be used to overcome limitations in the demonstration dataset. One typical limitation is dataset sparsity, since demonstration from every world state is infeasible in all but the simplest domains. Other limitations include the issue of correspondence between the teacher and learner or deficiencies in the performance of the teacher, who may in fact provide suboptimal or ambiguous demonstrations. Overcoming potential dataset limitations is key to good policy performance, since a LfD policy depends heavily on the quality of the demonstration data from which it is derived.

Under LfD, a variety of approaches incorporate information gathered from experience in order to *refine* a policy. For example, execution experience is used to update reward-determined state values [17] and learned state transition models [1]. Other approaches provide more demonstration data, driven by more teacher-initiated demonstrations [7] as well as by learner requests for more data [8, 9]. In this work, we also provide more data, but through a different mechanism

than that which provides the initial teacher demonstrations. Policy *reuse* under LfD occurs most frequently with behavior primitives, or simpler policies that contribute to the execution of a more complex policy. Hand-coded behavior primitives are used within tasks learned from demonstration [14], demonstrated primitives are combined into a new policy by a human [16] or automatically by the algorithm [2], and demonstrated tasks are decomposed into a library of primitives [4]. The focus of our approach is instead on adapting an existing policy to accomplish a *different* task, rather than incorporating the existing behavior as a subcomponent of a larger task.²

2.3 Tactile Corrections

A fundamental consideration in our work is the mechanism used to provide feedback to the learner; for example, feedback provided verbally to indicate successful task completion [11]. The feedback form we explore is this work is that of *tactile corrections*.

To correct poor policy predictions is a particularly direct approach to addressing potential LfD limitations. In comparison to overall performance evaluations or state rewards, that can provide an indication of the quality of a policy prediction, corrections furthermore provide guidance on what might have been a more suitable alternate prediction. While more teacher demonstration can populate sparse datasets or provide an improved behavior example, it also requires visiting the state in which the policy requires improvement, which can be impractical for real world domains.

Within LfD, policy correction has seen limited attention. In most approaches a human teacher indicates the correct prediction from a discrete set of actions with significant time duration [8, 14]. Even fewer provide corrections within *continuous* state-action spaces sampled at a *rapid* rate: both characteristics of low-level motion control, which is the target application domain for our work. To accommodate these constraints, our approach translates feedback from a tactile sensor into continuous-valued modifications on the current pose as the robot executes. In contrast to other works with continuous-valued corrections [2], here corrective feedback is offered *online*, instead of post-execution, and through a *tactile interface*, instead of a high-level corrective language.

We posit that tactile feedback furthers many of the strengths of demonstration-based learning. Namely, humans use touch to instruct other humans in certain contexts, for example when demonstrating a pose or motion that requires a particular position or trajectory in 3-D space, like a ballet posture or tennis swing. To augment demonstration learning with tactile feedback therefore is one natural extension to the idea of teaching robots as humans teach other humans. Another attractive feature of demonstration-based policy development is accessibility to those who are not robotics experts. Policy development by non-experts furthermore strongly suggests robot operation around humans, in which case the detection of tactile interactions can be crucial for safe robot operation. Tactile sensing thus gains importance on a very fundamental level, and the detected tactile interactions may be exploited further for knowledge transfer from human to robot.

Within the field of *Human-Robot Interactions (HRI)*, a handful of works utilize human touch for the development of robot behaviors. For example, during humanoid behavior learning, tactile feedback is detected in order to minimize the support forces provided by a

² We emphasize that the initial validation of TPC provided in this paper addresses policy refinement only. When used for reuse instead of refinement, however, the algorithm remains unchanged. Rather, it is only the intent of the teacher that changes, who guides the learner motion to exhibit a different behavior from that which was demonstrated.

teacher [12], and during interactions between a robotic pet-surrogate and elderly patients, tactile reward signals are used within RL to adapt behavior selection [19].

2.4 Motivation for our Approach

In summary, the approach presented in this paper employs *tactile* corrections to modify a policy learned through demonstration, for the purposes of policy *refinement* and *reuse*. The factors motivating this approach include the following.

The first factor is our target application domain: low-level motion control for high DoF robots. Policy development for high DoF robots is difficult. To specify a target behavior for each joint is complicated, and systems typically are under-constrained, resulting in many joint configurations mapping to a single end-effector pose. Demonstration-based policy development thus has many advantages. Correspondence differences motivate the use of teleoperation for demonstration, though we note that from an implementation standpoint teleoperation becomes more challenging when many degrees of freedom must be controlled.

The second factor is to overcome potential limitations in the demonstration dataset. For example, suboptimal behavior examples can result from a poor interface for controlling a demonstration, and other potential drawbacks include dataset sparsity. We aim to overcome dataset limitations through policy refinement. To accomplish policy refinement, the approach presented in this paper provides new behavior examples. The source for these examples, however, is *not* more teacher demonstration. Instead, we have the student respond online to corrections indicated by a teacher, and the resultant trajectory is treated as new training data for the policy.

The final factor is how to indicate the policy corrections that result in the new behavior examples. Our approach provides corrections through a tactile interface. We argue that information transfer through human touch is a natural extension of human demonstration as an intuitive and effective mechanism for the transfer of knowledge from human to robot, that furthermore is relatively unaddressed within the LfD literature to date.

3 The Tactile Policy Correction Algorithm

We introduce *Tactile Policy Correction (TPC)* as an algorithm for the refinement and reuse of motion policies, accomplished via tactile feedback from a human teacher.

Under TPC, a policy is initially derived from demonstrations of a task by a teacher. We formally define the world to consist of actions $A \in \mathbb{R}^l$ and observations $Z \in \mathbb{R}^{(m+n)}$ of world state. An observation $z \in Z$ consists of two components, $z = (z_\varphi, z_{-\varphi})$, where $z_\varphi \in \mathbb{R}^m$ describes the robot pose, and $z_{-\varphi} \in \mathbb{R}^n$ describes any other observables that are of interest to the policy.³ We define a *demonstration* to consist of a sequence of observations $\{z^j\}_{j=1}^T$, recorded during teacher execution of the task. The collected set $D = \{z^j\}_{j=1}^N$ of demonstrations is then provided to the robot learner. From this set a policy $\pi : Z \rightarrow A$ is derived, that enables the selection of an appropriate action given the observed state.

Following demonstration and policy derivation, the robot executes with its policy and receives tactile corrections from the human teacher. Tactile corrections are used within two capacities, either to refine the existing policy or to build a new policy bootstrapped on

the demonstrated policy. Pseudo-code for this approach is provided in Algorithm 1.

Algorithm 1 Tactile Policy Correction

```

1: Given  $D$ 
2: initialize  $\delta^0 \leftarrow \mathbf{0}$ 
3: derive  $\pi \leftarrow \text{policyDerivation}(D)$ 
4: while correcting do
5:   Policy  $\pi$  execution:
6:     predict  $\hat{z}_\varphi^t \leftarrow \text{regression}(z^{t-1})$ 
7:     execute  $z_\varphi^t \leftarrow \text{controller}(\hat{z}_\varphi^t + \delta^t)$ 
8:     if detect touch then
9:       map  $\delta_\epsilon^t \leftarrow M(\text{touch})$ 
10:      correct  $z_\varphi^t \leftarrow \text{controller}(z_\varphi^t + \delta_\epsilon^t)$ 
11:      record  $\delta^{t+1} \leftarrow \delta^t + \delta_\epsilon^t$ 
12:     end if
13:     set  $w^t$ 
14:     record  $D \leftarrow D \cup (z^t, w^t)$ 
15:   end while
16: rederive  $\pi \leftarrow \text{policyDerivation}(D)$ 
17: return  $\pi$ 

```

3.1 Algorithm Execution

The first phase of the TPC algorithm consists of task demonstration by the teacher, producing dataset D from which the learner derives an initial policy π . The second phase of the algorithm involves learner execution with the policy π , and corrective tactile feedback which is used to update π . This *execution-correction-update* cycle continues to the satisfaction of the teacher.

A single execution-correction-update cycle is presented in lines 4-16 of Algorithm 1. Policy execution (lines 5-7) at timestep t consists of two phases: prediction of a target pose $\hat{z}_\varphi^t$, and the selection of an action to accomplish that pose. Pose prediction is accomplished via regression techniques, based on state observation z^{t-1} (line 6). Action selection is accomplished via a robot-specific controller, and its execution results in a new robot pose z_φ^t (line 7).

The human teacher may choose to offer a tactile correction at any timestep of an execution. If detected, the robot learner translates the tactile feedback into an incremental shift $\delta_\epsilon^t \in \mathbb{R}^m$ in the robot pose, according to mapping M (line 9). Note that the form taken by the tactile feedback is platform-specific, depending both on the tactile sensors employed to detect human touch and how the sensor feedback is processed.

The robot controller is then passed the new *adjusted* pose, for which the incremental shift δ_ϵ^t is added to the current robot pose z_φ^t (line 10). The influence of this incremental shift is maintained over multiple timesteps, through an offset parameter $\delta \in \mathbb{R}^m$ that maintains a sum of all adjustments seen during the execution (line 11) and is added to the pose prediction at each execution timestep (line 7).

The timestep concludes with the recording of observation z^t , along with a weight $w^t \in [0, 1]$ for the new datapoint (details in Sec. 3.3.2), into the set D (line 14). The tactile correction thus also is recorded, since the current pose has been corrected by tactile feedback and is recorded into observation z^t through component z_φ^t . Upon completion of the entire execution, policy π is rederived from demonstration set D (lines 16); the corrected execution thus is treated as new data for the policy.

Important to note is that the TPC algorithm is agnostic to the techniques used for pose prediction (*regression*) and action selection

³ Pose information is necessary for the TPC algorithm, and so $z_\varphi \neq \emptyset$. The presence of additional observation information however is application-dependent, and possibly absent such that $z_{-\varphi} = \emptyset$.

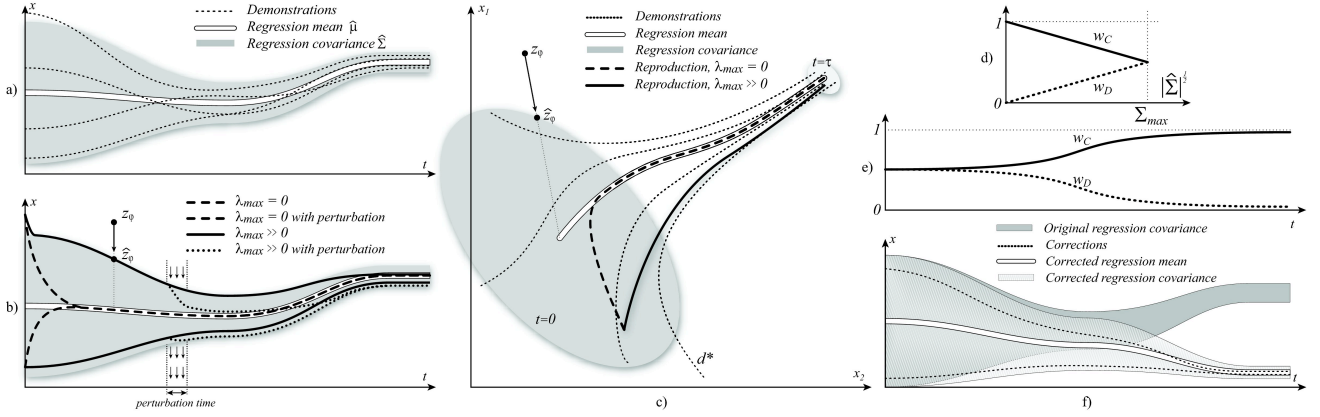


Figure 2. a) Demonstration data and resulting GMR regression mean and covariance envelope. b,c) Illustration of our offset formulation for GMR that allows for deviation from the regression mean, showing adaptability with respect to perturbations (b, dotted lines) and starting position (c, bold line). d) Our weight function formulation, as a function of covariance envelope size. e,f) Illustration of an example weight function and the resulting shift in regression signal.

(controller) during policy execution, as well as to the technique that translates tactile feedback into a pose adjustment (mapping M). The following sections describe the particular techniques we employ for the implementation of the TPC algorithm within this work.

3.2 Policy Execution

This section describes the specific techniques used for policy execution under our implementation of the TPC algorithm. For pose prediction, the *GMM-GMR* regression technique is employed (Sec. 3.2.1), with a modification to allow for variability in the resulting policy (Sec. 3.2.2). For action selection we use an inverse kinematic controller (Sec. 3.2.3).

3.2.1 Pose Prediction

Target poses are predicted through the *GMM-GMR* algorithm [7], which first encodes demonstrations in a *Gaussian Mixture Model* (*GMM*) and then predicts a target pose through *Gaussian Mixture Regression* (*GMR*). The recorded demonstrations are modeled probabilistically in a GMM, whose parameters are trained under a weighted version of the *Expectation-Maximization* (*EM*) algorithm (details in Sec. 3.3.2). Our implementation defines observation component z_φ as Cartesian position $x \in \mathbb{R}^3$ and orientation $q \in \mathbb{R}^4$ (as a quaternion, $\|q\| = 1$) of the end-effector in a robot-centric reference frame. Thus $z_\varphi \equiv [x, q] \in \mathbb{R}^7$. We further define component $z_{-\varphi} \equiv t \in \mathbb{R}$ as the timestep of recorded observation. The GMM thus models the joint probability of the temporal and spatial aspects of the demonstrations. To make a pose prediction, GMR estimates the conditional expectation of z_φ given $z_{-\varphi}$, i.e. $p(x, q|t)$.

3.2.2 Deviating from the Mean Trajectory

We additionally take advantage of the probabilistic nature of the regression to generate variability, and thus flexibility, in the predicted trajectory. Under GMR, a target pose $\hat{z}_\varphi^t$ is predicted with mean $\hat{\mu}^t$ and covariance $\hat{\Sigma}^t$ (Fig. 2a). We modify the pose prediction by

$$\hat{z}_\varphi^t = \hat{\mu}^t + \delta_\lambda^t \quad (1)$$

and thus apply to the regression mean offset $\delta_\lambda^t \in \mathbb{R}^m$

$$\delta_\lambda^t = \begin{cases} \Delta_\lambda^t & \text{if } \lambda^t \leq \lambda_{\max} \\ \Delta_\lambda^t \frac{\lambda_{\max}}{\lambda^t} & \text{otherwise} \end{cases} \quad (2)$$

$$\Delta_\lambda^t = z_\varphi^t - \hat{\mu}^t, \quad \lambda^t = \|(\hat{\Sigma}^t)^{-\frac{1}{2}} \Delta_\lambda^t\| \quad (3)$$

where δ_λ^t is defined by the difference between the current robot pose and regression mean (Δ_λ^t), and whether the magnitude (λ^t) of this difference (inversely scaled by standard deviation $(\hat{\Sigma}^t)^{\frac{1}{2}}$) exceeds a threshold ($\lambda_{\max}$). Without this offset, the pose predictions follow exactly the mean trajectory. An example where this would not be desired is if the execution starting position is actually closer to the target position than is the start of the mean trajectory. Predicting the mean trajectory in this case will cause the learner execution to backtrack unnecessarily. With our offset, the pose predictions are no longer constrained to only follow the mean trajectory (Fig. 2b,c, $\lambda_{\max} \gg 0$). Instead, variations are allowed within the constraints of the regression covariance.

The amount of allowable deviation is dictated in terms of an acceptable number ($\lambda_{\max}$) of standard deviations from the regression mean, where $\lambda_{\max} \geq 0$ is a constant parameter set by hand (in our empirical validation, $\lambda_{\max} = 2$). For execution points (including starting positions) within this threshold (i.e. within $\lambda_{\max}$ standard deviations of the regression mean $\hat{\mu}^t$), the execution proceeds with its current pose (i.e. $\hat{z}_\varphi^t = \hat{\mu}^t + \Delta_\lambda^t = z_\varphi^t$). Execution points outside of this threshold are first projected (e.g. Fig. 2c, z_φ to $\hat{z}_\varphi$) to the envelope (shaded region) defined by $\lambda_{\max}$ standard deviations around the regression mean. The result is more flexible learner executions, that take advantage of the variability present in the teacher demonstrations.

One gain of this regression formulation, that will be confirmed in Section 4.2, is allowing the learner execution to choose a more direct path to the goal, that perhaps deviates from the mean trajectory but is still within the bounds of what was demonstrated. Another potential gain in using offset δ_λ^t is online adaptation to external perturbations, for example a compliant response to a perturbation caused by unexpected contact with a human. Consider an external force applied to the robot arm during execution, (i) to which we would like the robot to respond (ii) while still attaining the goal (e.g. Fig. 2b). Without

our offset, the learner execution will attain the second objective, but not the first; that is, the learner will continue to follow the mean trajectory, and thus attain the target position, but will not be responsive to the force ($\lambda_{\max} = 0$, dashed lines). With our offset, however, both objectives are attained ($\lambda_{\max} \gg 0$, dotted lines). The offset value is the execution's response to the perturbation, and since the offset is constrained by the covariance envelope (more specifically, to keep the execution within $\lambda_{\max}$ standard deviations of the regression mean), the target position also will be attained.⁴

3.2.3 Action Selection

Given a target pose $\hat{z}_\varphi$, action selection is accomplished via an inverse kinematic controller. Our action space A consists of velocities $\dot{\theta} \in \mathbb{R}^7$ for the joint angles of a robot arm. The manipulator of our implementation (Sec. 4.1) is redundant, as the number of degrees of freedom (7) exceeds the number of constraints (6, position and orientation). We therefore compute desired joint angle velocities $\dot{\theta}$ according to the distance between the target pose $\hat{z}_\varphi$ and the current robot pose z_φ by using a pseudo-inverse method that both avoids joint limits and is robust to singularities [3].

3.3 Tactile Corrections

The tactile interface consists of five *Ergonomic Touchpads* located on the manipulator arm. The pads detect contact presence and relative motion, which we map to a change in end-effector position and orientation (Sec. 3.3.1). The movement that results is recorded into the dataset and incorporated into a policy update (Sec. 3.3.2).

3.3.1 Online Modification of Policy Execution

Four touchpads, $T_0 \dots T_3$, encircle the lower forearm of the robot arm (near the wrist), and one, T_4 , is located on the back of the robot hand (Fig. 3a,b). In practice, we decompose the mapping M into two distinct parts, that operate separately on the wrist (end-effector) and on the hand of the robot arm, which seemed a more intuitive mapping for the experimenters providing corrections.⁵

The first part of the mapping operates on the first 5-DoF leading to the wrist of our 7-DoF manipulator. The pads $T_{\{0..3\}}$ can be seen as an interface for controlling the manipulator along the 6-DoF Cartesian space (position and orientation). Sliding the fingers along two opposite touchpads can either lead to a translational or rotational motion command, depending on whether the sliding directions agree or not (Figure 3c,d). When motion is detected on the touchpads, the commands are mapped to a target velocity for the end-effector in Cartesian space, and then resolved to target joint velocities by an inverse kinematic controller [3]. The second part of the mapping relates to the final 2-DoF controlling the robot hand. Since the last pad has 2-DoF as well, here the mapping to motion commands is one to one.

⁴ With a reasonably-sized factor $\lambda_{\max}$ (i.e. reasonably small, such as $\lambda_{\max} = 2$), attaining the target position with our formulation depends primarily on the covariance envelope being small at the target position, which the unmodified formulation of GMR relies on as well.

⁵ Touchpad feedback is somewhat limited in comparison to more sophisticated tactile sensors, for example that provide force information or a finer spatial resolution. In practice corrective repositioning is not always as responsive as the teacher requires, and so we pause policy execution such that pseudo-code lines 9-10 loop until repositioning is complete. Note that this limitation results from a deficiency in *hardware*, not the algorithm, and validation with a more sophisticated tactile sensor is a target for future work.

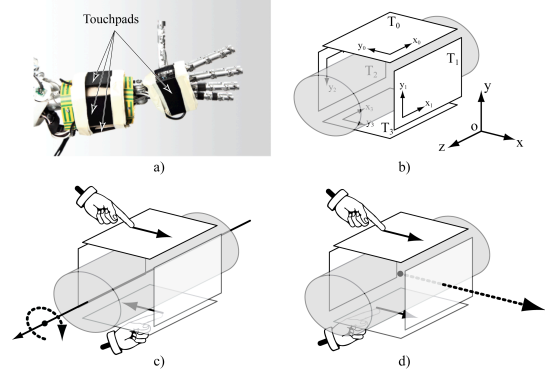


Figure 3. a,b) Schematic of the touch pads controlling the robot wrist and hand. c,d) Fingers sliding on opposite pads produces rotational (c) or translational (d) motions.

3.3.2 Incorporation into a Policy Update

Upon completion of an execution corrected with tactile feedback, new data is incorporated into the policy. To incorporate feedback, the policy is rederived taking into account the new datapoints. Policy execution within the TPC algorithm consists of a pose prediction via regression techniques, followed by action selection by a controller. Under our implementation the controller is statically defined, and so policy derivation consists of regression parameter estimation only. Policy rederivation thus consists of re-estimating the regression parameters, which is accomplished through a weighted version of the EM algorithm (details in Tbl. 1).

Table 1. Weighted Expectation-Maximization (EM)

Our weighted version of the EM algorithm modifies the EM implementation of [7] to include weight w^j . The algorithm loops between the *E-step* and the *M-step* until the overall likelihood $\sum_{k=1}^K E_k$ is maximized:

$$p_{k,j}^{(i+1)} = \frac{\gamma_k^{(i)} \mathcal{N}(z^j; \mu_k^{(i)}, \Sigma_k^{(i)})}{\sum_{i_k=1}^K \gamma_{i_k}^{(i)} \mathcal{N}(z^j; \mu_{i_k}^{(i)}, \Sigma_{i_k}^{(i)})}$$

$$E_k^{(i+1)} = \sum_{j=1}^N w^j p_{k,j}^{(i+1)}$$

M-step:

$$\gamma_k^{(i+1)} = \frac{E_k^{(i+1)}}{\sum_{j=1}^N w^j}$$

$$\mu_k^{(i+1)} = \frac{\sum_{j=1}^N w^j p_{k,j}^{(i+1)} z^j}{E_k^{(i+1)}}$$

$$\Sigma_k^{(i+1)} = \frac{\sum_{j=1}^N w^j p_{k,j}^{(i+1)} (z^j - \mu_k^{(i+1)}) (z^j - \mu_k^{(i+1)})^T}{E_k^{(i+1)}}$$

Datapoint weights are assigned based on the covariance envelope of the original GMM derived from the demonstration data. In particular, we define weight functions for corrected executions $w_C(t)$ and

demonstrated executions $w_D(t)$ as

$$w_C(t) = 1 - \frac{|\hat{\Sigma}^t|^{\frac{1}{2}}}{2 \cdot \Sigma_{\max}} \quad , \quad \Sigma_{\max} = \max_t |\hat{\Sigma}^t|^{\frac{1}{2}} \quad (4)$$

$$w_D(t) = 1 - w_C(t) \quad (5)$$

where $|\hat{\Sigma}^t|$ is the determinant of the GMR prediction covariance matrix at time t . We then assign weight w^j for datapoint z^j with functions $w_D(t)$ or $w_C(t)$, based on whether z^j was part of a demonstrated or corrected execution (respectively) and the time ($t \equiv z_{-\varphi}^j$) of the observation recording.

With this weight formulation, we assume that the teacher demonstrations provide an accurate portrayal of the variability profile of the task. That is, in areas of low covariance, little variability is allowed (or equivalently, high precision is required) in the target task behavior, while in areas of high covariance, much variability in the resulting behavior is acceptable, even expected. With our weight formulation, in areas of low covariance ($|\hat{\Sigma}^t|^{\frac{1}{2}} \ll \Sigma_{\max}$), corrected datapoints are given a high weight, and the regression signal accordingly shifts strongly. By contrast, in areas of high covariance ($|\hat{\Sigma}^t|^{\frac{1}{2}} \rightarrow \Sigma_{\max}$), it is not unexpected that executions might differ from the demonstrated behavior, and so demonstrated and corrected execution points are given approximately equal weight. This weight formulation is shown in Figure 2d, followed by an example weight function (e) and resulting shifted regression signal (f).

4 Empirical Validation

Here we provide initial validation results of TCP on a high DoF humanoid (Sec. 4.1), highlighting in particular the flexibility of our regression formulation (Sec. 4.2) and refinement of the demonstrated behavior (Sec. 4.3). A discussion of these results, and future research directions, are also provided (Sec. 4.4).

4.1 Robot Task and Domain

Initial validation of the TPC algorithm is performed with a 57-DoF humanoid robot, the *iCub* robot, performing a grasp positioning task. The task consists of positioning the robot end-effector (7-DoF) to grasp a cylindrical object. Demonstration is performed via teleoperation, accomplished with a *3D mouse* able to control all 6-DoF of the end-effector position and orientation. Closing the hand for grasping is handled by a static controller.⁶ Demonstrations are provided from multiple starting end-effector positions with respect to the object. Tactile corrections are then performed using the touchpad setup described in Section 3.3.1.

The focus of this validation is two-fold: (i) to explore policy *flexibility* with respect to the variability present within the teacher demonstrations and (ii) to confirm the ability of TPC to *refine* a demonstrated behavior. The validation of policy *reuse* is left for future work. Three policies therefore are developed for comparison:

π : Derived from the demonstration set D using GMR as in [7].

π_λ : Derived from the demonstration set D and using our modified version of GMR with offset δ_λ .

$\pi_{\lambda,c}$: Produced from tactile correction of Policy π_λ using TPC.

⁶ The focus of the task objective is on end-effector positioning, rather than the grasp itself, since the *iCub* hand has no force sensors or tactile feedback. Note also that if controlling the hand is also a part of the demonstrations, then the operational space is 15-DoF and a more complex teleoperation system is required.

4.2 Flexible and More Efficient Policies

Flexibility in the policy execution is enabled through offset factor δ_λ . As illustrated in Figure 2c, without this offset the execution trajectory (dashed line) will follow the regression mean (white trajectory), regardless of whether a more appropriate path (e.g. a shorter path such as demonstration d^*) is contained within the demonstration set executions. With the offset, however, the learner execution is free to follow a more direct path to the goal (bold line), providing this is within $\lambda_{\max}$ standard deviations of the regression mean. The executions in Figure 4 confirm this behavior with real robot data.

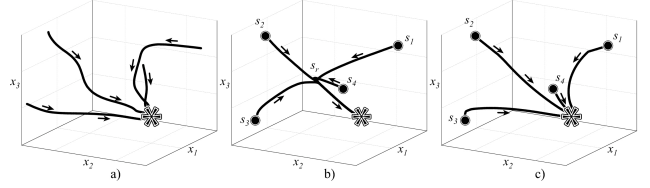


Figure 4. a) Demonstration executions to target position *. b,c) Executions from starting positions $s_1 \dots s_4$, performed by policy π (b) and policy π_λ (c). Note that executions by policy π_λ proceed directly to the target position, while those of π first visit the start of the mean trajectory (s_r).

Table 2 provides the lengths of the execution trajectories (as fractions of the distance traveled by policy π) from 4 starting positions ($s_1 \dots s_4$) for all policies. Indeed, from all positions the incorporation of offset δ_λ allows for execution paths that approach the target position more directly, shown through shorter trajectory lengths (π_λ vs. π). The most dramatic improvement is seen with starting position s_4 , whose position is such that the execution must travel explicitly away from the target position (*) to reach the start of the mean trajectory (s_r). In this case overt backtracking is the result if offset δ_λ is not employed (Fig. 4).

Table 2. Execution Length (from multiple starting positions, as a fraction of the execution length of policy π)

Starting Position	π	π_λ	$\pi_{\lambda,c}$
s_1	1	0.69	0.66
s_2	1	0.88	0.88
s_3	1	0.64	0.67
s_4	1	0.35	0.27

In the original demonstration set D , variability in the starting position was present. The task also allows for some variability in the *target* position, as the hand may be positioned at various locations on the cylinder and still successfully grasp the object. Variability in target position was minimally present in the demonstration set however, as navigating the end-effector to various grasp locations on the cylinder required a high level of precision that was difficult to achieve with the mechanism used for teleoperation (the *3D mouse*). Through tactile corrections, however, the teacher was able to convey variability with respect to target position.

As shown in Figure 5, following correction, the covariance envelope at the target position broadened with respect to location *on* the cylinder; that is, along the principle axis of the cylinder (dimension x_3). Note further that the envelope by contrast narrowed within the plane supporting the cylinder (dimensions x_1, x_2), appropriately indicating less flexibility with respect to the location *of* the cylinder

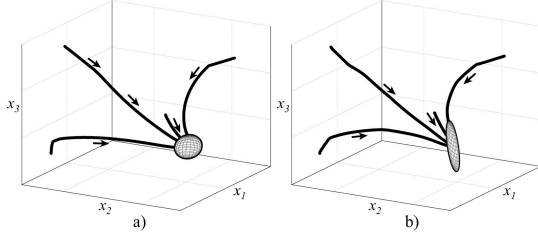


Figure 5. Covariance envelope (shaded region) at the target position a) before tactile corrections (i.e. with policy π_λ) and b) after tactile corrections (i.e. with policy $\pi_{\lambda,c}$).

(i.e. refinement). In general, allowing for additional variability at the target position resulted in similar execution lengths (Tbl. 2, $\pi_{\lambda,c}$ vs. π_λ), which is not unexpected given that the length of the cylinder is quite small compared to the distances between the starting and target positions. Rather, the benefit of allowing for variation at the target position is more flexible placement of the hand on the object. Flexible hand placement has many potential advantages, for example if part of the object is unreachable due to obstruction by an obstacle.

4.3 Refined and More Successful Behavior

Tactile corrections were able to refine the policy of the demonstrated behavior, as shown through the adaptation of the regression signal in response to the tactile feedback. For this validation task, corrections were provided near the end of the task execution, to adjust the end-effector position once it was close to the cylinder (though in theory corrections may be provided at any point during the execution). Figure 6 highlights the shift in regression envelope at the target position (discussed in the previous section), with the callouts showing the envelope to narrow in dimension x_1 , to slightly shift in x_2 and to broaden in x_3 .

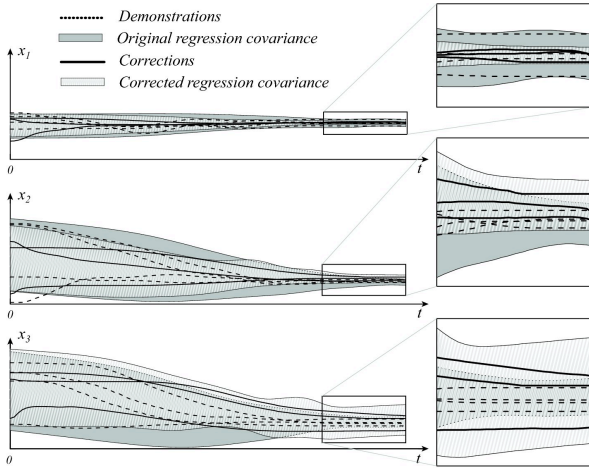


Figure 6. Shift in regression signal covariance with corrections.

Policy refinement resulted in improved execution success. Here success is measured by a score of 1 given to executions that successfully pick up the object, $\frac{1}{2}$ to those that grasp the object but not

stably enough to prevent slipping as it is picked up, and 0 to those that are unable to grasp the object (sum of 10 executions from random starting positions). One trade-off to not following the regression mean exactly, and thus to using our offset δ_λ , is that executions reach the target position less reliably (Tbl. 3, π_λ vs. π). Tactile corrections are able to buy back some of this lost reliability however, and thus improve policy performance. The corrected policy achieves the goal more reliably (Tbl. 3, $\pi_{\lambda,c}$ vs. π_λ) but with execution trajectories of similar length (Tbl. 2, $\pi_{\lambda,c}$ vs. π_λ), and thus without sacrificing the execution flexibility gained through offset δ_λ .

Table 3. Execution Success

	π	π_λ	$\pi_{\lambda,c}$
Score	10	6.5	8.5

4.4 Discussion

Here we provide discussion on two key aspects of the approach presented in this work: the preservation of demonstration variability within learned policy, and the choice of weight formulation for corrected datapoints. We follow by highlighting some promising directions for future research.

4.4.1 Reflecting Demonstration Variability in the Policy

This work employed a variant on the GMM-GMR regression formulation, that allowed for deviation from the mean trajectory of the demonstrations. The goal of such a formulation was to allow for flexibility in the resulting policy execution. Noted benefits of such flexibility included the possibility of following more direct trajectories to the target position, and online adaptability to perturbations. As a trade-off, potential detriments included reaching the target position less reliably however.

This formulation may equivalently be seen as using the differences seen between demonstrations as a template by which to infer those parts of the state space in which the task permits variability in the execution. Likeminded approaches have aimed to infer the crucial aspects of task execution by extracting what is similar between multiple demonstrations or demonstrators (e.g. [6, 10, 15]).

We highlight that acceptable variability in the task execution was effectively conveyed by the teacher through multiple modalities; namely, teleoperation and tactile corrections. Moreover, the modalities were individually better suited for different areas of the state space. In particular, to indicate generality in starting position, teleoperation was very effective. To provide generality over starting positions with tactile feedback we expect would have been quite tedious in comparison, as the tactile interface is best suited for small iterative movements. By contrast, to indicate generality along a single axis at the target position was best provided through the tactile interface, which was more responsive to precise positioning.

4.4.2 Weighting New Datapoints

We also employed this idea of demonstration variability within our weight formulation for new datapoints. In particular, in areas that exhibited little variability during teacher demonstration, the new behavior examples produced as a result of tactile corrections were considered to be very significant. By contrast, in areas that exhibited much

variability during demonstration, the presence of additional variability in the form of new corrected behavior examples was more expected, and thus considered to be less significant.

Our weight formulation relied on the assumption that the variance present within the demonstrations matched the target variance of the adapted task. Such an assumption is reasonable for policy *refinement*, as the target behavior of the adapted policy is the same as that of the demonstrated policy. Such an assumption is less appropriate for policy *reuse*, however, when the target behavior of the adapted policy *differs* from that of the demonstrated policy, and therefore so also might the variance profile. Consider for example the demonstration of a reaching task that passes through a large hole in order to grasp an object. Now consider adapting the resultant policy to achieve the same grasping objective, but by reaching through a much smaller hole. The demonstrated task will be constrained only at the target position, where it must grasp the object. By contrast, the adapted task will be constrained additionally in the middle of the execution trajectory, when it must pass through the small hole. The comparatively large covariance in the demonstration data at this phase of the execution trajectory therefore will *not* reflect the allowed variance in the adapted task.

We therefore expect the development of suitable weight functions for corrected datapoints to be an active area for future research. Many formulations are potential candidates, and their suitability depends at a higher level on what the designer wants to see come out of the learning. For example, if a separate weighting function is employed for refinement versus reuse, should the robot be notified when reuse is taking place, or statistically infer as much from differences between the distributions of demonstrated versus corrected datapoints? Another learning objective could be to infer the worth of particular datapoints, according to some utility function, and therefore not rely on the assumption that corrected datapoints are better examples (than the demonstrated datapoints) of the target task behavior.

4.4.3 Future Directions

There are many promising extensions to this work. From an algorithmic standpoint, to consider alternative paradigms for setting the weight on the influence of new data in a policy update will be at the forefront of our future investigations. Already discussed is how an alternate formulation would be more appropriate for policy reuse; another possibility could be to have a distinct weighting function for each teacher, for example, as demonstrators might differ in their ability to perform the task. Correcting within the action space is another area of interest, where for example human touch indicates changes in joint speed instead of, or in addition to, changes in pose.

From an implementation standpoint, we have found tactile corrections to be effective for the purpose of refining a demonstrated behavior, and to show their effectiveness for the purposes of policy reuse is currently under development. To validate TPC on a more sophisticated tactile sensor, that provides a richer set of feedback signals, is another direction that we are actively pursuing. An additional direction for future research is to expand the application influence of the tactile corrections, for example to correct the entire arm pose instead of just end-effector position.

5 Conclusion

We have presented *Tactile Policy Correction (TPC)* as an algorithm for the adaptation of policies through tactile feedback from a human

teacher. With tactile corrections, we aimed to improve the performance of a demonstrated behavior in response to execution experience. Multiple teaching modalities - namely, teleoperation and tactile corrections - were employed to provide examples of behavior execution, and we have highlighted the differing suitability of each for providing information about acceptable variability in the task behavior at different points during the task execution. We have provided an initial validation of policy refinement with TPC on a humanoid performing a grasp positioning task. Future work will validate TPC for policy reuse and with a more sophisticated tactile sensor.

ACKNOWLEDGEMENTS

This research was funded by EU Projects IST-2004-004370 (RobotCub), IST-04169 (felix-growing) and FET-7 grant number 231500 (RobotSkin).

REFERENCES

- [1] Pieter Abbeel and Andrew Y. Ng, 'Exploration and apprenticeship learning in reinforcement learning', in *Proceedings of ICML*, (2005).
- [2] Brenna D. Argall, *Learning Mobile Robot Motion Control from Demonstration and Corrective Feedback*, Ph.D. dissertation, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, March 2009.
- [3] P. Baerlocher and R. Boulic, 'An inverse kinematics architecture enforcing an arbitrary number of strict priority levels', *International Journal of Computer Graphics*, **20**, (2004).
- [4] Darrin C. Bentivegna, *Learning from Observation Using Primitives*, Ph.D. dissertation, College of Computing, Georgia Institute of Technology, Atlanta, GA, July 2004.
- [5] Aude Billard, Sylvain Calinon, Rudiger Dillmann, and Stefan Schaal, 'Robot programming by demonstration', in *Handbook of Robotics*, eds., B. Siciliano and O. Khatib, Springer, New York, NY, USA, (2008).
- [6] S. Calinon, F. D'halluin, D. G. Caldwell, and A. Billard, 'Handling of multiple constraints and motion alternatives in a robot programming by demonstration framework', in *Proceedings of Humanoids*, (2009).
- [7] Sylvain Calinon and Aude Billard, 'Incremental learning of gestures by imitation in a humanoid robot', in *Proceedings of HRI*, (2007).
- [8] Sonia Chernova and Manuela Veloso, 'Learning equivalent action choices from demonstration', in *Proceedings of IROS*, (2008).
- [9] Daniel H. Grollman and Odest Chadwicke Jenkins, 'Dogged learning for robots', in *Proceedings of ICRA*, (2007).
- [10] Michael Kaiser, Holger Friedrich, and Rudiger Dillmann, 'Obtaining good performance from a bad teacher', in *Programming by Demonstration vs. Learning from Examples Workshop at ML'95*, (1995).
- [11] Andrea Lockerd and Cynthia Breazeal, 'Tutelage and socially guided robot learning', in *Proceedings of IROS*, (2004).
- [12] Takashi Minato, Yuichiro Yoshikawa, Tomoyuki Noda, Shuhei Ike-moto, Hiroshi Ishiguro, and Minoru Asada, 'CB2: A child robot with biomimetic body for cognitive developmental robotics', in *Proceedings of IROS*, (2007).
- [13] Chrystopher L. Nehaniv and Kerstin Dautenhahn, 'The correspondence problem', in *Imitation in Animals and Artifacts*, eds., Kerstin Dautenhahn and Chrystopher L. Nehaniv, chapter 2, MIT Press, Cambridge, MA, USA, (2002).
- [14] Monica N. Nicolescu and Maja J. Mataric, 'Methods for robot task learning: Demonstrations, generalization and practice', in *Proceedings of AAMAS*, (2003).
- [15] P. K. Pook and D. H. Ballard, 'Recognizing teleoperated manipulations', in *Proceedings of ICRA '93*, (1993).
- [16] Joe Saunders, Chrystopher L. Nehaniv, and Kerstin Dautenhahn, 'Teaching robots by moulding behavior and scaffolding the environment', in *Proceedings of HRI*, (2006).
- [17] Martin Stolle and Christopher G. Atkeson, 'Knowledge transfer using local features', in *Proceedings of ADPRL*, (2007).
- [18] John D. Sweeney and Roderic A. Grupen, 'A model of shared grasp affordances from demonstration', in *Proceedings of Humanoids*, (2007).
- [19] Kazuyoshi Wada and Takanori Shibata, 'Social effects of robot therapy in a care house - change of social network of the residents for two months -', in *Proceedings of ICRA*, (2007).

An Autonomous Proxemic System for a Mobile Companion Robot

Mohammadreza Asghari Oskoei, Michael L. Walters, and Kerstin Dautenhahn¹

Abstract. – This paper presents an Autonomous Proxemic System (APS) for a mobile robot. It detects people in the surroundings and manipulates the robot's motions to approach them keeping an acceptable proxemic distance. The APS sensing functions include face and upper body detection, leg detection, and motion detection using camera, laser, and infra-red sensors respectively. The control functions consist of approach a human and obstacle avoidance. APS uses the sonar and laser range devices to keep an accurate proxemic distance with the human. Initial system tests indicate that the APS keeps desired proxemic distances to within an acceptable error margin.

1. Introduction

In order for a mobile service or domestic robot to be a socially acceptable and effective companion, it must exhibit appropriate socially acceptable behaviours. The study of social spaces between people is termed *Proxemics*. Within the wider research field of Human-Robot Interaction (HRI), Human-Robot Proxemics (HRP) studies how humans and robots use and manipulate distances between each other with regard to social behaviour and human perceptions. Breazeal [1] has found that humans responded socially to expressive zoomorphic robots in some very fundamental non-verbal ways, including respecting the robot's interpersonal space. Nomura *et al.* [2] found that both participants' negative attitudes and anxiety towards a small size humanoid robot had statistically significant effects on users' preferred (comfortable) robot approach distances. The main aims for our HRP research are to empower domestic or service robots to be able to:

- Detect the presence and position of people in its surroundings
- Approach, pass or avoid people as necessary, while dynamically controlling for socially acceptable HRP distances
- Take account of both the robot's and user's physical situations, and the robot's task context.

The particular HRP distance taken will also depend upon other factors, including each individual human user's preferences, the physical and social situation, and also task context [3]. Some of these other factors which affect HRP are known, but have only been roughly quantified using essentially static measurement methods, such as the HRP framework presented by Walters *et al.* [4]. It is therefore desirable to carry out more comprehensive research to see if some of the richness apparent in human proxemics interactions can also apply to other HRP interactions.

Human Proxemics

In human-human interactions, Hall [5] observed that human social spatial distance varies by the degree of familiarity between interacting humans and the number of participants. Later, Hall [6] provided a framework which categorized the main social spatial zones by interaction and situation. Hall estimated these distances visually but later researchers [7] have assigned numerical values for human-human personal space zones:

- Intimate zone $< 0.45\text{m}$
- Personal zone $\geq 0.45\text{m}$ and $< 1.2\text{m}$
- Social zone $\geq 1.2\text{m}$ and $< 3.6\text{m}$
- Public zone $\geq 3.6\text{m}$

In the field of human proxemics research, other factors which can also affect proxemic distances between interacting humans have also been proposed. For instance, Stratton *et al.* [8] suggested that uncertainty (or slight perceived threat) can affect human proxemic distances, and makes them take up slightly greater distances from the source of the perceived potential "threat". In a study, for a robot which used different voice styles, participants initially encountering the robot took significantly different comfortable approach distances [9] and it was suggested that these differences may be caused by participants' slight initial uncertainty due to perceived inconsistencies between the robot's appearance and voice styles.

Gillespie and Leffler [10] concluded that much of the observed variation in social distance between communicating humans is accounted for by the relative status of the interactants. Burgoon and Jones [11] explained many seemingly contradictory aspects of human-human proxemic behaviour by suggesting that relatively small (dynamic) manipulations of the distance between participants were a social "reward and punishment" mechanism. This theory can also explain how high status interactors can "reward" lower status interactors by moving closer, but lower status interactors can "reward" higher status interactors by keeping a greater distance.

Human-Robot (HR) Proxemics

Hüttenrauch *et al.* [12] concluded that in HRI user trials most participants kept inter-personal distances from a PeopleBotTM robot corresponding to Hall's Personal spatial zone (0.45m to 1.2m). Previously in HRI trials run using semi-autonomous robot control techniques in HRI trials, we found that children tended to approach a similar robot to similar distances [13] but for individual adults approaching the same robot, the approach distances were more ambivalent and inconclusive [14][15].

¹ Adaptive Systems Research Group, School of Computer Science, University of Hertfordshire, College Lane, Hatfield, AL10 9AB, UK, Emails: {M.Asghari-Oskoei, M.L.Walters, K.Dautenhahn} @ herts.ac.uk

We (Koay *et al.* [16]) also found that people generally allow robots to approach more closely during physical interactions (handing over an object etc.) than for verbal or no interaction conditions. Syrdal *et al.* [17] found people generally prefer more humanoid appearance robots to keep a further distance away than mechanoid appearance robots. Walters *et al.* [18] found that participants' preferences for particular robot attributes (both appearance and height) affected participants' comfortable approach distances with regard to whichever robot type they interacted with. The results from our previous HRP trials [4] are summarized in Table 1, where all distances have been compensated to satisfy a standard measurement between the human and the robot's closest body trunk parts (i.e. not including arms or manipulators). These distance measurements (as best as we can tell from the published details) are also roughly comparable to those made by Hall for his spatial zone distances and also by Stratton *et al.* [8]. Takayama and Pantofaru [19] found that other factors including robot head orientation, gender of participants, and previous experience interacting with both pets and robots also affected peoples comfortable HRP distances.

Factor	Context(s) – Approach	Base Distance = 57cm Estimated Adjustment for Factor ($\pm 0.5\text{cm}$)
Attribute or Factor of Robot		
Mechanoid Robot	All – RH	-3
	All – HR	-7
Humanoid Robot	All – RH	+3
	All – HR	-1
Verbal Communication	Verbal Interaction – RH	+3
Giving object	Physical Interaction – RH	-7
Taking object	Physical Interaction – RH	-7
Passing	No Interaction – RH	+4
Direction from:	Front – RH	+2
	Right/Left – RH	-2
Attribute or Factor of Human		
Preferred Robot Humanoid	All Private – RH	-3
Preferred Robot Mechanoid	All – RH	+3
Preferred Height Tall	All – RH	-1
Preferred Height Short	All – RH	+2
Uncertainty or perceived Inconsistency	Initial Encounter – HR	+13
Verbal Communication	Verbal Interaction – HR	+3
Giving object	Physical Interaction – HR	-7
Taking object	Physical Interaction – HR	-7
Passing	No Interaction – HR	+4

Table 1. Factors affecting HR proxemics and corresponding adjustments for Base HRP Distance (57cm) [4]

In order to confirm and extend these findings, investigate whether other factors might apply to HRP interactions, and also effectively measure and quantify any effects, it is necessary to first develop autonomous robot HRP sensing and control capabilities. Haasch *et al.* [28] have presented a mobile companion robot that employs multi modal person tracking, attention mechanism, speech recognition, and dialog manager to interact with a human, but not studied HRP in their work. This paper therefore presents a state-of-art Autonomous Proxemic System (APS) for sensing and control of HRP distance. It discriminates humans from objects, automatically measures the HRP distance, controls for a given desired HRP distance. The robot is also able to follow people around, keeping a desired HRP distance (to the best of the robot's capabilities, as it moves rather slowly compared to most people). The rest of paper is organized as follows. The next section explains the APS main components and its implementation details. Section 3 describes the experiment that evaluates the performance of APS in keeping the desired HRP distances. Finally, the last section gives a conclusion and prospect of future work.

2. Autonomous Proxemic System

The APS is designed to detect a human in the mobile robot's surroundings, and enable the robot to approach and keep a desired HRP distance in both static and dynamic states. The APS employs a range of sensors common to mobile robots consisting of a low resolution camera, passive infra-red (IR) sensor, laser, and sonar range finders. It uses computer vision techniques to detect either a face or upper body of a person within its camera range, but also applies a leg detection algorithm to laser range finder data. Meanwhile, it uses infra-red (IR) and sonar sensors to perceive and track human motions and obstacles, respectively. The rest of this section provides more details of how these sensors and associated algorithms are implemented in the APS.

Face Detection

The APS uses face detection to detect and localize people in the focus range of the camera. The main aim of this face detection is to determine whether or not there is actually a human face in the current captured camera frame, and if so return the location. However, face detection is challenging due to variability in scale, location, orientation (up-right, rotated), and pose (frontal, profile). Facial expression, occlusion, and lighting conditions also change the overall appearance of the faces [20].

In our application, we take advantage of an object detector which uses Haar-Like features of an image [21]. This is a refined version of the widely known algorithm created by Viola & Jones [22]. This algorithm is already implemented and trained for face and upper body detection in *OpenCV* [23], the open-source computer vision library adopted for use in the APS. According to a comparative survey in [20], the chosen solution shows a good balance between performance and computational speed. It is also proven that this method is colour independent (i.e., adapt for different skins) and robust to varying light conditions.

Briefly, the chosen face detection algorithm deploys Haar-like features that consist of two or three jointed rectangular regions (Figure 1). The value of a Haar-like feature is the difference between sums of grey level values of pixels within the two rectangular regions.

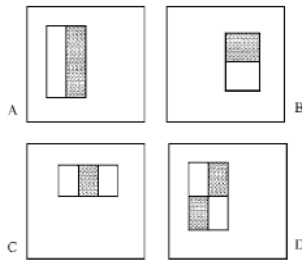


Figure 1. Haar-like features: two or three jointed rectangle regions [22]



Figure 2. Haar-like features are extracted from sub-windows for face detection [22]

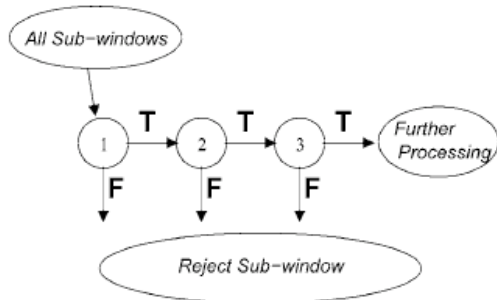


Figure 3. Cascade of simple classifiers applied to the Haar-like features of the sub-windows [22]

Compared with raw pixel values, Haar-like features can reduce in-class, and increase out-class variability, thus making more distinguishable data and easier classification. The Haar-like features are computed from sub-windows of an image (Figure 2). Given an image resolution of 320x240, sub-window resolution of 24x24, and 15 frames per second, the total number of sub-windows with one Haar-Like feature is about 1 million per second which has a relatively large computation cost. To optimize this computation, a cascade of pre-trained simple classifiers (i.e. AdaBoost [21]) with a threshold structure is applied to the features computed from sub-windows. The first classifier eliminates a large number of negative sub-windows and passes almost all positive sub-windows (high false positive rate) with very little processing effort. Subsequent layers eliminate additional negative sub-windows (passed by the first classifier) but which require more computation. After several stages of processing, the number of negative sub-windows has been reduced greatly (Figure 3). Finally, the remaining relatively few sub-windows may contain a face passed as the output of the algorithm.

Bellotto *et al.* [24] presented an adapting regulation for parameters of the face detection method to improve the fast tracking performance in real-time applications. It starts with an image at normal size (320x240) and once it has detected a face (or faces), selects the nearest one, and then scans just a sub-image containing the selected face. Meanwhile, it reduces the sub-windows size into 80% of the selected face size. This significantly increases the detection speed (~4 times) and keeps track of one face as long as it can be detected.

The proposed algorithm [23] is capable to be extended to distinguish different visual patterns. Then, as mentioned, we have extended the visual object detection to detect upper body rather than the exclusive face detection. Moreover, we deployed the face detection classifier with different profiles. By this means, the robot can perceive people even when it is behind or to the side of the person. More details about the implementation are discussed in the next sub-section. To supplement the face detection system, additional sensing tools are required to detect humans in the surrounding area. This is because the applied face detection is limited in both performance (false positives, negatives and lost targets) and the area of scanning coverage.

Leg Detection

Leg detection is a pattern recognition terminology that can discriminate and localize people legs using laser readings [29]. The leg detection system processes the range data collected by laser, extracts the edges produced by the objects, and localizes the patterns of edges that match with the human leg patterns'. Although, the detected patterns are not guaranteed to belong to human legs, they provide potential directions to explore to confirm whether people are in the environment. In APS, leg detection is employed to provide the turning direction for the robot when no one was detected by the visual detection object.

The laser sensor provides range data from 180° covering the front and sides of the robot, at a height of about forty centimetres from the floor, and with half degree resolution. The scanning area is semicircular with a radius of 8 meters. The laser range data, according to the manufacturer's specifications, are very accurate with errors of a few millimetres. Figure 4 depicts a snapshot of the range data in the presence of a person in the scanning area.

Belletto *et al.* [24] [25] presented a novel detection algorithm to find human legs by using laser scans. It is designed to work either in large empty environments or small cluttered rooms, and is able to distinguish among different leg postures, thus improving the discrimination of false positives. The leg detection algorithm extracts the necessary features (edges produced by the objects) from a single laser scan, and identifies typical patterns (relative to particular leg postures) that, in most of the cases, are distinguishable from the other objects in the environment. The desired leg patterns, shown schematically in Figure 5, correspond to three typical situations: two legs-apart (LA), forward straddle (FS), and two legs-together or SL. The first pattern is usually very common when a person is standing in front of the robot. The second is most likely to happen when the person is walking. The last pattern covers most of the remaining postures. However, it can also be generated by other objects in the environment, giving rise to false positive detections [25].

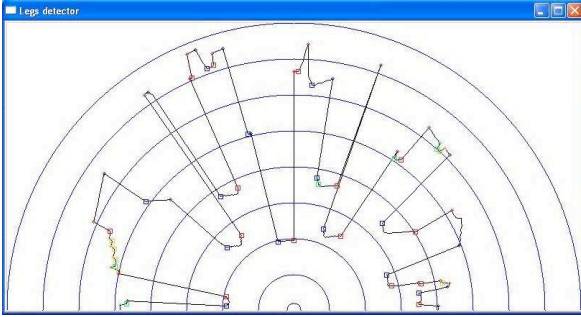


Figure 4. Range data collected from the laser sensor with marked edges

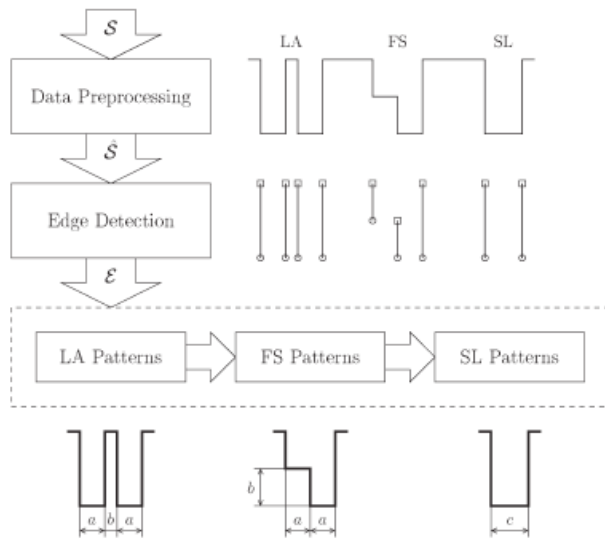


Figure 5. Leg patterns and the Leg Detector's schematic diagram [25]

As shown in the schematic presentation of Figure 5, the algorithm is divided into three main parts: data pre-processing, detection of vertical edges, and finally extraction of leg patterns. The laser range data are pre-processed by applying a local minimization operator to remove possible spikes due to reflections on sloped surfaces, and a local maximization operator to discard thin objects such as table legs. Suppose the angular step between two consecutive laser scans is constant, and the range data after pre-processing is stored in an array $S = [r_1 \dots r_i \dots r_M]$, where r_i is the range measured on the direction θ_i , and M is the total number of readings. If we represent S on a Cartesian graph, we can identify a sequence of vertical edges defined as follows. The doublet $\{r_i, r_{i+1}\}$ can be considered an almost vertical edge if the distance $|r_{i+1} - r_i|$ is greater than a given threshold. Moreover, we can distinguish a left edge, when $r_i > r_{i+1}$ from a right edge, when $r_i < r_{i+1}$, and refer to them as L_i and R_i , respectively.

The resulting vertical edges are initially queued into a list $E = \{e_1 \dots e_n \dots\}$, where each element e_n can be either an L or R edge. If they are very close and almost aligned, adjacent edges of

the same type are connected to form a longer one. After that, from the updated list of connected edges, we extract all the subsets that might belong to one of the three leg patterns described before. The order of patterns that we look for is as follows.

- The LA pattern is a quadruplet $\{L, R, L, R\}$.
- The FS pattern is a triplet $\{L, L, R\}$ or $\{L, R, R\}$.
- The SL pattern is a doublet $\{L, R\}$.

Every edge is removed from E as soon as it contributes to form one of the aforementioned sequences. Therefore, all the LA patterns, which are normally the most reliable, are extracted first, while the SL patterns, which are the easiest to misinterpret, are left at the end. During the search for the patterns we consider some constraints and spatial relations between edges, including maximum normal distance between legs and limits on their size. With reference to Figure 5, some dimensional constraints are fixed for the measures a , b , and c , which are, respectively, the leg's width, the maximum step length, and the width of two legs together. These are used by the algorithm's procedures to recognize LA, FS, and SL patterns. Finally, the distance and direction of the detected legs are calculated from the midpoint of each pattern [25].

Motion Detection

Many objects normally emit IR radiation, invisible to human eyes that can be detected by electronic devices designed for such a purpose. The APS motion detector is designed to perceive a human as the robot is moving. It is based on a passive IR sensor, which is an electronic device that measures IR light radiating from objects in its field of view. It is a passive sensor, which means that it does not emit an IR beam but merely passively accepts incoming IR radiations. Intensity of the emitted radiation is proportional to the objects' temperature [26], and apparent motion is detected when an IR source with one temperature, such as a human, passes in front of an IR source with another temperature (e.g. a wall etc.). Furthermore, a motion can also be realized when an IR sensor moves relative to an IR source with one temperature, such as a human, standing in front of an IR source with another temperature. This feature can be adopted to detect the positional change between a human and a mobile robot carrying the IR sensor.

In our application, we set up an IR sensor connected to an analogue to digital converter (ADC, Phidget Interface Board) on top of the mobile robot to perceive the presence of a human in its surroundings. The sensor was fixed at a height approximately 1.2 meters from the floor to read IR radiations emitted from the human body (i.e. trunk) and avoid typical non-human IR sources (e.g. heating radiators) in a room. The sampling frequency was 5Hz, and the recorded data was passed through a low-pass filter to remove noises produced by the vibration. The gradient of the filtered data represents a change in successive IR radiation measurements and indicates a motion when it exceeds a certain threshold. The threshold was chosen after preliminary tests, and adjusted to detect the motions produced by a human within a range of 5 metres.

According to the desired scenarios in our application, we deliberately ignored any fast or weak motions from the passive IR sensor by adjusting the sampling rate, low-pass filter and threshold parameters. After tuning, the robot detected the

presence of a static human when the robot passes by, or when a human moves in front of the robot.

System Implementation

The APS is a real-time controller developed for the Pioneer PeopleBot™ robot. It has a modular design and employs the above-mentioned detectors as modules running in multi-thread mode. The APS is implemented in C++ using ARIA and OpenCV libraries. It runs on a Dual Core PC (with Windows XP) connected directly to the robot via the USB serial port adaptor.

The APS includes both control and sensing functions. The controlling commands are made up of 'Turning Left/Right' and 'Moving Forward/Backward', both implemented as Actions in ARIA. The Turning command has a priority over Moving; this means the robot first turns toward the person and then starts approaching. The Moving command checks regularly the range devices (i.e. Sonar and Laser) to keep the HRP distance and avoid obstacles. The cycle time of the robot controller, at which the sensors and commands are regularly updated, was set to 200ms. This is a sensible rate for a companion robot in normal daily activities.

APS sensing functions include face and upper body detection, leg detection, and motion detection using the camera, laser, and Infra-Red (IR) sensor, respectively. Moreover, the obstacle avoidance Action also uses the sonar and laser range data. APS gives a higher priority to the camera-based functions over the laser [30]. This means APS manipulates the robot using vision-based information, as long as it detects a face or an upper body in view of the camera. When a face or an upper body is detected, APS uses its relative horizontal location in the image to adjust the robot's bearing angle towards the human. In the absence of a detected human by the camera system during a limited time, APS then uses the information provided by the leg detector to manipulate the robot. Motion detection, due to its particular characteristics, just stops the robot turning (i.e., interrupt the Turning command) when it spots a human in its range.

To make a sensible integration of the detectors, we require considering their features along with the characteristics of the controller. The laser-based leg detector is very accurate and in most of the cases is much more reliable than face detection [25]. Moreover, the computational time needed by the leg detector is much less than that one required by the face detection module. Considering the modules running asynchronously in real-time operation, the face and leg detection processes take about 500 and 250msec, respectively. The range covered by the laser device is much wider than the camera view. While the laser covers a semicircular area with a radius of 8 metres, the camera view is limited to approximately 40°. As mentioned, the face detection is featured to discriminate faces with different profiles as well as the upper body. However, even if the camera is fixed at about 1.4m from the floor (which is about a normal person's face height), there are cases when a face or upper body cannot be detected because a person is too tall, too short, or very close to the robot.

3. Experiments and Results

According to Section 1, a domestic robot should be able to keep an acceptable HR proxemic distance during interaction

with a human. The APS is supposed to manipulate the robot in such a way that it keeps reliably the desired HRP distances corresponding to different settings. We conducted an experiment, to evaluate the performance in HR proxemic control produced by APS.

Desired distance in APS									
	450	550	650	750	850	950	1050	1150	1250
Recorded distances for subject 1									
1	385	500	595	725	715	885	1015	1055	1055
2	395	495	645	730	825	885	1005	1005	1125
3	395	485	635	735	735	825	960	1065	1115
4	405	475	620	705	765	845	1015	1015	1135
5	385	505	595	675	755	865	985	1055	1145
Recorded distances for subject 2									
1	365	540	560	650	760	820	895	1005	1085
2	375	450	570	650	760	830	885	1005	1105
3	355	480	570	660	750	825	915	995	1095
4	345	480	565	655	755	835	895	985	1105
5	335	480	575	665	765	825	925	1005	1115

Table 2. Recorded HRP distances corresponding to different settings (in millimeters)

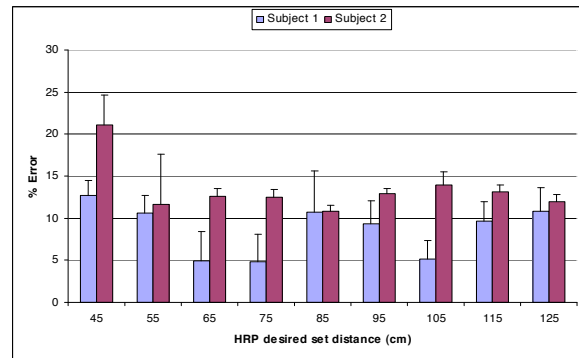


Figure 6. Error (%) of raw HRP distances

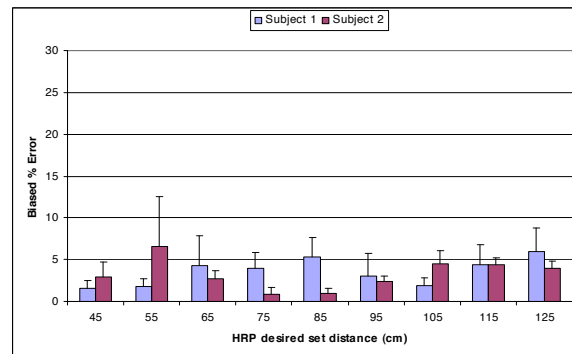


Figure 7. Error (%) of HRP distances after applying bias values (6cm for Subject 1 and 10cm for Subject 2)

To have a realistic evaluation, the experiment with two participants was conducted in a real living room with usual furniture in the University of Hertfordshire 'Robot House'. The 'Robot house' is a house near the University, based in a domestic area, which appears to be like any typical UK house, but has been adapted so that HRI experiments and user trials can be

performed in an ecologically valid, real home environment, rather than a laboratory or simulated home surroundings. In the experiment, we measured directly using a tape measure the actual distance between static human and robot corresponding to different desired HRP values set in APS. The experiment was designed to evaluate the reliability and repeatability of the APS performance; Hence, we repeated five trials for each setting, and also recorded the error. The error refers to the difference between desired and actual HRP distances. Some error is inevitable, since the APS records the nearest distances read by the sonar and laser at a fixed height (i.e. 0.3 and 1.2m for sonar and 0.4m for laser) from the floor as HRP, while we measured the actual distance at a height that includes the closest point between human and robot's trunks. However, it was expected to record errors with constant bias and low variance. The bias mainly depends on particular subjects' body shape.

The experiment examined HRP settings from 45 to 125 centimetres, and repeated trials for each setting five times. Table 2 demonstrates the recorded distances for two participants and nine settings. Figures 6 and 7 illustrate the mean and standard deviation for error of HRP distances for each subject and setting. The former is for the raw data and the later depicts the error rate after subtracting the bias value to reduce the constant part of the error.

The bias is a constant value that exists in the differences between measured values by the human and robot. It is primarily caused by the difference in the height of the point of measurement. The bias value therefore depends mostly on individual humans' body shapes and is worked out for each subject. Considering the errors with normal distribution, it can be calculated using the 'three-sigma' rule in Statistics [27]. The bias is worked out by subtracting the one-third of the standard deviation (STD) of the recorded errors from the absolute value of their mean. It was found to be 60 and 100 millimetres for Subject 1 and 2, respectively.

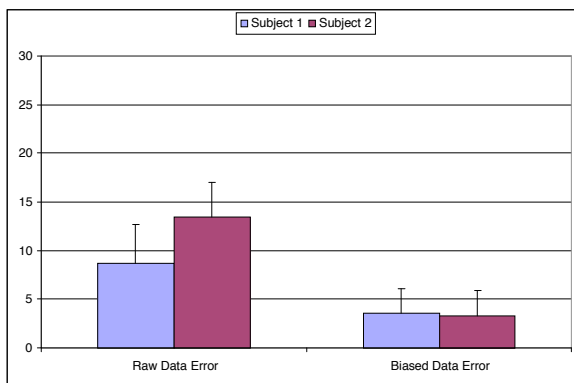


Figure 8. Error (%) decreases significantly by applying the bias values to raw HRP distances

Figure 8 shows that applying the individual subjects' biases to the raw HRP distances from the APS, increased significantly the accuracy of measurement. The adjusted (biased) mean HRP error is about $\pm 1.5\%$ with repeatability of $\pm 1\%$. This implies an approximate error of $\pm 0.75\text{cm}$ in HRP distance measurements close to the 57cm base distance [4] from the Proxemic Framework from Table 1. This is acceptable for our future work

which will focus on HRP interactions within the near Personal Zone distances (40 – 100cm). We hope to refine the measurement accuracy in the light of more data from a wider range and number of participants in future trials. We also intend to improve the APS in future work by incorporating a learning mechanism that can learn users' proxemic preferences and individual HRP parameters during run-time.

4. Conclusion

The social behaviour of a mobile robot can make it socially acceptable and effective as a companion. In HRI, the study of how human and robot use and manipulate distances between each other with regard to social behaviour and perceptions is called human-robot proxemics (HRP). An aim of HRP research is that a social service robot should be able to detect presence of people in its surrounding and approach them to an acceptable proxemic distance. The particular HRP distance depends on individual humans' preferences, the robot's task and services, social and physical situation, and possibly other factors. In this paper we proposed an autonomous proxemics system for a mobile robot. Experimental findings indicate that the proposed system works reliably and keeps a desired HRP distance with a total error variance about $\pm 1.5\%$. We intend to improve the APS in future work by refining the HR measurements and incorporating a learning mechanism that automatically can adapt to individual users' HRP preferences and parameters.

Acknowledgment

The work described in this paper was conducted within the EU Integrated Projects LIREC (Living with Robots and intERactive Companions) and was funded by the European Commission under contract numbers FP7-215554.

References

- [1] C. Breazeal, B. Scassellati, 'Infant-like Social Interactions Between a Robot and a Human Caretaker', *Adaptive Behavior*, 8(1), 49-74, (2000).
- [2] T. Nomura, T. Shintani, K. Fujii & K. Hokabe, 'Experimental Investigations of Relationships Between Anxiety, Negative Attitudes, and Allowable Distance of Robots (HCI 07)', *Proceedings of the Second International Conference on Human-Computer Interaction (HCI 07)*, Beijing, P.R. China, 13-18, (2007).
- [3] Topp, E. A., and Christensen, H. I. Detecting structural ambiguities and transitions during a guided tour. In *2008 IEEE Intl. Conf. on Robotics and Automation* (Pasadena, CA, May 2008), pp. 2564-2571.
- [4] M. L. Walters, K. Dautenhahn, R. te Boekhorst, K. L. Koay, D. S. Syrdal, and C. L. Nehaniv, 'An Empirical Framework for Human-Robot Proxemics', *Journal of Autonomous Robots*, 24(2), 159-178, (2008).
- [5] E. T. Hall, *The Hidden Dimension*, Doubleday, NY, (1966).
- [6] E. T. Hall, 'Proxemics', *Current Anthropology*, 9(2-3), 83-108, (1968).
- [7] D. Lambert, *Body Language*, Harper Collins, London, (2004).
- [8] L. O. Stratton, D. J. Tekippe, G. L. Flick, 'Personal Space and Self Concept', *Sociometry*, 36, 424-429, (1973).

- [9] M. L. Walters, K. L. Koay, K. Dautenhahn, R. te Boekhorst & D. S. Syrdal, 'Human Approach Distances to a Mechanical-Looking Robot with Different Robot Voice Styles', Proceedings of the 17th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN 2008), Munich, Germany, 707-712, (2008).
- [10] D. L. Gillespie, A. Leffler, 'Theories of Non-Verbal Behavior: A Critical Review of Proxemics Research', Sociological Theory, 1, 120-154, (1983).
- [11] J. K. Burgoon, S. B. & Jones, 'Toward a Theory of Personal Space Expectations and Their Violations', Human Communication Research, 2(2), 131-146, (1976).
- [12] H. Hüttenrauch, K. Severinson Eklundh, A. Green, E. A. Topp, 'Investigating Spatial Relationships in Human-Robot Interaction', Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 06), Beijing, China, 5052-5059, (2006).
- [13] M. L. Walters, K. Dautenhahn, K. L. Koay, C. Kaouri, R. te Boekhorst, C. L. Nehaniv, I. Werry, D. Lee, 'Close Encounters: Spatial Distances Between People and a Robot of Mechanistic Appearance', Proceedings of IEEE-RAS International Conference on Humanoid Robots (Humanoids2005), Tsukuba, Japan, 450-455, (2005).
- [14] M. L. Walters, K. Dautenhahn, K. L. Koay, C. Kaouri, S. N. Woods, C. L. Nehaniv, R. te Boekhorst, D. Lee, I. Werry, 'The Influence of Subjects' Personality Traits on Predicting Comfortable Human-Robot Approach Distances', Proceedings of Cog Sci 2005 Workshop: Toward Social Mechanisms of Android Science, Stresa, Italy, 29-37, (2005).
- [15] M. L. Walters, K. Dautenhahn, S. N. Woods, K. L. Koay, R. te Boekhorst, D. Lee, 'Exploratory Studies on Social Spaces between Humans and a Mechanical-looking Robot', Journal of Connection Science, Special Issue on Android Science, 18, 429-442, (2006).
- [16] Koay, K. L., Syrdal, D. S., M. L. Walters, K. Dautenhahn, 'Living with Robots: Investigating the Habituation Effect in Participants' Preferences During a Longitudinal Human-Robot Interaction Study', Proceedings of the 16th IEEE International Workshop on Robot and Human Interactive Communication (ROMAN 2007), South Korea, 564-569, (2007).
- [17] D. S. Syrdal, K. L. Koay, M. L. Walters, K. Dautenhahn, 'A Personalised Robot Companion? - The Role of Individual Differences on Spatial Preferences in HRI Scenarios', Proceedings of the 16th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN 2007), Korea, 26-29, (2007).
- [18] M. L. Walters, D. S. Syrdal, K. Dautenhahn, R. te Boekhorst, K. L. Koay, 'Avoiding the Uncanny Valley – Robot Appearance, Personality and Consistency of Behavior in an Attention-Seeking Home Scenario for a Robot Companion', Journal of Autonomous Robots, 24(2), 159-178, (2008).
- [19] Takayama, Leila; Pantofaru, Caroline, "Influences on proxemic behaviors in human-robot interaction," Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on , vol., no., pp.5495-5502, 10-15 Oct. 2009
- [20] M. Yang, D. J. Kriegman, and N. Ahuja, Detecting Faces in Images: A Survey, IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 24, No. 1, January 2002, pp. 38-58
- [21] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in Proc. the IEEE Int. Conf. on Image Processing 2002, vol. 1, New York, USA, 2002, pp. 900-903.
- [22] P. Viola and M. J. Jones, "Rapid object detection using a boosted cascade of simple features." in CVPR (1), Kauai, HI, USA, 2001, pp. 511-518.
- [23] OpenCV, "Open Computer Vision Library." [Online]. Available: <http://sourceforge.net/projects/opencvlibrary>
- [24] N. Bellotto and H. Hu, Multisensor Integration for Human-Robot Interaction, The IEEE Journal of Intelligent Cybernetic Systems, Vol. 1, 2005
- [25] N. Bellotto and H. Hu, Multisensor-Based Human Detection and Tracking for Mobile Service Robots, IEEE Transactions on Systems, Man, and Cybernetics: Part B, Vol. 39, No. 1, pp. 167-181, 2009.
- [26] Wikipedia http://en.wikipedia.org/wiki/Passive_infrared_sensor
- [27] Wikipedia http://en.wikipedia.org/wiki/68-95-99.7_rule
- [28] A. Haasch, S. Hohenner, S. Huwel, M. Kleinhagenbrock, S. Lang, I. Toptsis, G. A. Fink, J. Fritsch, B. Wrede, and G. Sagerer, "BIRON – The Bielefeld Robot Companion", Proc. Int. Workshop on Advances in Service Robotics, pp. 27-32, 2004
- [29] J. Xavier, M. Pacheco, D. Castro, A. Ruano, and U. Nunes, "Fast line, arc/circle and leg detection from laser scan data in a player driver", Proceedings of IEEE International Conference on Robotics and Automation (ICRA) , 2005
- [30] Z. Zivkovic, and B. Krose, "People Detection Using Multiple Sensors on a Mobile Robot", Unifying Perspectives in Computational and Robot Vision, pp. 25, Springer Verlag, 2008.

Towards a Skin-Like Embedded Tactile Sensor for Robotic Hands

Andrea Biggio and Perla Maiolino and Giorgio Cannata¹

Abstract.

A novel embedded tactile sensor designed to be integrated in future robotics hands and the related production process are presented. The sensor is based on a multilayer rubber structure bound to a rigid thermoplastic support frame which encloses the embedded sensor electronics. In the project design, each sensor features up to 12 pressure sensitive tactels distributed on a phalange-like curved surface. Tactile data are read by a CDC (Capacitance-to-Digital Converter) and forwarded to a microcontroller via I2C bus. Each microcontroller module handles up to 4 sensor modules (one finger) and provides the data over CANbus.

The construction process has been designed to be as cheap and simple as possible. It is divided in four macro-steps, which relies on different technologies: elastomers for dielectric and conductive layers, thermoplastic materials and thermoforming for rigid support frame, flexible PCB for the embedded electronics and adhesive for the layers binding.

1 INTRODUCTION

Companion or Social robots represent a new class of robots specifically designed to interact and cooperate with humans not only to perform manufacturing tasks, but also for service applications even in domestic environments (e.g. house keeping, healthcare support, entertainment, etc.).

A fundamental requirement for a this type of robot is the capability of safely interact with the environment including humans. For this purpose a major feature is the capability of measuring touch information, in order to implement control strategies, for reacting to external contact stimuli.

Although in recent years, robots have become increasingly part of human life, to allow a greater diffusion of social robots (i.e. a large scale production), it is necessary to lower building costs of the robots and to make the production processes simple and repeatable.

Tactile sensing for robotic hands has been deeply studied and many types of transducers have been adopted, as widely shown in literature [3], [8].

The development of tactile sensors is a very hard problem involving several engineering issues and often also special processing equipment and this has limited the development of embedded solutions. Furthermore most of the solutions proposed are strongly dependent on the mechanical robot design and often specialized for a particular platform [1], [4], [5], [6], [7], [10].

In this paper a novel tactile skin-like sensor for fingertip is proposed. In particular we have focused on the study of production pro-

cesses, materials and component selection issues, in order to define design guidelines for a low-cost embedded tactile sensory system. On the basis of these guidelines, we have developed some proof-of-concept prototypes.

The starting point has been the work of Maggiali et al. [9] and we have taken in account the following issues:

- Spatial Resolution: the tactile elements (tactels) number must allow the reconstruction of the contact profile all over the surface.
- Reliability: the sensor must give reliable measures and avoid false contacts.
- Mechanical toughness: as the sensor has to physically interact with the environment, it must be tough enough to not be easily damaged by these interactions.
- Miniaturization: as the sensor has to be deployed on a human-size like robotic hand, it must be self contained and fully embedded in a phalange-like volume.
- Production Costs: the sensor must be cheap enough to be disposable, therefore the related production process and the material used have to be optimized to have the most high-quality sensor at the lowest cost possible.

2 DESIGN CONCEPT

Capacitive sensor technology has been adopted thanks to the availability of miniature CDC (Capacitance-to-Digital Converter) used in last generation consumer electronics.

In this work a matrix of capacitors is integrated in the fingertip tactile sensor structure, and the contact pressure is detected thanks to the dielectric deformation, which results in a capacitance variation (Fig. 1). The tactile sensor elements (tactels) are, in fact, capacitive transducers distributed all over the fingertip surface.

The tactile sensor is made by 5 principal elements (Fig. 2):

- Embedded Electronics: a flexible PCB (polyimide + copper) is used to provide the tactels shape and to bring the signal to the CDC (section 3).
- Thermoplastic Support Frame: which gives the shape and the structural support to the sensor. The flexible PCB is bound to its internal surface (section 4).
- Dielectric Layers: made of a soft silicone rubber which deforms when it is subjected to the contact forces (section 5).
- Conductive Ground Layer: made of conductive silicone rubber (section 6).
- Protective Layer: made of a thin substrate of silicone rubber. It is designed both to protect the conductive layer from mechanical stress and to improve the surface grip (section 7).

¹ University of Genova, Department of Communications, Computer and Systems Science email: biggio@ieee.org, perla.maiolino@unige.it, giorgio.cannata@unige.it

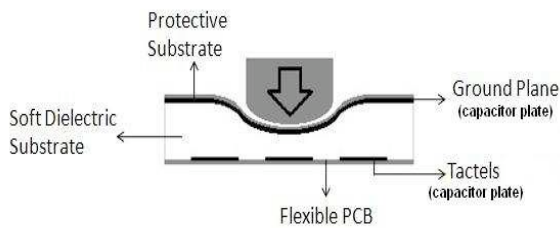


Figure 1. Layers Deformation Concept

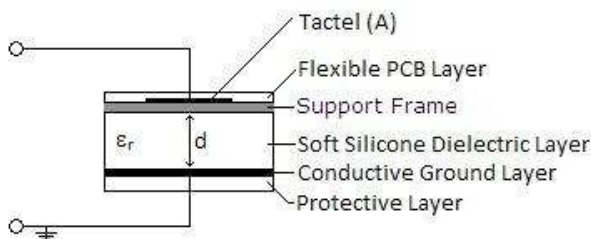


Figure 2. Tactile sensor concept

3 EMBEDDED ELECTRONICS

The embedded electronics is integrated on a flexible copper-on-polyimide (Polyimide film, Rogers Corporations) printed circuit board (PCB), capable to bend and adapt to the small and curved surfaces of the fingers. The flexible PCB can be divided in two parts:

- Tactels PCB: provides tactels shape and distribution over the surface.
- Circuitry PCB: which hosts the Capacitance-to-Digital Converter (CDC).

The fingertip sensor unit hosts 10 round tactels with a 3mm diameter and covers a human-size thumb surface, having about 2.5 tactels per square centimeter resolution (Fig. 3).

The selected CDC Analog Devices (AD7147-1, produced by Analog Devices) can read up to 12 tactels capacitance with a Femto-Farad ($10^{-15}F$) resolution up to 16-bits. The whole acquisition cycle is automatically completed in 20 ms.

4 THERMOPLASTIC SUPPORT FRAME

The Thermoplastic support frame gives the shape and the structural support to the sensor.

Since the flexible PCB must be supported by the support frame (flexible PCB is bound on its internal surface), various materials and manufacturing process has been investigated in order to obtain a tactile sensor structure with embedded electronic in easy and repeatable way. For this purpose the support frame consists of a mechanically thermoformed thermoplastic material sheet.

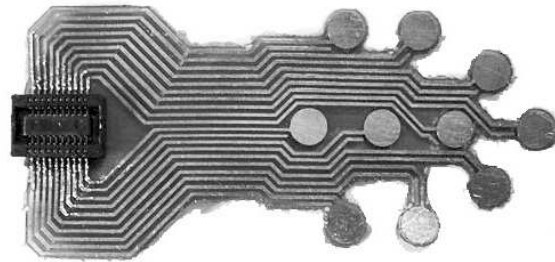


Figure 3. Tactels Pattern Flexible PCB.

Thermoforming is one of the oldest and most common methods of processing plastic materials. It is a very versatile process used to manufacture a wide range of products.

For our purposes mechanical thermoforming has been selected and performed using coupled moulds (positive and negative) which are pressed together forcing a preheated plastic sheet to the desired shape. The main advantages of this method are:

- No special thermoforming equipment needed to generate positive/negative air pressure.
- Formed material thickness can be a design parameter.
- Both the internal and external surfaces can be formed with the desired details.

The main critic part of this process is related to the mechanical stress which can damage the plastic sheet or the conductive layer attached to it.

The equipment used for thermoforming experiments and prototype frame support production was composed of:

- Electric oven with temperature regulation.
- Coupled PTFE molds.

The mechanical thermoforming process is made of 3 steps:

1. Mold Assembly preparation: the plastic sheet is blocked within aluminum clamps and placed on the female mold, using the guides to have the best alignment. Then the male mold is placed on the sheet and everything is blocked on the support aluminum bar (Fig. 4).
2. Heating and Forming: the Mold Assembly is introduced into the oven at the desired temperature (within the material thermoforming range). Material glass transition temperature is reached when its edges start to bend. When the central part of the plastic sheet starts flexing, the sample is ready to be formed. The male mold is

- pressed (manually or using a weight) into the female mold and the mold assembly is extracted from the oven.
3. Cooling: mold assembly is placed into a clamp which continues to press male and female molds together. When the mold is cold, it is possible to disassembly it and extract the formed plastic sheet (Fig.5).

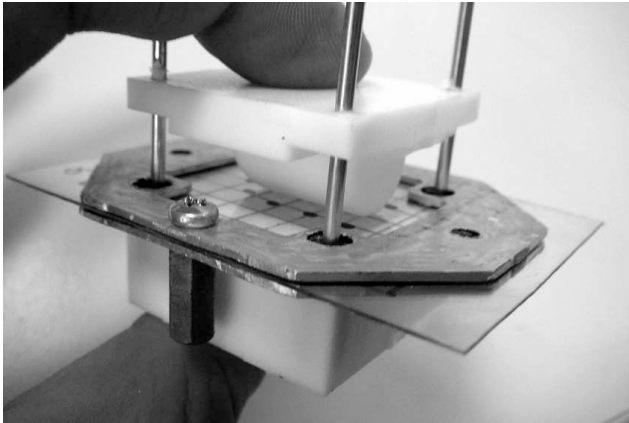


Figure 4. Mold assembly.

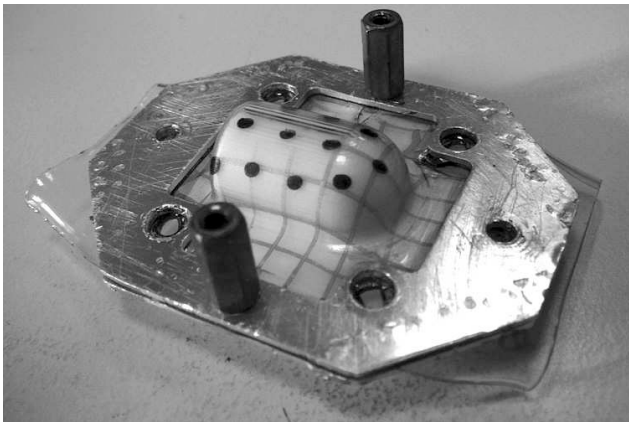


Figure 5. Thermoformed fingertip chassis.

We investigated on various type of thermoplastic materials and in particular: ABS (Acrylonitrile-Butadiene-Styrene), PETG (Polyethylene terephthalate glicole), PETGAG (3-layer structure of PETG-PETA-PETG), Acetate, White Styrene and Polypropylene.

The first issue, that has influenced the material choice, was related to the thermoforming temperature range, since the flexible PCB, and then the electronic, was bound to the thermoplastic sheet before the thermoforming process. The second issue was related to the commercial available thickness of the thermoplastic sheets, since thermoforming support frame is part of the dielectric layer of the equivalent

capacitor, and it is not thinned by the forces applied on the sensor surface.

On basis of this analysis we have decided for PETGAG that has a large thermoforming range: between $100^{\circ}C$ and $160^{\circ}C$, and it has been available in sheets with thickness of 0.150 mm, which was the smallest provided by suppliers. Furthermore, during the material investigation which led to the selection of PETGAG, we tried to empirically assess the locations predictability of the tactels and the sheet deformation repeatability. The former aspect refers to the needing to know where each tactel will be located after the thermoforming process, while the latter refers to the needing of produce small series of identical sensor units.

Experiments were done by manually drawing a 5x5 mm grid and the planar tactels pattern over the plastic sheet and then visually inspecting the formed part.

These experiments were conducted for all the materials investigated (ABS, PETG, PETGAG, Acetate, White Styrene and Polypropylene) with different results. In particular PETGAG showed very good deformation repeatability.

5 DIELECTRIC LAYER

The main part of the dielectric layer (dielectric layer is composed of silicone rubber substrate plus thermoformed frame support) is made of commercial soft silicone rubber platinum-catalyzed RTV (Room Temperature Vulcanization), with shore hardness A 00-30 and available as a two components (A, B) product (Ecoflex Series 00-30, Smooth-On, U.S.A).

To produce the fingertip-unit prototype we used the thermoformed frame support as male mould to have the best possible coupling, and a calibrated negative mould to obtain an even 2 mm layer. The material was processed according to the manufacturer specifications (curing agent B was added to the component A at the ratio 1:1 by volume), then the Ecoflex compound has been placed inside the moulds and cured at room temperature for 24 hours.

5.1 Improvement of Dielectric constant of Silicon Rubber

The thickness of dielectric layer is a trade-off between the deformation needed to have detectable capacitance variations (more the deformation, more the capacitance variation) and the capacitance needed to work in the CDC linear zone (20pF).

According to the capacitor law:

$$C = \epsilon_0 \cdot \epsilon_r \cdot \frac{A}{d} \quad (1)$$

and taking into account that, due to design requirements, pads area is fixed ($A = 6\pi \times 10^{-6}mm$), we performed various experiments in order to increase the dielectric permittivity of the silicone rubber matrix in order to get higher thickness (more deformation) and capacitance value under 20 pF (linear zone of CDC).

The performed experiments has been made on the basis of Carpi et al. [2] work. In particular a ceramic filler with high dielectric permittivity ($\epsilon_r \approx 110$), 99%-pure rutile-type titanium dioxide TiO_2 (Titanio biossido - code 488257, Carlo Erba, Italy) has been dispersed into the silicone rubber matrix Ecoflex 00-30 in order to increase its dielectric permittivity ($\epsilon_r \approx 3$).

Samples having 10%, 20%, 30% and 40% in weight dispersion ratio were separately fabricated. We got the same results as reported in [2]: an increase of 30% at low frequency in the composite resulting

from a dispersion of 30% in weight TiO_2 , and worst mechanical properties in composite samples with TiO_2 dispersion ratio of 40% in weight; at our work frequencies (250 KHz) the increase of silicone rubber dielectric permittivity is under 30% and also we noticed a loss of elasticity of the compound.

On basis of these considerations we discarded this ceramic filler and we performed the same experiments with another ceramic filler powder, 99% strontium titanate $SrTiO_3$ (Strontium titanate - code 396141-500G, sigma-aldrich).

With this filler we got better results, but not enough satisfying for our application. Others fillers and methods to increase dielectric permittivity of elastomers are under analysis, but for this prototype of fingertip tactile sensor we decided to use pure silicone rubber substrate as dielectric layer.

6 CONDUCTIVE GROUND LAYER

The conductive ground layer acts as common plate for all the tactels. We investigated on several conductive silicone rubbers, inks and adhesives in order to obtain a perfect binding between the dielectric layer and the conductive one.

On basis of this investigation we chosen a two components (A, B) black conductive silicone rubber (Silopren 2345, Momentive, U.S.A.).

The material was processed as follow: curing agent B was added to the component A at ratio 1:1 by volume, then in order to obtain a thin layer, the compound was deposited on the base rubber by dip-coating; Finally the layer was vulcanized in oven at 120°C for 2 hours.

Momentive Silopren 2345 has electric surface resistance of 370 ω for 50 μ m layer thickness. The surface resistance limits the sensor bandwidth, but in our application the sensor bandwidth is wide enough to let it work properly.

7 PROTECTIVE LAYER

The protective layer has the function of to separate and to insulate the ground plane from the touched objects, protecting it from wears; furthermore it provides good grip to the surfaces.

The protective layer has been made by a thin layer of Smooth-On Ecoflex Series 00-30 silicone rubber deposited by dip-coating on the tactile sensor surface.

8 LAYERS BINDING

In order to obtain a one-piece sensor all the layers must be bound together. We investigated several adhesives. In order to minimally affect the capacitance value, thin films of adhesives are required.

Thermoformed support frame (section 4) and silicone rubber layers (section 5) have been bound together using a one-component silicone adhesive specifically made for bonding RTV silicone rubber to other silicone rubber or other materials (Smooth-on, Sil-Poxy, U.S.A.). This adhesive has similar consistency as other silicone adhesives, but it is possible to obtain thin films by spreading it with a squeegee.

Flexible PCB (section 3) has been bound to the thermoformed support frame (section 4) using a cyanoacrylate instantaneous elastomeric adhesive (Loctite, Loctite 480). It has good elasticity after curing thanks to its native silicone addition. Its glass transition temperature is 150°C, therefore it is suitable to be used during the thermoforming

process (section 4). Finally, thanks to its liquid consistency it is easy to obtain thin adhesive films.

Silicone Rubber dielectric substrate (section 5) and Conductive silicone (section 6) perfectly bind together, furthermore they show excellent deformation and elongation characteristics (Fig. 6).

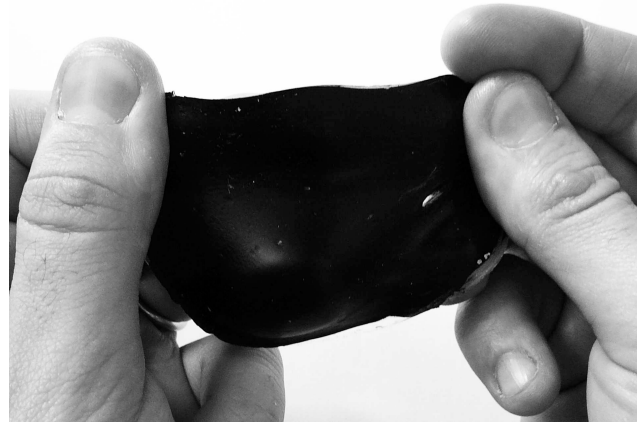


Figure 6. Ecoflex and Silopren binding

9 EXPERIMENTAL RESULTS

In order to validate the functioning of the system we performed experiments on three intermediate phases of implementation process of the tactile sensor, and in particular in the first experiments we tested the sensor response in the case of planar flexible PCB covered by the dielectric and conductive silicone substrates (Fig. 9.1). In the second experiments we tested the sensor response in the case of planar flexible PCB bound to a PETGAG sheet and covered by dielectric and conductive silicone substrate (Fig. 9.2). The third experiments is about the functioning of the fingertip prototype (Fig. 9.3).

Data acquisition has been done using MATLAB RTW (Real-Time Workshop): the sensor has been connected via CANbus to an XPC-target (with a CAN-PCI card), then sent over LAN to a Host PC.

In order to avoid spurious data, due to inappropriate bending of flexible PCB, the sensor has been fixed on a rigid support in order to obtain reliable and repeatable measurements

9.1 Planar Version Operative Test (rubber only)

In this first experiments we tested the sensor response in the case of planar flexible PCB covered by the dielectric and conductive silicone substrates. we activated a zone of the sensor using an indenter loaded with a known weight (45g, 95g, 250g, 500g).

In this phase the sensor pads were covered with a pure Ecoflex 00-30 layer of 2mm thickness. This test aims to understand which is the sensor response to external forces without the rigid support frame.

Test results are shown in Fig. 7: each line refers to a tactel signal; when a load is applied, the 8-bit output stabilizes around a value between 0 and 255. More is the force applied to the tactel, more the line will fall to 0. As the indenter round contact area was about 4mm diameter we have a force distribution which excite more a small

group of three tactels and less the other perihperal ones. Each rectangle corresponds to the application of the different loads. Sensor saturation is over 500g.

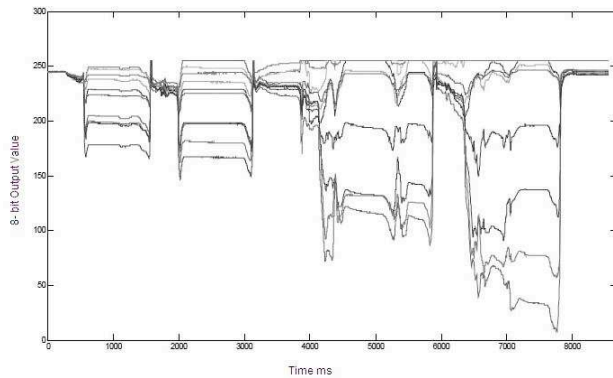


Figure 7. Planar sensor covered with rubber layer

9.2 Planar Version Operative Test (planar PETGAG + rubber)

In this second experiments we tested the sensor response in the case of planar flexible PCB bound to a PETGAG sheet and covered by dielectric and conductive silicone substrate. This intermediate test consists in activating a zone of the sensor using an indenter loaded with a known weight (45g, 95g, 250g, 500g). In this phase a 0.150 mm PETGAG sheet was interposed between the sensor pads and a pure Ecoflex 00-30 layer of 2mm thickness with. This test aims to understand how the chassis affects the sensor behavior. Test results are shown in Fig. 8 from the comparison between this experiment and the previous one, we notice that the measures are only slightly affected by the PETGAG layer introduction.

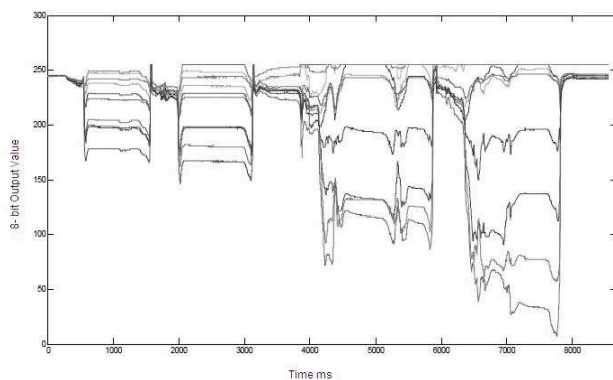


Figure 8. Planar sensor covered with PETGAG and rubber layers

9.3 Fingertip Sensor Operative Test

In this final test the sensor was in its final fingertip shape. The fingertip has been placed over the positive fingertip mold and fixed in order to perform the measurements. The sensor has been activated using a pen as indenter to qualitatively compare the sensor response to previous tests. Test results are shown in Fig. 9; when the fingertip is pressed it has a fast and suitable response to the exerted force.

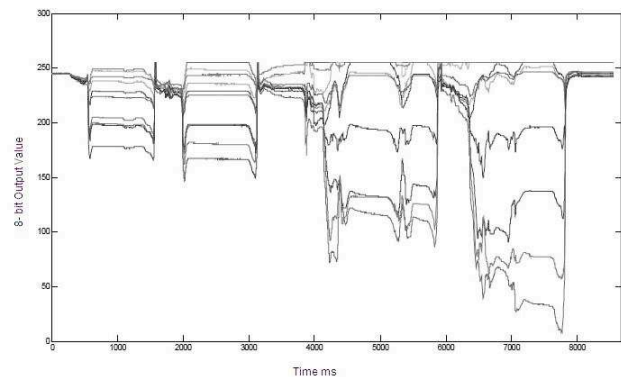


Figure 9. Fingertip sensor qualitative test

10 CONCLUSION AND FUTURE WORK

The fingertip-sensor prototype was produced and tested successfully. It is important to notice that it has not shown any hysteresis due to material deformation. Planned future tests will be performed using a load cell based equipment, in order to better characterize sensor response.

Miniaturization will be taken into account also for embedded electronic circuit which will be manufactured on flexible circuit in order to have a sensor effectively embeddable onto a robotic hand.

The capacitance augmentation did not give the desired results, actually we have not planned to perform experiments with other materials.

Finally the production process can be improved by adopting the vacuum forming technique, using a vacuum forming machine and proper designed molds because the flexible circuit related to the tactels pattern will be attached to the internal support frame before the thermoforming process and mechanical thermoforming could in some case to damage the flexible PCB. In this way the only external intervention at the end of the production will be the binding of rubber layers and the sensor frame support.

This will be the foundation of a process capable of produce small series of sensor units, such as the sensors for a complete robotic hand, using a single customized female mold.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 231500 (Project ROBOSKIN).

REFERENCES

- [1] R.O. Ambrose, T.B. Martin, and M.J. Butzer, 'Tactile gloves for autonomous grasping with the nasa/darpa robonaut', *Proc. of 2004 IEEE International Conference on Robotics and Automation*, 1713–1718, (2004).
- [2] F. Carpi and D. De Rossi, 'Improvement of electromechanical actuating performances of a silicone dielectric elastomer by dispersion of titanium dioxide power', *Proc. of the 2005 IEEE Transaction on dielectrics and Electrical Insulation*, **12**(4), 835–843, (2005).
- [3] R.S. Dahiya, G. Metta, M. Valle, and G. Sandini, 'Tactile sensing: From humans to humanoids', *IEEE Transactions on Robotics*, 2009, (2009).
- [4] B.B. Edin, L. Beccai, L. Ascari, S. Roccella, J.J. Cabibihan, and M.C. Carrozza, 'A bio inspired approach for the design and characterization of tactile sensory system for a cybernetic prosthetic hand', *Proc. of the 2006 IEEE international conference on Robotics and Automation*, 13541358, (2006).
- [5] I. Gaiser, S. Schulz, A. Kargov, H. Klosek, A. Bierbaum, C. Pylatiuk, R. Oberle, T. Werner, T. Asfour, G. Bretthauer, and R. Dillmann, 'A new anthropomorphic robotic hand', *Proc. of the 2008 IEEE-RAS International Conference on Humanoid Robots*, (2008).
- [6] D. Goger, N. Gorges, and H. Worn, 'Tactile sensing for an anthropomorphic robotic hand: Hardware and signal processing', *Proc. of the 2009 IEEE international conference on Robotics and Automation*, 2972–2978, (2009).
- [7] H. Kawasaki, T. Komatsu, and K. Uchiyama, 'Dexterous anthropomorphic robot hand with distributed tactile sensor: Gifu hand ii', *Transaction on Mechatronics*, **7**, 296303, (2002).
- [8] M.H. Lee and H.R. Nicholls, 'Review article tactile sensing for mechatronics -a state of the art survey', *Mechatronics 1999*, **9**(1), 1–31, (1999).
- [9] M. Maggiali, G. Cannata, P. Maiolino, M. Metta, G. and Randazzo, and G. Sandini, 'Embedded distributed capacitive tactile sensor', *11th Mechatronics Forum Biennial International Conference 2008, University of Limerick, Ireland, June 2008*, (2008).
- [10] T. Mouri, H. Kawasaki, K. Yoshikawa, J. Takai, and S. Ito, 'Anthropomorphic robot hand: Gifu hand iii', *Proc. of 2002 Int. Conf. on Control, Automation and Systems*, 1288–1293, (2002).

Towards HRI on the Move with Mixed Initiative

Cristian Bogdan (*), Dominik Ertl (o), Michael Göller (+),
Anders Green (*) and Kerstin Severinson Eklundh (*)¹

Abstract. When a human and a robot are traveling in each other's vicinity, they typically have reduced recognition capabilities while they interact or communicate. In this paper we posit that the design of a human-robot interface has to consider such reduced capabilities to allow for robust human-robot interaction on the move. As we will illustrate, this brings an important set of design challenges, even more so when the user interface is multimodal and takes mixed initiative into account. In beginning to address such challenges, we are charting a design space for UIs that are dedicated to HRI on the move. The design space is based on chronological communication phases. Furthermore, we point out several modality specifics related to movement, for modalities such as GUI, speech input, speech output and gesture recognition. Finally, we discuss several UI constraints for HRI on the move: walking cognitive workload, time constraints and reduced perception capabilities for the human or the robot. Our example is a semi-autonomous robot trolley in a simulated shopping environment.

1 Introduction

A lot of efforts in human-robot interaction (HRI) today assume a more or less static human-robot positioning, whereby a robot is standing still and expects commands from a human who is also stationary while giving these commands. After these commands are issued, movement of the full robot or human body, or just body parts may occur. In this paper we are interested not just in intermittently moving robots or users (like the one described in [5, 6]), but settings where both robot and user are traveling simultaneously and continuously, and have to interact while on the move. In the course of a European funded project (<http://www.commrob.eu>) working on a semi-autonomous robot we found the need to re-consider HRI in regard to the full body movement of both the human and the robot. For example, using a touchscreen or speech input is much harder for a user in a crowded environment when the robot is allowed to move (displace itself) during interaction. It is easier to interact with a static robot in a such an environment, if the robot moves from time to time but may not be interacted with when “on the move”. This is an even more challenging task when both the human and the robot can propose own initiative for interaction (mixed initiative) in a socially acceptable manner, while on the move.

So, to begin addressing some of these “HRI on the move” challenges in the context of mixed initiative we present chronological communication phases charting a design space. We will take our examples from a robot usage situation which we have worked on in

the CommRob project, which corresponds to a well-known everyday scenario: traveling with a semi-autonomous robot through a crowded supermarket while shopping. We have worked with a robot capable of supporting such usage, the interactive Behavior Operated Trolley (InBOT, [2]) that we use as development platform and demonstrator in the course of our project. This robot allows for multimodal communication (GUI, speech input and output, and gesture recognition) and safe navigation in a crowded human every-day environments. Figure 1 shows an improved version of the robot trolley.



Figure 1. Robot Trolley with multimodal UI.

During shopping, both the user and the robot can take initiative. We have taken this particular kind of mixed initiative as our *perspicuous instance* of “HRI on the move”. Multiple modalities can be used artfully to allow the robot to indicate its initiative in ways that do not disturb the user, yet manage to get the message across. Within a lab-simulated supermarket environment, multimodal, mixed initiative interaction is used by our robotic trolley to help the customer to find the desired products without extensive search and to relieve the customer from the burden to push the shopping trolley from his own force all the time, especially if the trolley is heavily loaded or the customer is an elderly person. The user can interact with the robot by way of a touchscreen, speech input (with a headset), and speech output. The robot has several sensors to localize the user.

We performed a number of user studies and experiments where

¹ (*) Royal Institute of Technology School of Computer Science and Communication, Stockholm, Sweden, {cristi.green,kse}@csc.kth.se, (o) Institute of Computer Technology, Vienna University of Technology, Vienna, Austria, erl@ict.tuwien.ac.at (+) FZI Research Center for Information Technology, Karlsruhe, Germany, goeller@fzi.de

a total of 30 users have done simulated shopping with our robot in a laboratory environment. The problems they encountered and the issues we found while trying to improve the interaction led us to delimit several categories of “HRI on the move challenges” which we will present below.

The remainder of the paper is as follows. First, we present mixed initiative in the context of a moving robot. Then we introduce the communication phases charting our HRI design space. Third, we present user input and output interfaces for HRI on the move. At last, we discuss constraints like walking cognitive load, time constraints and reduced perception capabilities for HRI on the move.

2 Mixed Initiative on the Move

Mixed initiative allows either the human user or the robot to start an interaction. This allows for more natural interaction between a human and a robot [3]. We present here two examples of mixed initiative from our shopping trolley context.

Suggestion of product: While on route to some product, the robot suggests an item that the user is probably interested in. This can be an item from a shopping list which is being passed by at the very moment. Such a shopping list is administrated via the touchscreen of the robot. For suggesting such a product the robot can choose between the different modalities available to it. In this case slowing down and uttering a verbal comment (e.g. ‘we are passing’ followed by product name) can be most familiar to the user. This could be assisted by turning the robot in the direction of the item which would catch more attention, thus, reinforcing speech and GUI output.

Suggestion of route: The robot suggests an alternative route which probably poses advantages for the user, e.g., because it is shorter than the originally planned route or the current route became crowded or blocked. Another interesting application can be leading a visually impaired person, who is holding on to the robot’s force sensitive handle bar, along a planned route and this way acting as an intelligent white cane. Here the most suitable suggestion might be slowing down a little and slightly turning in the new intended direction while uttering a speech output. This provides the user with an extra modality to react to the robot’s action: the resistance force on the handle bar.

3 Communication Phases Charting a Design Space

To illustrate the challenges of mixed initiative while on the move, we propose a design space conceptualization to support a robot initiative around the chronological phases of the human-robot communication: the robot initiative, the human perception of initiative, the human reaction, the robot detection of that reaction, and the robot response to the human reaction. The design space can be understood as a form of help and guidance for a robot interaction designer who wishes to design socially acceptable human-robot interaction for robots who will accompany their users while on the move. We will now show how the design space for supporting robot initiative can be supported by considering the chronological communication phases and by examining each phase in turn. A similar charting can be done for the symmetrical case, when the initiative comes from the human. Throughout, we will consider our theme, “HRI on the move” and the strains put by robot movement, time and other dynamic constraints on the HRI design for our and similar settings. The chronological phases of the human robot communication are depicted in Figure 2.

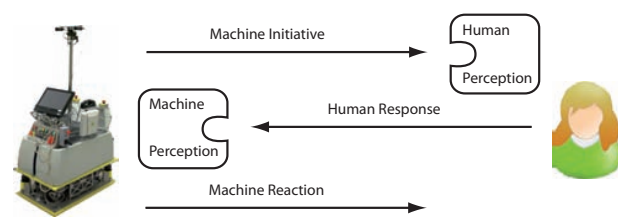


Figure 2. Communication Phases for HRI on the Move.

Machine initiative is the first phase of the process. The design considerations for this phase are depending on the application, i.e. what the initiative is taken about. We have exemplified above two such applications and the possible design decisions in regard to modalities used (speech, GUI) and how to use them. In a lot of cases of interest in our shopping case, this first communication is done on the move.

Human perception of machine initiative follows, after a certain time, the initial initiative. The time needed for the user to notice the initiative, as well as other aspects of the human perception are important factors to consider in a human-robot interaction set-up where human attention may be directed to other components of the environment. If the human is walking along with the moving robot, their attention is affected by the need to look at where to walk. Further robot actions may come too early for the user to comprehend these actions in their logical relation to the initiative.

Human response to the machine initiative can be of different natures. It can range from *tacit perception* whereby the user notices the initiative but does not react in any way that can be understood by the machine, to *acknowledgement* whereby the user reacts to the initiative without attempting to explicitly respond by interacting with the machine, to *interactional response* in which the user interacts with the machine to accept, strengthen, weaken, or reject the machine initiative. For example, for a product suggestion of our robot, the user can acknowledge by slowing down their walking speed, and/or looking at the product, without explicitly interacting with the trolley, yet both such acknowledging cues can help the trolley in shaping its further actions. An interactional response would occur when the user is stopping the trolley via a GUI touchscreen or stopping the trolley by scanning the offered product with a bar code scanner. For route suggestion the user can amplify or reduce the robot turn using the robot’s haptic handle. Let us re-emphasize that all such responses are made on the move.

In case of user acknowledgment without further interaction, a *default action* may be defined during design. In some cases acknowledgment without explicit interaction may be enough to consider the robot initiative as accepted, while in other cases it may only account as partial acceptance, thereby weakening the initiative or canceling it altogether. The difference between acknowledgment and explicit interactional response has played a role for us also in evaluating the movement designs with mock-up robots [1] and currently with the InBOT platform. We found that it is much easier to detect whether the initiative was agreed upon by the user (since in that case the user explicitly responds to it) than it is to detect whether it is noticed or acknowledged, especially if the measuring has to be done on the move.

Design plays a role also in *facilitating interactional response*. Not only the launching of the initiative can be designed but also the levers offered to the user for responding to it can be made more available. In the case of a product offering, the user could not only respond by speech or touchscreen GUI interaction, but also by scanning the offered product with a portable scanner, which may be more handy than reaching for the touchscreen. To reject an offer, the user could also employ movement by pushing a trolley which slows down while offering a product or by turning back to the initial direction a trolley that turns to suggest a route.

Such prompting for user interaction is also useful when the user has the initiative, e.g. when the user asks for the trolley to move to a destination, the trolley can start slowly and accelerate progressively, to allow for the user to stop the action or alter the action. While slowly accelerating the robot sensing capabilities are already affected by the robot moving.

Machine detection of human response is fundamental in providing a socially competent robot behavior. Previous work motivates the need for and gives an initial characterisation of a model for communicative contact and perception feedback [7]. While explicit interactional response is straightforward to detect, passive acknowledgment may play an important role in the initiative interaction, yet it may be harder to detect, especially while moving. For example in the case of product offering via slow-down, the user slowing down themselves can mean that the offer is acceptable (though not surely accepted), while the user not slowing down is almost certain to mean offer rejection or lack of notice about the offer. Similarly in our evaluation work (both with a mock-up robot and with the InBOT) we have seen users looking at the offered product but not following this up with any explicit interaction. This inspired us the idea that user gaze detection (while the robot is moving) can be a useful indicator to automatically detect acknowledgment, which can further help in continuing the slow-down robot action for an eventual explicit interaction response. Another detection of acknowledgment (and acceptance) is the posture of the user reaching out to the shelf to grab the product, which can be detected by image analysis. None of these acknowledgment cues represent certainties, yet probability-based techniques similar to those used in speech or gesture recognition can be employed to combine these cues in order to provide a more natural coordination between user and robot.

The product-offering example thus poses at least three important sensing challenges: sensing user moving speed, sensing the user posture in relation to the shelf, and sensing the user gaze; all these are important challenges since detections need to take place while the robot itself moves. Furthermore this example is among the most illustrative to suggest that the interaction design of machine initiative does not end with the design of the initiative itself, but also includes detection of user response or uptake for the machine to be able to continue the interaction ‘properly’.

Machine reaction to human response involves the machine follow-up of the initiative based on the user response and its machine detection. Several cases need to be considered here: the user may be acknowledging, accepting, strengthening, weakening or rejecting the machine initiative. All these possibilities may lead to design decisions for a certain application. Furthermore, machine detection of acknowledgment can be followed by detection of acceptance or strengthening (or indeed weakening or rejection) so the machine can react to acknowledgment in a way that is consistent with the following reaction in case of subsequent explicit user interaction.

For example, in case of product offering acceptance, we might want to design the trolley to stop, or to do even more by placing itself in a position which facilitates easier product moving from the shelf to the trolley. Performing such movement may introduce time constraints which may or may not be acceptable in the given communication situation. In case of sensing user acknowledgment without any detection of acceptance or rejection, we could decide to make the robot prolong its slowed-down course, to give more time to the user to decide on the machine initiative.

4 User Input and Output Interfaces on the Move

We have already introduced some challenges for implementing mixed initiative while on the move, in chronological approach, which also describes how an ideal implementation would look like. In what follows, we will take a modality-oriented approach to comment on such “on the move” difficulties as they result from our partial (and less than ideal due to issues we will present) implementation of these mixed-initiative design considerations. We will now illustrate them, as well as the difficulties we expect with the design space aspects that we have not yet implemented.

4.1 Using a GUI with a Touchscreen

Our robot has a touchscreen for GUI input and output. Such a GUI can be used while both the human and the robot are standing. However, it can be used as well when the human and the robot are both moving which poses several problems. We encountered a large number of instances where, although walking close to the robot, the users repeatedly missed targets on the touchscreen. We soon realized that the touchscreen transparent surface has a certain thickness and if it is not looked at from a proper angle (due to walking a bit behind), there is a refraction which makes the users touch the screen lower than they should, hence missing the target. Furthermore, when trying to correct this action, users move their fingers on the screen but it is the first touch that matters for the touchscreen device driver, so the only way to correct the action is to take the finger off the screen and touch again, at the right spot, which no user managed to learn.

This rather simple example shows one interesting issue: our touchscreen is not designed for use on the move. While on a static usage scenario the required viewing angle (with low refraction) is easy to “find” through a correct sitting position, or, lacking that, through trial and error, hitting the targets right on the move is much less successful. While this may indicate that a new touchscreen generation could address such issues for use while moving/traveling, our point with this example is to illustrate that “use on the move” poses specific requirements to input and output devices.

4.2 Using Speech Input and Output

It may be argued (and in HRI it has been argued) that the touchscreen is not the most appropriate device to use on the move, and that stopping the trolley could be performed via, e.g., speech. We soon realized that speech recognition while on the move is a quite different matter from static speech input, as it is made more difficult by the sounds of the robot engines and wheels, by the sound of user steps, by the words being uttered differently by the walking users, environmental sounds, etc. Speech output needs to be adjusted as well, with higher volume due to the sounds specific to robot moving and user walking.

4.3 Using Gesture Recognition

Recognition of gestures performed by a human user is still a matter of research, even more when the user and robot are moving [4]. Such a system recognizes symbolic gestures like “Stop” or “Come To Me” as well as deictic gestures like “Go There”, accompanied by showing the direction. This modality relies on the vision system capabilities that are mounted on the moving robot that has to detect and recognize gestures of a (moving) human. It is not only a matter of the underlying computing vision algorithms, as the constraints have to be considered in the interaction design of a robots’ multimodal UI as well. Typically, the user has to be in front of the cameras and may not walk around just waving her hands. When the human and the robot are moving (at least one of them), the system may not recognize the gestures correctly. We recognized the required light settings for the vision system to distinguish and recognize colored objects properly as another issue for gesture recognition systems. Optimal light can be achieved comparably easy in a laboratory environment. However, in a real-world example this can lead to a malfunctioning gesture recognition when the human and the robot enter, e.g., a shadowy area.

5 Constraints for HRI on the Move

Besides the modality oriented set of difficulties for mixed-initiative HRI on the move, we identified more constraints for designing such interaction, which we discuss below.

5.1 Walking Cognitive Workload

Slowing down when the robot “shows” something to the user is a functionality of our robot, yet only later we came to reflect on why this is appropriate. Our current understanding is that walking poses a certain cognitive workload which reduces the attention that the user can invest for looking in other directions than the walking direction, looking at specific details like the touchscreen, etc. Reducing the movement speed can thus be necessary if important aspects of the environment must be considered, or if other important facts must be considered in the human-robot communication subject. Again, like for other aspects of the interaction on the move, we wish to consider this cognitive load aspect in a more general way, as a generic condition for HRI on the move.

5.2 Time Constraints

In order for the human-robot communication to take place in a socially acceptable manner, some operations need to be performed within a certain time window. The robot slow-down exemplified a situation where the robot tries to give more time for the user to consider an offer and react.

We can exemplify “latest” time constraints. Some interactions have to be performed in a specific short time frame (“let’s go to the right”, “oh, stop a second”, “no - left, not right”). These are cases where the complete interaction cycle has to take place in few seconds. This means: detection of the need of interaction, taking the initiative and start a communicative act, perception of the communicative act by the other partner, interpretation of the act, performing an action to respond to the communication, detection of the action by the first partner, interpretation by the first partner.

“Earliest” time constraints exist too. Sometime there has to pass a time span which is needed before a reaction to a communicative act is possible or before taking the initiative to start a communicative act

is possible. This depends on the available modalities. For example: the available modality is a GUI on a touchscreen: *Robot*: “here is a special offer, if you want me to stop, press the screen!”, but the user is 3m away from the robot therefore it takes some seconds for the user to reach the screen.

If both types of constraints (earliest and latest) appear in the same situation, the robot designer must provide heuristics for determining which one is more important, and how does that affect the communication to the user.

5.3 Reduced Machine Perception Capabilities

We have already introduced the reduced human perception capabilities on the human side due to walking. On the technical side, many sensors behave differently (generally more poorly) while on the move. Cameras give fuzzy images, they may give shaky pictures when on a moving robot due to a shaking platform. As mentioned already, noise from the robots drive system disturbs speech recognition. And TimeOfFlight (3D) cameras have great problems with moving objects. Such camera system creates distance data with help of the time-of-flight (TOF) principle. Summed up, electro-mechanical systems like robots have reduced perception capabilities while on the move.

6 Related work

The Autonomous City Explorer project (<http://www.ace-robot.de>) aims to create a robot that will navigate in an unstructured urban environment and find its way through interaction with humans. Interaction while moving is not as much in focus as for our project yet we believe that the ACE project can constitute a further instance of the “HRI on the move” problematic.

In other instances that illustrate challenges posed by “communication on the move” Leet et al. [10] consider speech-based email during vehicle driving and its effects on attention to the road, and Keskinpala et al. [9] look at interacting with robots via portable devices such as PDAs.

7 Conclusion

In this work we present the challenges we identified when designing mixed-initiative multimodal human-robot interaction while on the move. A major ingredient of these challenges is movement: robot and user movement, their relative distance, relative speed, their slowing down, speeding up, etc. We presented chronology (interaction flow)-oriented, modality-oriented and constraint-oriented charting and illustration of such challenges. Based on such reflections, we believe that considering interaction on the move is an important theme to be addressed by HRI in the future by e.g. working on more intelligent/more robust communication/interaction models, which anticipate communication failures and constrain violations on the move.

ACKNOWLEDGEMENTS

This research has been carried out in the CommRob project (<http://www.commrob.eu>), partially funded by the EU (contract number IST-045441 under the 6th framework programme).

REFERENCES

- [1] Bogdan, C.; Green, A.; Hüttenrauch, H.; Räsänen, M.; Severinson Eklundh, K.: Cooperative Design of a Robotic Shopping Trolley, In: Proceedings of COST-298, 2009
- [2] Göller, M.; Kerscher, T.; Zöllner, J. M.; Dillmann, R.: Setup and Control Architecture for an Interactive Shopping Cart. In: Proceedings of the 14th International Conference on Advanced Robotics (ICAR), 2009
- [3] Adams, J. A.; Rani, P.: Mixed-initiative interaction and robotic systems. In: Proceedings of AAAI Workshop on Supervisory Control of Learning and Adaptive Systems, 2004
- [4] Burger B.; Ferrane, I.; Lerasle, F.: Multimodal Interaction Abilities for a Robot Companion. In: Proceedings of International Conference on Computer Vision Systems, ICVS 2008
- [5] Kosuge, K.; Hirata, Y.: Human-Robot Interaction. In: IEEE International Conference on Robotics and Biomimetics, ROBIO 2004, pp.8-11, 22-26 Aug. 2004
- [6] Goodrich, M. A.; Schultz, A. C.: Human-robot interaction: a survey, In: Journal of Found. Trends Hum.-Comput. Interact., vol. 1, num. 3, 2007, p. 203–275, Now Publishers Inc.
- [7] Green, A.: The need for contact and perception feedback to support natural interactivity in human-robot communication. In Proceedings of IEEE 16th International Symposium on Robot and Human Interactive Communication (RO-MAN 2007), pages 552557, Jeju, Korea, August 26-29 2007.
- [8] Kaindl, H.; Falb, J.; and Bogdan, C.: Multimodal communication involving movements of a robot. In CHI '08 Extended Abstracts on Human Factors in Computing Systems (Florence, Italy, April 05 - 10, 2008). CHI '08. ACM, New York, NY, 3213-3218.
- [9] Keskinpala, H.K., Adams, J.A., Kawamura, K.: PDA-based human-robotic interface, IEEE International Conference on Systems Man and Cybernetics, 2003
- [10] Lee, J.D., Caven B., Haake S., Brown T.: Speech-based interaction with in-vehicle computers: The effect of speech-based e-mail on drivers' attention to the roadway. Human Factors, 43 (4), 2001

Touch-Triggered Protective Reflexes for Safer Robots

Torbjørn S. Dahl and Andrew Palmer¹

Abstract. Covering robots with a skin-like sensor can enable a range of skills that will improve human-robot interaction (HRI) both in terms of safety and communication. This position paper presents a series of concepts and problems related to developing middleware technologies to support robot skin. We define middleware as generic tactile skills that will be useful to the HRI community independent of application area and focus in particular on touch-triggered protective reflexes that can make HRI safer for both humans and robots. We review current literature on protective reflexes in biological systems and present the design of two experiments for capturing protective reflexes in humans during reaching and balancing.

1 Introduction

Collision detection has, together with low momentum and compliance, been identified as a main requirement for *safe robots* [15]. Collision detection is, in humans, closely related to a set of protective reflexes that translate the detected collisions into actions. The purpose of these actions is to keep both the person sensing the touch and his/her environment safe by stopping potentially dangerous movements and moving parts of the body away from potentially harmful positions. This paper presents concepts, technologies, problems and potential solutions for the realization of touch-triggered protective reflexes on humanoid robots. The reflexes are modeled on human protective reflexes and implemented using *robot skin*, a new sensor hardware technology that provides touch sensitivity for large areas of a robot body.

1.1 The ROBOSKIN Project

The ROBOSKIN project² is funded by the European Commission's Framework Programme 7 (FP7) and its activities are centered around new sensor technologies that can cover large areas of a robot's body with a touch sensitive material, i.e., *robot skin*. The activities of the ROBOSKIN project span a range of levels from sensor hardware to application specific skills. The project studies hardware for sensing not only the force and location of touch, but also the direction of the force and other non-touch phenomena such as heat and proximity. The ROBOSKIN project is also developing a level of problem domain and robot platform independent middleware such as protective reflexes and touch recognition and production skills. Finally, the ROBOSKIN project will study the application of robot skin technology in several domains, including programming by demonstration [5] and HRI for therapeutic play [9]. This paper focuses in particular on the protective reflexes activities of the ROBOSKIN project.

¹ Robotic Intelligence Lab, University of Wales, Newport, UK, email: torbjorn.dahl|andrew.palmer@newport.ac.uk

² The ROBOSKIN project web site can be found at <http://www.roboskin.eu>.

2 Tactile Sensor Technology

Human skin has a wide range of sensory capabilities provided by different receptor cells. The sensory modalities of the skin include temperature (thermal receptors), pressure (e.g., Ruffini corpuscles), vibration (Meissners and Pacinian corpuscle end-organs) and pain (nociceptors) [7]. An illustration of the human skin with its most common receptor cells is provided in Figure 1.

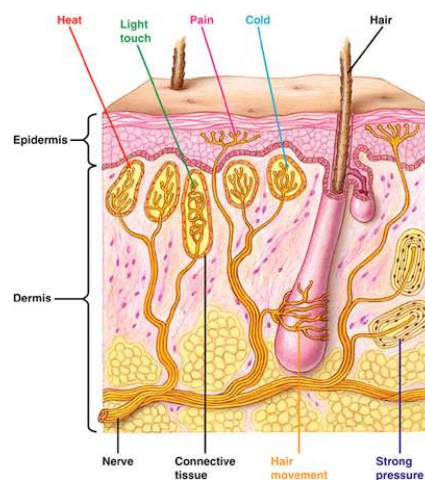


Figure 1. The different sensory receptors embedded in the human skin, from the University of Miami
(<http://www.bio.miami.edu/cmallery/bil150.htm>).

Our work will provide study protective reflexes. These can be triggered by both thermal receptors and mechanoreceptors. However, the first generation robot skin prototype will only have pressure sensory capabilities. Historically a wide range of sensory technologies has been studied in the context of robot tactile sensing including micro-electromechanical systems [3], force-sensitive resistors (FSRs) [17] and a quantum tunneling compound [4]. Dahiya *et al.* have published a comprehensive review of human sensory capabilities and potential technologies for robot skin [8]. It is an explicit objective for the ROBOSKIN project to develop robot skin technology further, including an integration of multiple sensory modalities into a single sensor. The initial work, however, will be focused on a new sensor technology that provides purely tactile feedback from large areas of the body.

A picture of the initial prototype skin sensor that will be used for the ROBOSKIN work is given in figure 2. Each of the golden disks on the interconnected triangular base plates is an individual touch sensor. Cannatta *et al.* [6] have presented this technology in detail.



Figure 2. The first robot skin prototype fitted on a human hand.

3 Human Protective Reflexes

The simplest level of sensor-motor integration in biological entities consists of reflex actions [7]. The simplest human reflex, the *monosynaptic stretch reflex*, relies on a single synaptic connection, within the spinal chord, between the receptor cells sensing the stimulus and the effector cells producing the response. The directness of this connection allows humans to respond to increases in muscle stretch within 50 milliseconds. Increasing the intensity of the reflex stimulus reduces the response time and increases the duration and amplitude in the response. The monosynaptic stretch reflex plays an important role in balance and posture control.

All human reflexes, other than the monosynaptic stretch reflex, are *polysynaptic*, i.e., the neural pathway or *arc*, between the sensors and effectors contain multiple synapses. Polysynaptic reflexes are subject to inhibition from the brain, e.g., the reflexive extension of the fingers as a response to painful stimuli can be inhibited when carrying a hot casserole. Among the simpler polysynaptic reflexes we find the *nociceptive withdrawal reflex*, i.e., the withdrawal of a limb as a response to *noxious* or painful stimulus produced by mechanical, thermal or chemical stimulation. This is the reflex we aim to study and reproduce in this work. Other polysynaptic reflexes are more complex and include other sensory modalities, e.g., the *pupillary light reflex* that reduces the diameter of the pupil in response to light intensity and the *primitive grasp reflex* that closes a baby's hand when something touches their palm.

3.1 Reflex Receptive Fields

Research on the *cutaneous* or skin-related nociceptive withdrawal reflex (NWR) has shown that a reflex has a related *reflex receptive field* (RRF) and that each RRF controls only a single or small group of muscles [22]. As the stimuli moves towards the edge of the RRF, the response is reduced. Further research on RRFs has shown that such fields are also clearly observable in humans [2]. Two RRFs on the sole of the human foot are illustrated in Figure 3.

The two RRFs illustrated in Figure 3 control two muscles in the leg, the tibialis anterior and the peroneus longus. The locations of these two muscles in the human leg are illustrated in Figure 4.

Another observed effect in relation to NWR is the increased excitability during voluntary movements compared to the excitability during static body postures [11]. Finally, research on infant humans and monkeys has shown that the ability to inhibit reflexes increases with age, allowing increased cortical control as a person matures [10].

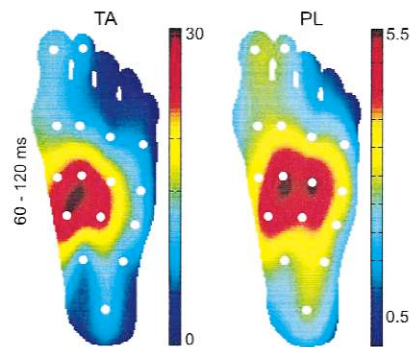


Figure 3. The receptive fields of the tibialis anterior (TA) and peroneus longus (PL) muscles, from Andersen *et al.* [2]



Figure 4. The muscles activated by the RRFs in Figure 3, tibialis anterior and peroneus longus, from the UCLA online Health Encyclopedia (<http://www.uclahealth.org/>).

In order to implement protective reflexes on a humanoid robot, we need to map the responses produced by a range of different stimuli. From such maps we can then develop an implementable reflex model. In addition to this mapping, however, there are also still several open questions w.r.t. the basic nature of protective reflexes. The two questions we see as the most critical for the successful production of an appropriate model are:

1. Are reflexes sensitive to the force direction as well as the force magnitude of a touch?
2. Are reflexes expressed by a single dominant muscle?

We plan to address the first question experimentally, by producing appropriate stimuli and capturing the resulting response. If reflexes are force direction insensitive, it will reduce significantly the size of the problem we are modeling. As our prototype technology cannot sense force directions, it will also not be possible to implement force direction sensitive reflexes until later generations of robot skin technology become available.

The second question is more difficult to answer. Most reflexes involve multiple joints, but it is possible that only one joint is dominant while the remaining joints are only compliant so as to accommodate the movement produced by the dominant joint. Using the techniques presented by d'Avella *et al.* [12], we will evaluate motion

data from human protective reflexes. The basic reflex arc may help explain muscle synergies such as those found by d'Avella *et al.* and we hope to be able to find similar synergies in human withdrawal reflexes.

4 Capturing Human Protective Reflexes

A range of methods have been used to elicit withdrawal reflexes in humanoids. These include the application of pin-pricks or heat [16], reaching with a transparent barrier [10], direct electronic stimulation of the sensory nerve [24] and electronic stimulation of the sole of the foot [23]. To capture the reflexes, a goniometer or angle-meter, is commonly used [1], measuring the reflex for a single joint.

Our work will use the Gypsy-6 mechanical full-body motion capture suit from Animazoo Ltd.³ to capture protective reflex responses. The Gypsy-6 suit is shown in Figure 5.



Figure 5. The Gypsy-6 mechanical full-body motion capture suit, from the Animazoo web-site (<http://www.animazoo.com/>).

The Gypsy-6 suit allows us to measure the angle of a response at multiple joints simultaneously. Understanding how a stimuli affects a range of effectors will increase our understanding of protective reflexes and will allow us to implement such reflexes on humanoid robots. Figure 6 shows a graphical representation of motion data capture during reaching.



Figure 6. Single frame from motion captured in the reaching scenario developed for analysis of protective reflexes in humans.

³ Further information on the Gypsy-6 motion capture suit is available from the Animazoo web-site at <http://www.animazoo.com>

To produce high-quality data for our models we have formalized a set of experiments over a range of parameters and scenarios. The parameters considered reflects issues such as whether the subjects are expecting to be touched and the direction of the touch force. The three main scenarios, reaching, standing and walking, have been formulated to cover situations where the expression of protective reflexes takes significantly different forms.

Two main factors differentiate the scenarios, the need to maintain balance and the need to recover from a perturbation to achieve a goal. With respect to goal orientedness, in the reaching and walking scenarios the robot is engaged in active, goal oriented behaviors when the unexpected touch takes place. This means that the perturbation produced affects the performance of those behaviors and raises questions about how to recover. In the standing scenario, on the other hand, there is no explicit active behavior and thus, no explicit need for recovery. With respect to balance maintenance, in the reaching scenario there is no need to maintain balance and perturbations do not have implications for other joints than those in the arm. In the standing and walking scenarios, on the other hand, the perturbation has consequences for several links in the kinematic chains around the touch point as the whole body posture must be moderated to avoid the robot falling over. Table 4 summarizes the different requirements of the different scenarios.

Scenario	Goal Oriented	Balance Maintenance
Reaching	Yes	No
Standing	No	Yes
Walking	Yes	Yes

Table 1. The differences in requirements for the three scenarios considered for analysis of human protective reflexes.

5 Formalizing the Reaching Experiment

When unexpected touches are experienced during reaching, the arm is generally moved quickly away from the stimuli, i.e., in the direction opposite to the side of the arm where the touch was made. We have noticed, during initial explorations that the perturbations produced in this scenario, rarely changes the configuration of the lower arm, but appears to be produced mostly by a single movement in the shoulder.

This experiment is designed to capture and measure the withdrawal reflex activated when a subject reaches for an object and encounters an unexpected and unpleasant stimulus. It also aims to capture the subsequent trajectory correction performed to successfully reach the target object. The experiment will measure the reflex reactions caused by stimulating several different locations on the arm during reaching. It is desirable to record the same reaching action both with and without a reflex reaction, as the final implementation of these reflexes on a humanoid robot will express them as a modulation of the underlying behavior.

Each trial will record the unmodulated and then the reflex-modulated actions resulting from reaching from a fixed starting point to a fixed target object and with a fixed stimulus location. These fixed locations are designed to stimulate a certain point on the subject's arm, and to allow for a consistent stimulus location in each trial across all the subjects. During one reaching trial, the subject will start from a set position and reach for a target object several times. During one of these reaching motions, a stimulus will be placed appropriately so that it will make contact with the desired area of the arm.

This will allow the natural, unmodulated action to be captured, along with the same action modulated with a reflex. In order to make the stimulus as unexpected as possible, the subject will be blindfolded during the second part of each trial when the stimulus is present. Also, there will be a random number of blindfolded reaching motions when a stimulus is not present. The subject will also wear some form of hearing protection during the second part of each trial, to prevent the subject from gaining any information about the presence of a stimulus by auditory cues.

5.1 Stimulus Locations

Six locations on the arm have been identified as being potentially interesting, as they are the most likely places that would encounter some kind of obstacle during reaching. All of these are described for the right arm, a trial involving a left-handed subject would be exactly the same except that all relative position references would be to the left, not the right.

Upper arm: For the upper arm we have identified two locations that will be stimulated during the trials:

1. **Upper arm 0 degrees** - the point midway between shoulder and elbow, on the biceps
2. **Upper arm 90 degrees** - 90 degrees to the right of location 1: the point midway between shoulder and elbow, on the outside of the arm

Lower arm: For the lower arm we have identified four further locations that will be stimulated during the trials:

3. **Lower arm 0 degrees** - the point midway between elbow and wrist, on the palm side of the lower arm
4. **Lower arm 90 degrees** - 90 degrees to the right of location 3: the point midway between elbow and wrist, on the outside of the lower arm
5. **Lower arm 180 degrees** - 90 degrees to the right of location 4: the point midway between elbow and wrist, on the same side of the lower arm as the back of the hand
6. **Lower arm 270 degrees** - 90 degrees to the right of location 5: the point midway between elbow and wrist, on the inside of the lower arm.

5.2 Experimental Setup for Reaching

Target Object: We want to measure not only the time and amplitude of the reflexes, but also the accuracy of the subject's reaching. This is because our data will be published as a protective reflex benchmark. It is not necessary for the subject to actually take or even grasp the target object, only to touch it. Therefore, the target object will be a small blackboard, approximately 15cm square, with a cross painted in the center. During the first part of each reaching trial, when the subject can see, they will be told to reach for the cross and touch it with their fingertips. The subject's fingers will be dipped in powdered chalk before each blindfolded reaching motion, allowing the end-point of the subject's reaching trajectory relative to the central cross to be recorded. This object will be held on a suitable stand with a sufficiently heavy base. A suitable stand could be made from a microphone stand, adapted with a heavier and more stable base.

Stimulus Object: The obstacle will be designed so that when encountered, it will provoke as strong a reflex reaction as possible. To achieve this, it will incorporate some kind of edge or point. The current design consists of a metal wedge, the point of which will come into contact with the subject, and incorporating an FSR to measure the force with which the stimulus was encountered by the subject. The obstacle will be held on an adjustable mechanical arm that has a sturdy base. As above, a modified microphone stand could be suitable. It will be placed so that the subject will still have an alternative path to the target object.

Starting Positions and Target Locations: For each of the stimulus locations discussed above we have formalized the experimental setup in terms of the starting position of the subject, the location of the target and the location of the stimulus.

Location 1 (upper arm 0 degrees - biceps):

- **Starting position:** Right arm held down, parallel with the body, palms facing inwards
- **Target object location:** At subject's elbow height, 10 cm away from fingertips when the lower arm is raised 90 degrees.
- **Stimulus location:** Placed close to the biceps muscle, on a mechanical arm parallel to the subject's lower arm when raised and fixed at the other end. This will allow the subject to reach the object to the left or right of the obstacle after encountering it.

Location 2 (upper arm 90 degrees outside of upper arm):

- **Starting position:** Right arm held down, parallel with the body, palms facing forwards.
- **Target object location:** At subject's shoulder height, at arm's length away.
- **Stimulus location:** Placed close to the desired stimulus location, held on a mechanical arm parallel with the shoulders. Again, this will allow the subject to reach the object to the left or right after encountering the obstacle.

Location 3 (lower arm 0 degrees palm-side of lower arm):

- **Starting position:** Right arm held down, parallel with the body, palms facing forwards.
- **Target object location:** At subject's elbow height, a little more than the length of the subject's lower arm away in front of the subject.
- **Stimulus location:** Held on a mechanical arm that is parallel with the subject's lower arm when it is raised.

Location 4 (lower arm 90 degrees - outside of lower arm):

- **Starting position:** Right arm held down, parallel with the body, palms facing inwards.
- **Target object location:** At subject's elbow height, a little more than the length of the subject's lower arm away in front of the subject.
- **Stimulus location:** Held on a mechanical arm that is parallel with the subject's lower arm when it is raised.

Location 5 (lower arm 180 degrees back of lower arm, same side as back of hand):

- **Starting position:** Right arm held down, parallel with the body, palms facing inwards
- **Target object location:** At subject's shoulder height, at arm's length away to the side of the subject
- **Stimulus location:** Similar to location 2

Location 6 (lower arm 270 degrees inside of lower arm):

- **Starting position:** Right arm held down, parallel with the body, palms facing forwards.
- **Target object location:** At subject's elbow height, at arm's length away but on the left hand side of the subject's body.
- **Stimulus location:** Close to the center line of the subject's body, a small distance in front of the subject, at elbow height.

5.3 The Complete Trial Procedure for Reaching

Each subject will be stood, wearing the motion capture suit, and with their sleeves rolled up. A target object will be placed a set distance front of them (in the appropriate position, as described above). They will then be told to touch the cross in the center of the target. This will be repeated three times, with the target in the same place. The subject will then be blindfolded, wear hearing protection, dip their fingers in powdered chalk, and be told to reach for the same target. This will again be repeated a number of times, but during one of these a stimulus will be put in the way (in the position described above). The reaching motion on which this occurs is randomly chosen to prevent any influence by the researcher on the day of the experiment. No subject has the same number for both trials. Also, the stimulus is never placed on the first blindfolded trial, to allow the blindfolded reaching accuracy to be measured. The subject's starting position and the locations of the target object and stimulus vary for each trial, as described above. Also, there will be a fixed time period between each reaching movement. This is to prevent the subject from guessing any information about when the obstacle has been placed. 20 seconds should be enough time.

5.4 Further Evaluation of the Reaching Experiment

Further to this experiment, we would like to assess whether reflexes are sensitive to touch force direction and whether there is a single dominant muscle/degree of freedom that produces the perturbations. The reaching scenario is the scenario in which such effects will be most visible as it does not require the robot to maintain balance, something that could potentially blur the effect of the perturbation itself.

Another interesting element to keep an eye on during these experiments is the differences between left and right-handed people. It would also be interesting to tell the subjects to reach with different speeds, fast versus normal versus slow/cautious. With fast reaching is that the stimulus becomes stronger and it would be interesting to see what the effect on the response would be.

6 Formalizing the Balance Reflex Experiment

Unexpected touches to the upper body will, when humans are standing still, make them shift the body away from the touch point. Touches to the legs will typically produce steps away from the touch point. When the body is shifted, it is typically shaped into new, stable position. The new position, however, typically places an increased

amount of stress on certain joints, i.e., the new poses are 'uncomfortable'. They require muscle activity and are difficult to maintain over a long period of time.

In this scenario we are particularly interested in the recovery from the perturbations produced. Without an active behavior to take over control, we are interested in seeing how recovery will take place and what the differences will be between recovery in this and the two other scenarios.

This experiment is designed to capture the unconscious movements performed when a human subject is pushed or poked off-balance (the balance reflex). During this experiment, each subject will be stood up in the motion-capture suit and blindfolded, and poked with a suitable stimulus in one of four locations, and with varying force. The resulting movements performed to stay upright will be recorded by the suit.

During this initial experiment into measuring balance reflexes, there will be four stimulus locations:

- **Left shoulder** a poke into the left shoulder from the side, in line with both shoulders.
- **Right shoulder** as above, for right shoulder.
- **Back** a poke into the middle of the back
- **Chest** a poke into the middle of the chest.

The Stimulus Object: The stimulus object will be based on the one used for the withdrawal reflex experiment described above, but mounted on a suitable boom.

The Complete Trial Procedure for Balancing The subject will be stood, wearing the motion capture suit, and blindfolded. They will then be poked in the appropriate location and with the appropriate force.

Further Evaluation of the Balancing Experiment It would be interesting to evaluate whether the force direction, i.e., incident angle of the stimulus with the skin, of the stimulus affect the response. Human skin can probably sense differences in force direction while the current robot skin technology cannot. It is also unclear whether we can design an experimental procedure that can produce different but consistent and accurate force direction stimuli.

Finally, it would also be interesting to note any differences for one given stimuli between those recorded on people who had that stimulus first and those who had it second. This would highlight any differences due to the subject's reactions after the first reaching measurement, e.g., location 2 is measured second on subject number 1, but third on subject number 6.

7 The Walking Experiment

The walking scenario is the most ambitious as walking is a difficult problem in its own right, made increasingly difficult by the addition of protective reflexes. Perturbations in this scenario typically include changing in posture, turning the body and stepping away from the touch point. These complex perturbations appear to require multiple forms of recovery. The natural body posture for walking must be recovered as must the orientation towards the goal the robot was heading for before the unexpected touch took place. The most interesting features of this scenario are the effects of perturbation on a goal oriented behavior that involves the whole body and also the interaction between the recovery of natural body posture and goal oriented behavior.

We have not yet formalized this experiment as we are expecting the lessons from the reaching and balance experiments to heavily influence our process for experimentation. Thus, we will wait for those lessons before investing more time in the formalization.

8 General Methodological Considerations

It is difficult to predict how much information each participant should be given about the experiment, e.g. whether to give them any forewarning at the start of the experiment that there will be an obstacle. If no warning at all is given, the reflex reaction should be more natural. However, the subject may then not continue reaching for the target, and immediately following the reflex reaction (whatever that may be), there will be some other motion, such as moving the arm completely away from the test area, gesturing at the experimenter, etc. If some warning is given, for example, by telling the subject, once blindfolded, to Continue reaching for the target object no matter what happens, then the reflex reaction could be less natural, as the subject now has some foreboding that something will happen, but then at least they will continue towards the target.

9 Representing Protective Reflexes

The purpose of our motion capture experiments is to collect data describing both unperturbed and perturbed behaviors in order to separate the motions produced by the reflexes from the motions produced by the behaviors that are active prior to the unexpected touches. From the isolated reflex data we hope to produce motions that we can simply add to ongoing motions whenever a robot encounters unexpected touches. Assuming that we can isolate and represent these reflex perturbations, questions arise as to how to represent the motion related to recovery and how to integrate these motions with the implementations of the domain specific behaviors. The integration mechanisms should make it as easy as possible to include reflexes across different problem domains and robot platforms. Our aim is to produce a platform independent library of reflexes for humanoids that can be easily integrated into existing robot behavior development frameworks [13, 21] in a way that hides the reflexes and allows roboticists to develop new robot behaviors without considering their potential effects.

9.1 Integrating Reflexes and Behavior

We have identified two distinct approaches to integrating a library of protective reflexes with higher level behaviors. First, we can integrate the reflexes at the lowest level possible, i.e., at the motor interface level where raw joint speeds and angles are specified, e.g., the YARP arm interface provided with the iCub software [13]. We call this approach *motor interface integration*. Motor interface integration has the advantage of being representation independent, leaving roboticists free to use whatever representation they think appropriate to develop the domain specific behaviors, whether this is operational space methods [18] or radial basis function neural networks [20]. A disadvantage in motor interface integration is that it becomes difficult to represent and integrate complex reflexes, e.g., the whole body posture changes observed in perturbed walking. For such reflexes, a representation level integration would be more natural, implementing reflexes using the same framework that is used to implement the walking in order to produce changes to the overall walking pattern rather than to individual joints or legs.

Representation level integration requires representation specific libraries of reflexes, e.g., a set of reflex behavior modules in a common

behavior-based development framework such as iB2C [21]. Representation level integration appears better suited for expressing reflexes with general effects on complex behaviors such as walking or when whole body constraints are present such as when balance must be maintained. The apparent disadvantage to this approach is that the reflex representations would no longer be generic. The reflexes would have to be re-implemented for each representational framework in order to support relevant interaction mechanisms.

We plan to implement and compare both of these approaches to integration. It is also possible that a generic low-level representation can be found that supports quick re-implementations of reflexes in novel representational frameworks in the same way that, in programming, C libraries are relatively easy to wrap up in thin interfaces making them quickly available to new programming languages.

10 Platform Independent Reflexes

Another issue related to the general applicability of a reflex library is the expression of the captured motion on different humanoid robot platforms. Different limb dimensions, relative joint positioning, weight and motor efficiency means that motor speeds or joint angles cannot be transferred directly between platforms. Several solutions to this problem has been proposed. We will be looking at two in particular, a constraints-based approach that has been successfully applied in computer graphics where this problem is referred to as *retargeting* [14]. This approach allows a developer to specify space, time and frequency constraints for the target motion. The constraint-based methods have the limitation that the exact kinematics of the target physiology must be known. This is typically the case in robotics, but a more generic approach would be to learn a mapping between the physiology that produced the original motion data and the kinematics of the target humanoid robot platform. This would alleviate the need for explicit kinematics models. Another problem with a kinematics-based transfer is that differences in the dynamic models of the human origin and the humanoid target can lead to unstable behaviors [19]. We will explore whether it is possible to learn mappings between a human physiology and a humanoid target in a way that produces stable behaviors. The different humanoid platforms we consider for this work, the iCub platform developed by the RobotCub EC FP7 project, the Nao platform from Aldebaran Robotics, and the KASPAR robot developed at the University of Herfordshire, are presented in Figure 7.



Figure 7. The robot platforms considered as targets for the protective reflexes, the iCub, the Nao and KASPAR

11 Conclusions

The work presented here aims to model human protective reflexes and to implement similar capabilities on humanoid robots in a way that is as generic as possible. In order to achieve this, we will study human protective reflexes in detail. We will also evaluate a range of technologies for representing and integrating reflexes with other, domain specific behaviors. Finally, we will evaluate technologies for transferring motions between different humanoid platforms in order to produce reflex libraries that, as far as possible, are platform independent. This paper has presented initial concepts and problems, the work is planned to go on until May 2012.

Acknowledgement

The research leading to these results has received funding from the European Commission's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 231500.

REFERENCES

- [1] Ole K. Andersen, Nanna B. Finnerup, Erika G. Spaich, Troels S. Jensen, and Lars Arendt-Nielsen, 'Expansion of nociceptive withdrawal reflex receptive fields in spinal cord injured humans', *Clinical Neuropsychology*, **115**(12), 2798–2810, (2004).
- [2] Ole K. Andersen, Finn A. Sonnenborg, and Lars Arendt-Nielsen, 'Modular organisation of human leg withdrawal reflexes elicited by electrical stimulation of the foot sole', *Muscle & Nerve*, **22**(11), 1520–1530, (1999).
- [3] D. J. Beebe, A. S. Hsieh, D. D. Denton, and R. G. Radwin, 'A silicon force sensor for robotics and medicine', *Sensors and Actuators A: Physical*, **50**(1), 55–65, (1995).
- [4] D. Bloor, K. Donnelly, P. J. Hands, P. Laughlin, and D. Lussey, 'A metal-polymer composite with unusual properties', *Journal of Physics D: Applied Physics*, **38**, 2851–2860, (2005).
- [5] Sylvain Calinon, Florent Guenter, and Aude Billard, 'On learning, representing and generalizing a task in a humanoid robot', *IEEE transactions on systems, man and cybernetics, Part B. Special issue on robot learning by observation, demonstration and imitation*, **37**(2), 286–298, (2007).
- [6] Giorgio Cannata, Marco Maggiali, Giorgio Metta, and Giulio Sandini, 'An embedded artificial skin for humanoid robots', in *Proceedings of IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pp. 434–438, Seoul, Korea, (August 20–22 2008).
- [7] N. R. Carlson, *Physiology of Behaviour*, Pearson Education, ninth edn., 2009.
- [8] Ravinder S. Dahiya, Giorgio Metta, Maurizio Valle, and Giulio Sandini, 'Tactile sensing - from humans to humanoids', *IEEE Transactions on Robotics*, **26**(1), 1–20, (2010).
- [9] Kerstin Dautenhahn, Chrystopher L. Nehaniv, Michael L. Walters, Ben Robins, Hatice Kose-Bagci, N. Assif Mirza, and Mike Blow, 'KASPAR - a minimally expressive humanoid robot for human-robot interaction research', *Applied Bionics and Biomechanics*, **6**(3-4), 369–397, (2009).
- [10] Adele Diamond, 'Developmental time course in human infants and infant monkeys, and the neural bases of, inhibitory control in reaching', *Annals of the New York Academy of Sciences*, **608**, 637–676, (1990).
- [11] J. Duysens, A. A. Tax, M. Trippel, and V. Dietz, 'Increased amplitude of cutaneous reflexes during human running as compared to standing', *Brain Research*, **613**(2), 230–238, (1993).
- [12] Andrea d'Avella, Philippe Saltiel, and Emilio Bizzi, 'Combinations of muscle synergies in the construction of natural motor behavior', *Nature Neuroscience*, **6**, 300–308, (2003).
- [13] Paul Fitzpatrick, Giorgio Metta, and Lorenzo Natale, 'Towards long-lived robot genes', *Robotics and Autonomous Systems*, **56**(1), 29–45, (2008).
- [14] Michael Gleicher, 'Comparing constraint-based motion editing methods', *Graphical Models*, **63**, 107–134, (2001).
- [15] Sami Haddadin, Alin Albu-Schäffer, and Gerd Hirzinger, 'Requirements for safe robots: Measurements, analysis and new insights', *International Journal of Robotics Research*, **28**(11-12), 1507–1527, (2009).
- [16] Eric Kugelberg, 'Demonstration of a and c fibre components in the babinski plantar response and the pathological flexion reflex', *Brain*, **71**(3), 304–319, (1948).
- [17] H. Liu, P. Meusel, and G. Hirzinger, 'A tactile sensing system for the dlr three-finger robot hand', in *Proceedings of the International Symposium on Measurement and Control in Robotics*, pp. 91–96, Smolenice Castle, Slovakia, (June 12–16 1995).
- [18] Jun Nakanishi, 'Operational space control: A theoretical and empirical comparison', *International Journal of Robotics Research*, **27**(6), 737–757, (2008).
- [19] Nirut Naksuk, C. S. George Lee, and Shirley Rietdyk, 'Whole-body human-to-humanoid motion transfer', in *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*, pp. 104–109, Tsukuba, Japan, (December 4–7 2005).
- [20] Francesco Nori, Lorenzo Natale, Giulio Sandini, and Giorgio Metta, 'Autonomous learning of 3D reaching in a humanoid robot', in *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1142–1147, San Diego, California, (October 29 - November 2 2007).
- [21] Martin Proetzsch, Tobias Lukscha, and Karsten Berns, 'Development of complex robotic systems using the behavior-based control architecture iB2C', *Robotics and Autonomous Systems*, **58**(1), 46–67, (2010).
- [22] Jens Schouenborg, Han-Rong Weng, and Hans Holmberg, 'Modular organization of spinal nociceptive reflexes: A new hypothesis', *Physiology*, **9**(6), 261–265, (1994).
- [23] Bhagwan T. Shahani and Robert R. Young, 'Human flexor reflexes', *Journal of Neurology, Neurosurgery and Psychiatry*, **35**(5), 616–627, (1971).
- [24] Jean Claude Willer, 'Comparative study of perceived pain and nociceptive flexion reflex in man', *Pain*, **3**(1), 69–80, (1977).

Reducing Speech Collisions by Using an Artificial Subtle Expression in a Decelerated Spoken Dialogue

– Should communication robots respond quickly? –

Kotaro Funakoshi¹ and Kazuki Kobayashi² and Mikio Nakano³
and Takanori Komatsu⁴ and Seiji Yamada⁵

Abstract. We argue that spoken dialogue systems or communication robots do not need to quickly respond verbally as long as they quickly respond non-verbally by showing their internal states by using an artificial subtle expression. This paper describes an experiment whose results support this point. In this experiment, 48 participants engaged in reservation tasks with a spoken dialogue system coupled with an interface robot using a blinking light expression. The expression is designed as an artificial subtle expression to intuitively notify a user about a robot's internal states (such as processing) for the sake of reducing speech collisions as consequences of turn-taking failures due to end-of-turn misdetection. Speech collisions harm smooth speech communication and degrade system usability. Two experimental factors were setup: the blinking light factor (with or without a blinking light) and the reply speed factor (moderate or slow reply speed), resulting in four experimental conditions. The results suggest that the blinking light expression can reduce speech collisions and improve user impression, and surprisingly that users do not care about slow replies.

1 Introduction

An important issue in spoken dialogue systems is the management of turn-taking. Failures of turn-taking due to systems' end-of-turn misdetection cause undesired speech collisions, which harm smooth communication and degrade system usability. Such speech collisions lead users to stop speaking and bring about troubles in spoken dialogue systems [11] because interrupted speech is hard to recognize automatically and dialogue states after collisions are unclear. Deterioration in user impressions is also a problem.

There are two approaches to reducing speech collisions due to end-of-turn misdetection. The first approach is using machine learning techniques to integrate information from multiple sources for accurate end-of-turn detection in early timing. The second approach is to make a long interval after the user's speech signal ends and before the system replies simply because a longer interval means no continued speech comes. As far as the authors know, all the past work takes the first approach (e.g., [2, 8, 12, 14]) because the second approach deteriorates responsiveness of dialogue systems. This choice

is based on the presumption that users prefer a responsive system to less responsive systems. The presumption is true in most cases if the system's performance is at human level. However, if the system's performance is below human level, high responsiveness might not be vital or even be harmful. For instance, a user study reported that the familiarity of a spoken dialogue system with back-channel feedback was inferior to that without feedback due to a small portion of errors even though the overall timing and frequency of feedback were fairly good (but did not come up to human operators) [8]. Technologies are advancing but they are still below the human level. We challenge to the past work that took the first approach.

The second approach is simple and stable against user differences and environmental changes. Moreover, it can afford to employ more powerful but computationally expensive speech processing or to build systems on small devices with limited resources. A concern with this approach is debasement of user experience due to poor responsiveness as stated above. Another issue is speech collisions due to users' following-up utterances such as repetitions because systems' late responses tend to induce such utterances, which are not desired from the viewpoint of systems.

Taking the second approach, we showed the possibility that non-speech feedback using a blinking light based on the concept of artificial subtle expressions (see the next section) can suppress undesired utterances (repetitions) from users [4]. However, our experiment was not with an automatic spoken dialogue system but with a human-operated Wizard-of-Oz system. We showed that our method reduced repetitions (potential causes of collisions) but did not show that our method reduced collisions. We also showed that the blinking light feedback improved user impression but did not show whether the improvement in user impression was enough to compensate for the debasement of user experience due to slow replies.

This paper shows the results of the experiment in which participants engaged in hotel reservation tasks with a spoken dialogue system equipped with an artificial subtle expression-based method proposed in [4], which intuitively notified a user about the system's internal states (such as processing or busy). The results suggest that the method can reduce speech collisions and provide users with a comfortable impression and a modesty impression. The comparisons of user evaluations between systems with a slow reply speed and a moderate reply speed suggest that users of spoken dialogue systems do not care about slow replies. These results indicate that taking the second approach, decelerating spoken dialogues, is not a bad idea.

Section 2 explains artificial subtle expressions. The experiment and results are described in Section 3 and Section 4 respectively. Sec-

¹ Honda Research Institute Japan Co., Ltd., 8-1 Honcho Wako, Saitama 351-0188, Japan, email: funakoshi@jp.honda-ri.com

² Shinshu University, Japan

³ Honda Research Institute Japan Co., Ltd., Japan

⁴ Shinshu University, Japan

⁵ National Institute of Informatics / The Graduate University for Advanced Studies, Japan

tion 5 concludes this paper.

2 Blinking Light as Artificial Subtle Expression

Although human communication is explicitly achieved through verbal utterances, non-verbal facial expressions, gaze, gestures, etc., also play an important role [6, 7]. Such non-verbal communication often influences the accuracy of utterance understanding [15].

Furthermore, researchers have reported that very small changes (called *subtle expressions*) in facial expressions and gestures might influence human communication. We believe that we can utilize such subtle expressions to make humans easily understand a robot's internal state because humans can intuitively understand subtle expressions. Some studies have been done on applying subtle expressions to human-agent interaction [1, 3, 13]. However, since they tried to enable subtle expressions on real faces and with real arms, their implementations were considerably expensive.

Equipping a robot to express its turn-taking intention by using body/eye movements as humans do may reduce speech collisions. Though, such an approach is technically difficult and uneconomical as mentioned above. Using spoken back-channel feedback (such as “well” and “uh”) is another option, however, it is not an easy matter because such back-channel feedback expressions are not arbitrarily used but require appropriate timing and situations to be used [18, 19]. Moreover, such approaches strongly restricts characters of robots, therefore, their applicabilities inevitably go low. Cultural differences also affect them.

In contrast with such human-like approaches, subtle expressions have been studied for artifacts like a robot or PC. Komatsu and Yamada [9] reported that an agent's subtle expression of simple beeping sounds with decreasing/increasing frequency enabled humans to interpret the agent's positive/negative states. Their work indicated the effectiveness of subtle expressions such as varying beeping sounds for a robot or agent. They named such subtle expressions “artificial subtle expressions (ASEs)” and defined them as expressions fulfilling the following four requirements [10].

- **Simple:** ASEs should be implemented on a single modality. It is expected that implementation cost also should be lower.
- **Complementary:** ASEs should only have a complementary role in communication and should not interfere with communication's main protocol. This means that the ASEs themselves do not have any meaning without any communication context.
- **Intuitive:** ASEs should be understood by humans who do not know about the ASEs beforehand.
- **Accurate:** ASEs should convey the specific meanings accurately. Specifically, ASEs should convey the internal states of the artifact like subtle expressions are doing.

Following their work, we proposed the use of a blinking light as a means of artificial subtle expression to intuitively notify a user about a robot's internal states (such as processing or busy) and showed that the blinking light expression potentially can reduce speech collisions, however, indirectly by showing the reduction of users' speech repetitions in a last-and-first game dialogue [4]. This paper verifies the effectiveness of the blinking light expression in a more practical task-oriented dialogue with an automatic spoken dialogue system to directly show this approach reduces collisions.

The use of a blinking light may not be the best way in terms of reducing speech collisions in comparison with other human-like approaches explained above. However, it will be easy and cheap to im-



Figure 1. The LCD monitor and robot with an embedded LED

plement, and applicable to wider conversational agents/devices not limited to robots.

3 Experiment

We conducted an experiment in which 48 participants engaged in hotel reservation tasks with a conversational robot. The dialogue system, robot, blinking light, experimental conditions and method are explained below.

3.1 Spoken Dialogue System

A spoken dialogue system (not a voice command system) that can handle user requests in a hotel reservation domain was built. The system was equipped with an LCD monitor to show reservation information and an interface robot with an LED attached to its chest (see Figure 1). Participants' utterances were recognized by a free automatic speech recognizer Julius⁶, and interpreted by using a domain-ontology-centered language understanding method [5]. The robot's utterances were voiced by a commercial speech synthesizer (NTT-IT FineVoice [17]). The LCD monitor was used to reduce the time in which the system just read up reservation details and to reduce participants' cognitive loads to catch the system's lengthy speech by showing reservation details on the monitor only when the system asked for confirmation.

The system was tuned to uplift its speech understanding performance as much as possible. For example, the language model used by Julius was built using the data collected in the same task domain in advance (4223 utterances / 140-minute speech of 47 users).

Default values were used for the parameters in Julius. Julius output a recognition result (a sequence of recognized words) to the dialogue system at 400 msec after an input speech signal ended, but the dialogue system awaited the next input for a fixed interval (we call this interval *wait interval*, whose length is given as an experimental factor). If the system received an additional input, it awaited the next input for the same interval again. Otherwise, the system concatenated recognition results, made an interpretation, and replied.

3.2 Robot and Blinking Light Expression

We adopted the same robot (WowWee RS-Media) and the same red LED (diameter: 4 mm) as [4]. The blinking pattern was also the same

⁶ <http://julius.sourceforge.jp/>

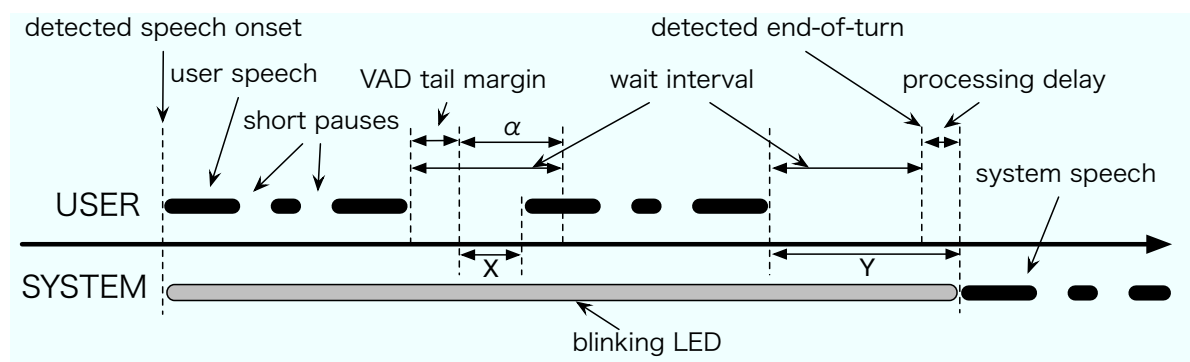


Figure 2. Behavior of the dialogue system along a timeline

(1/30 sec even-intervals). The LED started blinking when a speech signal was detected and stopped when the system started replying.

The basic function of the blinking light expression is similar to hourglass icons used in GUIs. A big difference is that basically GUIs can ignore any input while they are showing those icons, but dialogue systems must accept successive speech while it is blinking an LED. What we intend to do is to suppress only collision-eliciting speech such as repetitions and rephrasings which are negligible but difficult to automatically distinguish from *barge-ins* when speech collisions happen (we call them *follow-ups*). Barge-ins are users' intentional interruptions to systems' speech and dialogue systems must accept them.

3.3 Experimental Conditions and Participants

Two experimental factors each having two levels were setup, that is, the blinking light factor (with or without a blinking light) and the reply speed factor (moderate or slow reply speed), resulting in four experimental conditions A, B, C, and D as below.

Condition A: slow reply speed, **with** a blinking light,

Condition B: slow reply speed, **without** a blinking light,

Condition C: moderate reply speed, **with** a blinking light,

Condition D: moderate reply speed, **without** a blinking light

We prepared a dialogue system for each group, four systems in total and randomly assigned 48 Japanese participants of 18 to 55 years old (mean age 30.9, SD = 10.2) to one of the four conditions, while genders and generations were controlled to be equally distributed among the conditions.

A reply speed depends on a wait interval for which the dialogue system awaits the next input. A user study [16] showed that the best reply speed for a conversational robot was one second. Thus we chose 800 msec as the wait interval for the moderate reply speed because an actual reply speed was the accumulation of the wait interval and a delay for processing a user request, and 800 msec is simply twice the default length (the VAD tail margin) by which the Julius speech recognizer recognizes the end of a speech. For the slow reply speed, we chose 4000 msec as the wait interval. Wait intervals include the VAD tail margin.

Figure 2 shows how the system and the LED works along with user speech. In this figure, a user utters a continuous speech with a rather long pause that is longer than the VAD tail margin but shorter than the

wait interval. If the system detects the end of the user's turn and starts speaking within the interval marked with an 'X', a speech collision would occur. If the user utters a follow-up within the interval marked with a 'Y', a speech collision would occur, too. We try to suppress the former speech collision by decelerating dialogues and the latter by using a blinking light.

3.4 Experimental Method

The experiment was conducted in a room for one participant at one time. Participants entered the room and sat on a chair in front of a desk as shown in Figure 1. They were asked to wear headphones to separate system voice and user voice. All dialogues were videotaped.

The experimenter gave the participants instructions so as to reserve hotel rooms five times by talking with the robot in front of them. All of them were given the same five tasks which require them to reserve several rooms (one to three) at the same time. The meaning of the blinking light expression was not explained to them. After giving the instructions, the experimenter left the participants, and they began tasks when the robot started to talk to them. Each task was limited up to three minutes. The experimenter aborted dialogues if they could not finish a given task in three minutes. They worked on the five tasks without intermission in a fixed order. For illustration, Figure 3 shows the instruction of one of the tasks, and Figure 4 shows an example dialogue observed with this task.

After finishing the tasks, the participants answered a questionnaire about their impressions of the dialogue and robot, and other particular questions. Participants were paid for their participation.

4 Results

4.1 Observed reply speeds

All dialogues were transcribed into texts with timestamps. Averages of observed reply speeds including the system's delays were calculated from those timestamps. They were 4.53 sec (SD = .75) for the slow conditions (A and B) and 1.42 sec (SD = .43) for the moderate conditions (C and D).

4.2 Task completion

The average numbers of completed tasks in the four conditions are shown in Table 1. A two-factor two-level ANOVA did not find any



Figure 3. An example of task instructions

S: Welcome to Hotel Wakamatsu-Kawada. May I help you?
 U: I want to stay from March 10th to 11th.
 S: What kind of room would you like?
 U: One non-smoking twin room and one non-smoking double room.
 S: Are your reservation details correctly shown on the screen?
 U: Yes. No problem.
 S: Your reservation has been accepted. Thank you for using us.

Figure 4. An observed dialogue (translated into English)

significant difference among the four conditions on the numbers of completed tasks.

Table 1. Numbers of completed tasks

reply speed	slow		moderate	
	w/	w/o	w/	w/o
blinking light				
condition label	A	B	C	D
mean	4.00	3.83	3.83	4.33
SD	1.28	1.34	1.27	0.98

4.3 Speech Collisions

As mentioned above, we use the term *follow-ups* to refer to all user utterances such as so called repetitions, corrections, and additions, which are potentially inducing speech collisions. In the experiment in [4], they counted repetitions and it was not difficult in their quite simple last-and-first game dialogues where each user utterance contains only one word. In our experiment, in contrast, it is hard to provide follow-ups with a concrete definition, which is necessary to count them in the dialogue data. Therefore, we do not count follow-ups but directly focus on the number of speech collisions.

We counted speech collisions for which the dialogue system was responsible, that is, the cases where the robot spoke while participants were talking (i.e., end-of-turn misdetections). Of course, there were speech collisions for which participants were responsible, that is, the cases where participants intentionally spoke while the robot

was talking (i.e., barge-ins). These speech collisions were not the targets of this paper, hence they were not included in the counts.

Speech collisions due to participants' back-channel feedbacks were not included, either. We think that it is possible to filter out such feedback because feedback utterances are usually very short and variations are small. On the other hand, as we mentioned above, it is not easy to automatically distinguish negligible speech such as repetitions from barge-ins. We want to suppress only such speech negligible but hard to distinguish from other not negligible speech.

Speech collisions were counted both at the experiment by an experiment supporting staff and after the experiment on the dialogue transcriptions while watching videos by a research supporting staff. The two counting results were checked and merged by one of the authors. The numbers are shown in Table 2 for each participant.

First we performed a two-factor two-level ANOVA on the numbers of collisions. A significant difference between the slow reply speed (A and B) and the moderate reply speed (C and D) was found ($F_{1,44} = 4.06, p < 0.005$). This result confirms that making a long interval after the user's speech ends and before the system replies reduces speech collisions.

The ANOVA test did not find a significant difference on the blinking light factor. This was, however, reasonable because the effect of the blinking light suppressing troublesome follow-ups from users were expected only in the slow reply speed conditions and the numbers of collisions in the slow reply speed conditions were fairly smaller than those in the moderate reply speed conditions.

Hence, we performed a Fisher's exact test (one-side) on the numbers of participants who had speech collisions between the two conditions of the slow reply speed (A and B). The contingency table is shown in Table 3. The test found a significant difference ($p < 0.05$). This result indicates that the blinking light can reduce speech collisions by suppressing users' unnecessary follow-ups in decelerated dialogues.

Table 3. Contingency table on the numbers of participants who had collisions in the slow reply speed conditions

condition	had collisions	had no collision	total
w/ a blinking light (A)	3	9	12
w/o a blinking light (B)	8	4	12

4.4 Impression on the Dialogue

Table 4 shows the results of participants' ratings for the dialogue. The adjective pairs in the table are translated from Japanese words that we used in the questionnaire. The ratings are based on a seven-point Likert scale (1:strong agreement with a negative adjective, 4: neutral, 7: strong agreement with a positive adjective). The highest mean score for each adjective pair is underlined.

We performed a factor analysis (principal factor method) under the varimax rotation and obtained five factors from the scree plot. Table 5 shows the factor loadings under the varimax rotation.

We interpreted the factors according to the adjective pairs. The first factor was named the *likability factor*. The second factor was named the *comfortability factor*. The third factor was named the *interest factor*. The fourth and fifth factors were named *casualness factor* and *comprehensibility factor*, respectively. Their factor scores by using a regression method were calculated and compared among the conditions. Table 6 shows the factor scores for the impression of the

Table 2. Numbers of collisions by each participant

reply speed	condition		participant ID in each condition												total
	blinking light	label	1	2	3	4	5	6	7	8	9	10	11	12	
slow	w/	A	1	2							2				5
	w/o	B	1		3		1		2	1	1	1	1		11
moderate	w/	C	2	5		5	6	3	3	1	14	5	1		45
	w/o	D	7	2		6	6		2		5		2		30

(All the 48 participants are different from each other. Blank slots mean zeros.)

Table 4. Rated adjective pairs for impression of the dialogue

reply speed		slow		moderate	
blinking light		w/	w/o	w/	w/o
condition label		A	B	C	D
positive	negative	mean (SD)	mean (SD)	mean (SD)	mean (SD)
casual	grave	3.42 (1.24)	3.33 (1.07)	3.50 (1.31)	4.00 (1.54)
smooth	rough	3.25 (1.29)	2.58 (1.00)	3.17 (1.59)	3.00 (1.35)
decent	indecent	4.42 (1.00)	4.08 (0.79)	4.08 (1.00)	4.42 (1.08)
exciting	dull	4.58 (0.67)	4.17 (1.03)	3.83 (1.95)	4.17 (1.11)
relaxed	tensional	3.25 (1.06)	3.25 (1.48)	3.42 (1.16)	3.75 (1.48)
easy	uneasy	3.33 (1.37)	2.42 (1.24)	2.75 (1.06)	2.58 (1.08)
warm	cold	3.25 (1.48)	3.25 (0.97)	3.67 (0.98)	3.67 (1.15)
pleasant	unpleasant	4.33 (0.89)	3.67 (1.15)	3.17 (1.03)	4.25 (0.97)
leisurely	hurried	4.75 (1.06)	4.92 (1.51)	4.83 (1.27)	4.67 (0.89)
informal	formal	3.08 (1.16)	2.67 (0.89)	2.67 (1.44)	3.17 (1.03)
light	dark	3.42 (1.00)	3.00 (0.85)	3.17 (0.94)	3.67 (0.78)
clear	confusing	4.00 (1.76)	3.58 (1.51)	4.17 (1.59)	4.50 (0.80)
likable	dislikable	4.42 (0.67)	3.75 (0.87)	4.08 (1.16)	4.33 (0.78)
good	poor	3.75 (1.22)	3.17 (1.11)	3.17 (1.19)	3.58 (1.24)
peaceful	annoying	4.33 (1.37)	3.08 (1.56)	3.83 (1.95)	3.75 (0.87)
interesting	boring	4.58 (1.24)	4.25 (1.14)	4.33 (1.67)	4.67 (1.30)
spirited	dispirited	3.42 (0.67)	2.92 (1.00)	3.00 (0.85)	3.25 (0.87)
settled	unsettled	4.50 (1.31)	3.50 (1.38)	3.83 (1.11)	3.58 (0.79)

Table 5. Results of factor analysis of impression of the dialogue (varimax rotation, factor loading matrix)

factor	item (positive)	factor loadings				
		1	2	3	4	5
1	decent	0.80	0.13	0.23	0.04	0.13
	likable	0.80	0.16	0.15	0.17	-0.23
	warm	0.58	0.01	-0.10	0.34	0.16
	peaceful	0.56	0.52	0.23	-0.25	0.15
	light	0.54	0.23	0.05	0.27	0.28
	informal	0.49	0.37	0.15	0.38	0.04
2	settled	0.35	0.86	0.06	-0.07	0.06
	easy	0.02	0.68	0.05	0.28	0.15
	good	0.13	0.62	0.37	0.42	0.14
	spirited	0.32	0.57	0.36	0.31	-0.04
	smooth	0.07	0.49	0.33	0.26	0.23
3	interesting	0.31	0.03	0.88	0.14	0.02
	exciting	-0.03	0.22	0.59	0.06	0.27
	pleasant	0.10	0.27	0.43	0.15	0.21
4	casual	0.20	0.08	0.14	0.68	-0.10
	relaxed	0.14	0.22	0.11	0.63	-0.05
5	clear	0.04	0.26	0.13	0.03	0.87
	leisurely	0.09	0.00	0.17	-0.13	0.34
contribution (%)		16.12	16.04	10.79	9.91	7.27

dialogue. A two-factor two-level ANOVA found a 10 % level significant main effect for the blinking light in the second factor's score ($F_{3,44} = 3.53, p = .07$). This result suggests that the blinking light provides users with a comfortable impression on the dialogue.

4.5 Impression on the Robot

We analyzed the impression on the robot with the same method as that on the dialogue. Table 7 shows the results of participants' ratings for the robot.

Table 7. Rated adjective pairs for impression of the robot

reply speed		slow		moderate	
blinking light		w/	w/o	w/	w/o
condition label		A	B	C	D
positive	negative	mean (SD)	mean (SD)	mean (SD)	mean (SD)
aggressive	defensive	3.67 (0.78)	3.58 (0.79)	3.92 (1.16)	4.17 (1.19)
innocent	wicked	4.08 (1.08)	3.92 (1.31)	4.17 (0.94)	4.33 (0.78)
respectful	impudent	4.75 (0.87)	4.17 (1.59)	4.42 (0.90)	5.25 (1.36)
accessible	inaccessible	3.08 (1.08)	3.42 (1.31)	3.50 (1.09)	4.00 (0.95)
pretty	provoking	4.17 (0.39)	4.17 (0.58)	4.17 (0.83)	4.08 (0.79)
tolerant	intolerant	4.17 (0.58)	3.75 (0.75)	3.92 (0.79)	4.08 (0.79)
sociable	unsociable	3.75 (1.22)	3.50 (1.17)	3.33 (1.15)	3.92 (0.90)
responsible	irresponsible	4.42 (1.08)	4.17 (1.34)	4.08 (1.24)	4.50 (1.00)
careful	careless	5.42 (1.08)	4.92 (0.90)	4.92 (1.24)	4.50 (1.31)
modest	shameless	4.08 (0.51)	4.08 (0.51)	3.83 (0.72)	3.83 (0.39)
serious	frivolous	5.00 (1.28)	4.67 (1.07)	5.00 (1.35)	4.50 (1.17)
excited	gloom	3.50 (0.52)	3.67 (1.07)	3.33 (0.78)	4.00 (1.04)
regal	servile	4.92 (1.51)	5.00 (1.21)	4.75 (0.97)	4.83 (0.72)
decent	indecent	4.25 (1.06)	3.67 (1.23)	3.58 (0.90)	4.17 (0.72)
discreet	indiscreet	4.67 (1.23)	4.33 (1.07)	3.92 (1.38)	4.33 (0.78)
friendly	unfriendly	3.75 (1.29)	3.33 (1.50)	3.17 (1.40)	4.42 (1.00)
active	inactive	3.83 (1.03)	3.58 (0.67)	3.25 (1.48)	4.17 (0.94)
confident	unconfident	4.50 (1.00)	4.25 (0.87)	4.00 (0.95)	4.67 (0.89)
patient	impatient	4.42 (0.79)	4.92 (1.31)	4.00 (1.04)	4.33 (1.07)
kind	unkind	3.67 (1.30)	4.08 (1.44)	3.33 (1.15)	4.00 (1.04)

Table 8 shows the factor loadings (principal factor method) under the varimax rotation. We interpreted the obtained factors by a factor analysis (principal factor method, varimax rotation) according to the adjective pairs. The first factor was named the *civility factor*. The second factor was named the *seriousness factor*. The third factor was named the *friendliness factor*. The fourth and fifth factors were named *aggressiveness factor* and *modesty factor*, respectively.

Table 9 shows the factor scores by using regression method for the impression of the robot. A two-factor two-level ANOVA among factor scores (regression method) found a 10 % level significant main effect for the slow reply speed in the fifth factor's score ($F_{3,44} = 3.39, p = .07$). This result suggests that the slow reply speed makes the robot look modest.

Table 6. Factor scores for impression of the dialogue

factor	reply speed	blinking light	mean	SD	ANOVA	$F_{3,44}$	p
1. likability	slow	w/o	-0.31	0.27	reply speed	0.45	0.51
		w/	0.13	0.27	blinking light	0.12	0.73
	moderate	w/o	0.21	0.27	reply speed	1.59	0.21
		w/	-0.03	0.27	× blinking light		
2. comfortability	slow	w/o	-0.25	0.26	reply speed	1.19	0.28
		w/	0.53	0.26	blinking light	3.53	0.07
	moderate	w/o	-0.25	0.26	reply speed	1.21	0.28
		w/	-0.04	0.26	× blinking light		
3. interest	slow	w/o	-0.03	0.28	reply speed	0.09	0.77
		w/	0.12	0.28	blinking light	0.00	0.96
	moderate	w/o	0.05	0.28	reply speed	0.35	0.56
		w/	-0.13	0.28	× blinking light		
4. casualness	slow	w/o	-0.08	0.25	reply speed	1.24	0.27
		w/	-0.20	0.25	blinking light	1.37	0.25
	moderate	w/o	0.38	0.25	reply speed	0.49	0.49
		w/	-0.10	0.25	× blinking light		
5. comprehensibility	slow	w/o	-0.18	0.27	reply speed	1.78	0.19
		w/	-0.18	0.27	blinking light	0.17	0.68
	moderate	w/o	0.29	0.27	reply speed	0.20	0.66
		w/	0.06	0.27	× blinking light		

Table 9. Factor scores for impression of the robot

factor	reply speed	blinking light	mean	SD	ANOVA	$F_{3,44}$	p
1. civility	slow	w/o	-0.19	0.28	reply speed	0.38	0.54
		w/	0.02	0.28	blinking light	0.32	0.58
	moderate	w/o	0.35	0.28	reply speed	1.71	0.20
		w/	-0.17	0.28	× blinking light		
2. seriousness	slow	w/o	-0.10	0.28	reply speed	0.30	0.59
		w/	0.26	0.28	blinking light	0.85	0.36
	moderate	w/o	-0.16	0.28	reply speed	0.12	0.73
		w/	0.00	0.28	× blinking light		
3. friendliness	slow	w/o	-0.11	0.26	reply speed	0.13	0.72
		w/	0.01	0.26	blinking light	1.42	0.24
	moderate	w/o	0.42	0.26	reply speed	2.68	0.11
		w/	-0.32	0.26	× blinking light		
4. aggressiveness	slow	w/o	-0.39	0.28	reply speed	1.15	0.29
		w/	0.09	0.28	blinking light	1.45	0.23
	moderate	w/o	0.05	0.28	reply speed	0.25	0.62
		w/	0.25	0.28	× blinking light		
5. modesty	slow	w/o	0.39	0.26	reply speed	3.39	0.07
		w/	0.09	0.26	blinking light	0.18	0.67
	moderate	w/o	-0.28	0.26	reply speed	0.54	0.47
		w/	-0.20	0.26	× blinking light		

Table 8. Results of factor analysis of impression of the robot (varimax rotation, factor loading matrix)

factor	item (positive)	factor loadings				
		1	2	3	4	5
1	respectful	0.85	0.12	-0.02	-0.10	-0.05
	innocent	0.79	0.14	0.07	0.27	0.23
	sociable	0.72	0.02	0.08	-0.18	-0.02
	decent	0.71	0.02	0.32	0.29	0.30
	accessible	0.69	-0.16	0.37	0.02	0.22
	pretty	0.65	-0.23	0.10	0.23	0.03
	kind	0.65	0.24	0.31	-0.02	0.20
	excited	0.63	-0.43	0.33	0.00	-0.01
2	tolerant	0.62	-0.01	0.05	0.19	0.01
	serious	-0.15	0.86	-0.08	-0.18	-0.27
	careful	-0.11	0.65	0.04	-0.01	0.24
	responsible	-0.14	0.61	0.52	-0.10	0.19
	discreet	0.41	0.54	0.21	-0.05	0.45
3	regal	0.17	0.47	0.07	-0.03	-0.12
	friendly	0.54	-0.19	0.71	-0.04	0.14
	active	0.26	0.20	0.43	-0.01	-0.07
4	confident	0.05	0.09	0.41	0.27	-0.22
	patient	0.16	0.32	0.15	0.78	-0.43
5	aggressive	0.34	-0.07	0.15	0.71	0.15
	modest	0.10	-0.03	-0.06	-0.03	0.56
contribution (%)		26.15	13.02	8.55	7.60	6.08

4.6 System Evaluations

The participants evaluated the dialogue system in two measures on a scale from 1 to 7, that is, the convenience of the system and their willingness to use the system. The greater the evaluation value is, the higher the degree of convenience or willingness.

The average convenience scores of the four conditions in Table 10. The average willingness scores of the four conditions are shown in Table 11. The largest mean values among the four conditions are shown in bold letters. ANOVAs did not find any significant difference among the four conditions both for the two measures.

Table 10. Convenience scores

reply speed	slow		moderate	
blinking light	w/	w/o	w/	w/o
condition label	A	B	C	D
mean	3.50	3.17	3.17	3.92
SD	2.02	1.53	1.47	1.62

Table 11. Willingness scores

reply speed	slow		moderate	
blinking light	w/	w/o	w/	w/o
condition label	A	B	C	D
mean	3.58	2.58	2.83	3.42
SD	1.73	1.31	1.34	1.56

4.7 Discussion on User Preference

The analysis of the questionnaire suggests that the blinking light expression gives users a comfortable impression on the dialogue. In addition, comparing mean scores of adjective pairs and system evaluations between the conditions A and B, a tendency giving higher scores to A is observed. These support the effectiveness of the blinking light expression. However, comparing scores between C and D, a tendency giving higher scores to D is observed. This intimates that the blinking light expression includes some negative effects, too. We are planning to investigate the blinking pattern from this aspect in future work.

The analysis also suggests that the slow reply speed gives users a modest impression on the interface robot. Meanwhile, no negative impression with a statistical significance is found on the slow reply speed. Although no statistically significant difference is found between the four conditions (that is, differences are small), numbers of completed tasks shown in Table 1 and convenience scores shown in Table 10 strongly correlate. However, users' willingness to use the systems, which is the most important measure for systems, is inverted between condition A and D as shown in Table 11. Convenience will be primarily dominated by to what degree a user's purpose (reserving rooms) is achieved, thus, it is reasonable that convenience scores correlate with numbers of completed tasks. On the other hand, willingness will be dominated by not only practical usefulness but also overall usability. Therefore, we can interpret that the improvements of impressions and reduction of aversive speech collisions let condition A have the highest score for willingness. These results indicate that decelerating spoken dialogues is not a bad idea in contradiction to the common design policy in human computer interfaces (HCIs), and they suggest to exploit merits brought about by decelerating interactions rather than pursuing quickly responding human-like systems.

Our finding contradicts not only the common design policy in HCIs but also the design policy in human robot interaction found by Shiwa et al. [16], that is, *the best response timing of a communication robot is at one second*. We think this contradiction is superficial and is ascribable to the following four major differences between their study and our study.

- They adopted a within-subjects experimental design while we adopted a between-subjects design. A within-subjects design makes subjects do relative evaluations and tends to emphasize differences.
- Their question was specific in terms of response timing, that is, "Answer your preference about the timing of the system response." Our questions were overall ratings of the system such as convenience.
- They assumed a perfect machine (Wizard-of-Oz experiment). Our system was elaborately crafted but still far from perfect.
- Our system quickly returns non-verbal responses even if verbal responses are delayed.

From these differences, we hypothesize that response timing has no significant impact on the usability of dialogue systems in an absolute and holistic context at least in the current state of the art spoken dialogue technology, even though users prefer a system which responds quickly to a system which responds slowly when they compare them with each other directly, given an explicit comparison metric of response timing appropriateness with perfect machines.

5 Conclusion and Future work

This paper showed the results of the experiment in which forty-eight participants engaged in hotel reservation tasks with a spoken dialogue system coupled with an interface robot equipped with an LED. The experiment aimed to investigate the influences of decelerating dialogues and the effects of a blinking light expression devised as an artificial subtle expression (ASE) on reducing speech collisions and compensating the deterioration of responsiveness due to decelerating dialogues. Decelerating dialogues is the simplest way to build stable spoken dialogue systems with speech collisions reduced, and affords to employ more powerful but computationally expensive speech processing or to build systems on small devices with limited resources. The blinking light expression as an ASE is quite simple, thus it is easy and cheap to implement, and applicable to wider conversational agents/devices not limited to robots.

The two experimental factors used in the experiment were the blinking light factor (with or without a blinking light) and the reply speed factor (moderate or slow reply speed). The analysis showed that speech collisions were reduced by slowing the reply speed, and they were reduced further by using the blinking light expression with statistical significances. The analysis of a questionnaire suggested that the blinking light expression gave participants a comfortable impression, and surprisingly that users did not care about slow replies. In addition, although a statistical significance was not found, the system with the slow reply speed and a blinking light obtained the highest score for users' willingness to use the system, which is the most important measure for systems.

While our method using an LED can apply to any other interfaces on wearable/handheld devices, vehicles, whatever, it is difficult to directly apply it to call-centers (i.e., telephone interfaces), which occupy a big portion of the deployed spoken dialogue systems pie. However, the underlying framework, that is, "decelerating spoken dialog with an artificial subtle expression", will be applicable even to telephone interfaces by using an auditory artificial subtle expression which is to be explored in future work.

So far, our conclusion is that spoken dialogue systems or communication robots do not need to quickly respond verbally as long as they quickly respond non-verbally by showing their internal states by using an artificial subtle expression, while many researchers try to make them verbally respond as fast as possible. Decelerating dialogue has many practical advantages as stated above. However, through the experiment, we suspect that this conclusion is not valid in some specific cases. That is, we think in some situations users are troubled by slow verbal responses primordially, and those situations are such as when users simply reply to systems' yes/no questions or greetings. Our hypothesis is that users expect quick verbal responses (and hate slow verbal responses) only when users expect that it is not difficult for systems to understand their responses or to decide next actions. If this hypothesis is valid, and if users' expectations can be easily estimated not from unreliable information such as speech recognition results, intonations, facial motions, etc but from solid information such as the speech act type of a system's preceding utterance, we will be able to immediately achieve better speech communication systems which respond quickly only when needed. We will seek this direction in future work.

REFERENCES

- [1] C. Bartneck and J. Reichenbach, 'Subtle emotional expressions of synthetic characters', *International Journal of Human-Computer Studies*, **62**(2), 179–192, (2005).
- [2] L. Bell, J. Boye, and J. Gustafson, 'Real-time handling of fragmented utterances', in *Proc. NAACL-2001 workshop on Adaptation in Dialogue Systems*, (2001).
- [3] S. Brave, C. Nass, and K. Hutchinson, 'Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent', *International Journal of Human-Computer Studies*, **62**(2), 161–178, (2005).
- [4] Kotaro Funakoshi, Kazuki Kobayashi, Mikio Nakano, Seiji Yamada, Yasuhiko Kitamura, and Hiroshi Tsujino, 'Smoothing human-robot speech interactions by using a blinking-light as subtle expression', in *Proc. 2008 International Conference on Multimodal Interfaces (ICMI)*, pp. 293–296, (2008).
- [5] Kotaro Funakoshi, Mikio Nakano, Yuji Hasegawa, and Hiroshi Tsujino, 'Semantic interpretation supplementarily using syntactic analysis', in *Proc. 2007 International Conference /Recent Advances in Natural Language Processing/ (RANLP)*, Borovets, Bulgaria, (Sep. 2007).
- [6] A. Kendon, 'Some functions of gaze direction in social interaction', *Acta Psychologica*, **26**, 1–47, (1967).
- [7] A. Kendon, 'Do gestures communicate?', *A Review. Research in Language and Social Interaction*, **27**(3), 175–200, (1994).
- [8] Norihide Kitaoka, Masashi Takeuchi, Ryota Nishimura, and Seichi Nakagawa, 'Response timing detection using prosodic and linguistic information for human-friendly spoken dialog systems', *Journal of The Japanese Society for Artificial Intelligence*, **20**(3), 220–228, (2005).
- [9] T. Komatsu and S. Yamada, 'How do robotic agents' appearances affect people's interpretations of the agents' attitudes?', in *Proc. 25th ACM International Conference on Human Factors in Computing Systems (CHI)*, pp. 2519–2524, (2007).
- [10] Takanori Komatsu, Seiji Yamada, Kazuki Kobayashi, Kotaro Funakoshi, and Mikio Nakano, 'Artificial subtle expressions: Intuitive notification methodology of artifacts', in *Proc. 28th ACM International Conference on Human Factors in Computing Systems (CHI)*, p. (to be appeared), (2010).
- [11] Mikio Nakano, Yuka Nagano, Kotaro Funakoshi, Toshihiko Ito, Kenji Araki, Yuji Hasegawa, and Hiroshi Tsujino, 'Analysis of user reactions to turn-taking failures in spoken dialogue systems', in *Proc. 8th SIGdial Workshop on Discourse and Dialogue*, (2007).
- [12] Tomoko Ohsuga, Masafumi Nishida, Yasuo Horiuchi, and Akira Ichikawa, 'Investigation of the relationship between turn-taking and prosodic features in spontaneous dialogue', in *Proc. 9th European Conference on Speech Communication and Technology*, pp. 33–36, (2005).
- [13] H. Prendinger, J. Mori, and M. Ishizuka, 'Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game', *International Journal of Human-Computer Studies*, **62**(2), 231–245, (2005).
- [14] A. Raux and M. Eskenazi, 'Optimizing endpointing thresholds using dialogue features in a spoken dialogue system', in *Proc. 9th SIGdial Workshop on Discourse and Dialogue*, (2008).
- [15] W. Rogers, 'The contribution of kinesic illustrators towards the comprehension of verbal behavior within utterances', *Human Communication Research*, **5**, 54–62, (1978).
- [16] Toshiyuki Shiwa, Takayuki Kanda, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita, 'How quickly should communication robots respond?', in *Proc. 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 153–160, (2008).
- [17] S. Takano, K. Tanaka, H. Mizuno, M. Abe, and S. Nakajima, 'A Japanese TTS system based on multi-form units and a speech modification algorithm with harmonics reconstruction', *IEEE Transactions on Speech and Audio Processing*, **9**(1), 3–10, (2001).
- [18] Wataru Tsukahara and Nigel Ward, 'Evaluating responsiveness in spoken dialog systems', in *Proc. International Conference on Spoken Language Processing*, (2000).
- [19] Nigel Ward, 'On the expressive competencies needed for responsive systems', in *Proc. the CHI2003 workshop on Subtle Expressivity for Characters and Robots*, (2003).

Evaluation of Robot Body Movements Supporting Communication

Helge Hüttenrauch¹ and Cristian Bogdan¹ and Anders Green¹ and Kerstin Severinsson Eklundh¹
Dominik Ertl² and Jürgen Falb² and Hermann Kaindl² and Michael Göller³

Abstract. In designing socially interactive robots we have focused on robot movement and its role in multi-modal human-robot communication. In this paper we describe a user-centred design and evaluation process, investigating the idea of using speed and orientation adjustments as design elements in human-robot interaction. The scenario studied includes a robotic shopping trolley that offers products to the user while both are moving in a supermarket-like environment. Our results show that if the robot slows down while making such offers, users are more prone to react upon them. However, in an early pre-study, performed only with a robot mock-up, we observed that users tended not to notice the robot's slow-down movements while offers are made, even if these movements were shown several times to them during a video-based debriefing. This phenomenon, that users react implicitly on the robot's movements without being consciously aware of them, was confirmed during an experimental study with a fully integrated robot prototype. We discuss our results by reflecting on human-robot interaction design methods, and we propose implications from the lessons learnt in the study of the design of robot behaviours.

1 Introduction

Service robots that carry out tasks in homes, for instance hovering or lawn mowing, often use only uni-modal interfaces such as push buttons for input, showing the effects on small liquid-crystal displays. These types of robots are functionally designed for practical purposes [4] and therefore lack actuators, like arms, heads or a face, that could be used to provide additional multimodal output to enhance communication. When today's service robots move, these movements are primarily intended to bring about physical actions that support the service task of the robot. The research on spatially situated robots is yet in its infancy, and has focused on themes like approach directions, engagement space, comfort zones and appropriate distances between robot and user, etc. [19, 21, 23, 24, 27]. Our research is focused on how active engagement using the body in the communication between user and robot can be designed and evaluated [7, 13]. The study presented in this paper challenges the current understanding of human-robot communication by extending the communication design to include full-body gesture intended to influence users as they are

planning and performing an everyday activity: shopping in a supermarket. Our approach is to integrate the output of full-body gestures into the interaction design of a human-robot interface. Here the immediate goal is to increase the communicative capability of the specific type of robots we are investigating: Robotic Shopping Trolleys.

The rest of this paper is structured as follows; initially we introduce some concepts relevant for communication with full body gestures between humans and robots. Then we describe and report the findings from two user studies: an exploratory design study and an experiment. The exploratory design study was carried out using an enacted robot prototype and was used to prepare for an experimental study for which a real robot was used.

2 Background and related work

Arguably human body movements may have communicative effects, and thus be interpreted as gestures. When people are turning towards something, change their position or speed of walking during conversation, they convey meaning. In some cases this might be conventionalised, such as turning towards some object to carry out a deictic reference. Such body gestures are interleaved with ordinary movements – without symbolic meaning or lexicalised content, like walking or standing.

Mobile robots are also able to perform gestures using their bodies. It is therefore plausible to analyse them using the same type of concepts as used in human-human communication. This means that robotic gestures, like human body gestures, may function like utterances on their own, used in alternation or conjunct with linguistic information [15]. Another dimension of human movement is the ability to adapt to each other's movements, something which also has been shown to be relevant for human-robot interaction [27].

Central to the notion of situated interaction is the notion of shared space, the so called o-space, or the transactional space which can be described as an area, normally located in front of the interlocutors, where the interaction is conducted [1]. Full body gestures characterised as body moves can provide meta-discursive information in co-occurrence with information from other modalities or by themselves [5]. This pragmatic dimension aggregates commitments, or social obligations [22] by the participants that are bodily involved in the activity at hand.

Human body gestures fall within a type of gesture that are less conventional than for instance emblems that provide conventionalised ways of signing commands like “stop” or extending the index finger to display that the gesture conveys some type of deictic in-

¹ KTH School of Computer Science, Royal Institute of Technology, 100 44 Stockholm, Sweden, email: {hehu, cristi, green, kse}@csc.kth.se

² Vienna University of Technology, Institute of Computer Technology, Vienna, Austria, email: {ertl, falb, kaindl}@ict.tuwien.ac.at

³ Research Center for Information Technology (FZI), Karlsruhe, Germany email: {goeller}@fzi.de

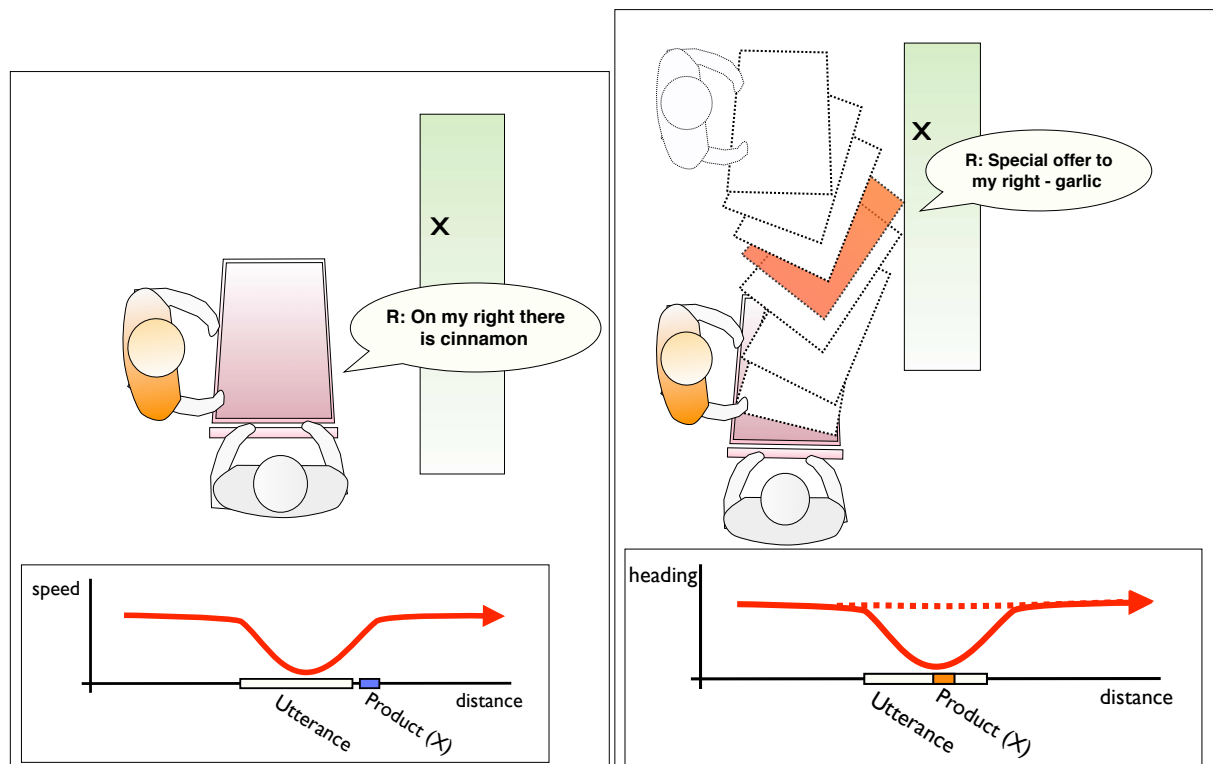


Figure 1. I: Slowing down: The figure shows how the robot slows down while providing the speech output (left).
II: Pointing: Trolley turns its body to point on the object while maintaining its trajectory (right).

formation. Instead, body movements convey additional information to an utterance given in some other modality, such as focusing on something by leaning into a work space [5].

2.1 Social spatial behaviour in Human-Robot Interaction

There are several research approaches for human-human interaction regarding social spatial behaviour that are relevant to study human-robot interaction, for instance Hall's work on interpersonal distances [10] has been used as a backdrop for studies on approach direction and comfort distances [23, 24]. Concepts drawn from theories on human communication such as work space [1] have been used to model situated interaction between humans and robots. One example of this is the tutoring system described by Sidner et al. [19], where the notion of a work space or engagement space is used as a fundamental configuration for the unfolding interaction between human and (robotic) tutor. The robot in this particular system was used as a pointing device, to guide the attention towards objects and to perform meta-conversational communicative actions, i.e. conversational feedback. Another approach that involves social models of spatiality is, for instance, the receptionist robot described by Michalowski et al [18], that estimated the degree of engagement based on spatial proximity. Tasaki et al. [21] describe a museum robot that uses a "Friendliness Space Map" to model the attitudes of visitors towards the exhibition. In their system, a humanoid robot in a fixed position, they varied the parameters for what type of communication the robot should engage in, based on a model of spatial proximity and po-

sitioning. Furthermore Hüttenrauch et al. [11] found that humans adjusted their spatial distance in relation to their imminent task.

In the approaches listed above body movements or bodily configurations are used by robots to spatially influence the interaction partner in an implicit manner rather than attempting to achieve some specific conventional communicative effect. The remainder of this paper will focus on robotic full body gestures that are carried out with the intent to evoke a communicative effect.

3 Methodological approach

As methodological approach we have chosen user-centred design adapted to human-robot interaction [8]. Preceding the work described in this paper we went out to study shopping using ethnographically inspired methods, observing shopping patterns and how shopping trolleys were used. We have also adapted video prototype techniques [17] to the human-robot interaction situation, as the interaction enactment which comes from creating video-prototypes captures the dynamic and embodiment aspects of interaction. This would otherwise only be attainable using more complex setups involving full-scale robot hardware using complex control mechanisms or hi-fi-simulation [2, 9]. Using video prototypes we were also able to invite users to reflect on their interaction with the robot [20]. In the final phase we have carried out a controlled experiment to evaluate our human-robot interaction design which is the main focus of this paper. In the following we describe two steps in this evaluation process:

- An initial, qualitative design study, focusing on user participation and reflection over some full body robot gestures.

- An experimental study of one specific robotic body gesture.

4 Initial exploratory design study

To explore the design space for robotic body gestures, we carried out an exploratory study. In the study we created a scenario and robot design involving robotic body movements that we enacted using an overt version of the Wizard-of-Oz simulation technique. Normally when performing a such a study, the operator is hidden and a real(istic) system is simulated without telling the user what is going on behind the scenes [9]. In our version the robot movements were controlled by a physically fully present robot navigator who operated a regular shopping trolley. We enacted the scenario in our laboratory, with props that would allow participants and experiment leaders to engage in a role-play that contain the conceptual components of shopping in a supermarket with a robotic trolley.

We invited 8 employees from the research institute where the study was carried out to participate in the study. The participants came from the same department, but they were not involved in the robot project. They all had experience of shopping, but not from any robotic shopping trolleys. Each participant was told about the general purpose of the study: to investigate the design of a supermarket trolley. Furthermore we informed them about the way the robot would be driven by a human operator, but said nothing about the robotic body gestures that they were going to experience. The instruction to the participants was that they would use a robotic shopping trolley to pick the ingredients needed to bake a cake.

The research questions we addressed in this design study were:

- Do robotic body gestures have a communicative effect?
- Do such gestures change the shopping behaviour of the users?

The robot design we wanted to evaluate involved two types of full body gestures:

- Slowing down: the robot slows down to allow the user to focus her attention to the immediate spatial context while uttering a phrase (i.e., “On my right there is cinnamon”, uttered right before the robot passed the product). (Fig. 1, I).
- Full body pointing: this gesture is a referential gesture that complements a verbal utterance, i.e., “Special offer to my right – sugar!” (Fig. 1, II).

To evaluate the appropriateness of these body movements we have analysed the scenarios in two ways. First we looked at what the users did when the robot displayed body gestures in the video recordings trying to see the effects of the gesture. This allowed us to judge to what extent the gestures influenced the interaction. We analysed where the participants looked, if they changed body direction and if they manipulated or mentioned any of the products suggested by the robot. Taken together such behaviour of the participants was used to evaluate the communicative success of the robotic gestures. Secondly we interviewed participants after the interaction sessions. During the interviews the participants were shown video recordings showing parts of the interaction and were asked to reflect on it.

Some reflections concerned the shopping behaviour in general. Some participants assumed the robot to have some common sense knowledge about their overall shopping plan: “It [(garlic)] does not really fit my cake”, or, “Why did special offer concern garlic, it does

not make sense for baking?”. The shopping list serves a manifest representation of the users’ shopping plans, rendering comments regarding the appropriateness of having the robot reminding them about items: “[It was] not on the [shopping] list”; “[I am] not usually buying things that are on offer”, and “If there was an offer every step it would have been annoying”.

When asked about the two gestures, users appeared to notice and comment on the referential (turning) gesture to a larger extent than the slowing-down gesture. However, even if they seem to notice the robot’s turning gesture, it is not identified as a referential gesture straightforwardly by the participants, indicated by reflections like: “Aha, and it didn’t go straight, it went a bit... to the right?”. In general our interpretation was that the strength of the confidence they gave their own observations of the gesture were quite weak: “Well, on the movie it turned slightly towards the [product]. I don’t know if that was supposed to happen?”

More important, and the primary motivation for selecting the slowing-down gesture to evaluate in the larger experimental study, was that it seemed this gesture was not interpreted as a communicative act in itself. This supports our understanding that this gesture works as a complementary gesture. It seems that when the trolley slows down, users pay more attention to the spoken prompt, but that the slowing-down did not appear as a discernible gesture. In the reflection sessions some participants only noticed the slowing down when watching the sequence for the second (or even third time). The main contribution of the initial study was that it allowed us to formulate a hypothesis to test in the quantitative study, described in the following.

5 Experimental study

In the main study we present in this paper we have narrowed our focus of the research questions somewhat. Instead of two robotic gestures, we are now investigating only one, namely slowing-down. The research question is:

- Does the robotic body gesture *slowing down* (accompanied by spoken output) positively affect the amount of products taken by participants during a shopping round?

To test this we created a “supermarket aisle” in our laboratory. In a large room we placed a long shelf to form a 10 meter aisle (Figure 3). In the shelf we put about 10 different products that the robot had stored in its product database and could be scanned with the bar code scanner. The position of each product was also entered in the database. With this supermarket aisle we intended to give the user a sense of being in a real shop. Along the aisle the robot had its starting point and the stored location of the three products in the shopping list for which we had pre-programmed the robot to utter a spoken prompt and perform the slowing down gesture. The products were a soft drink (“Cola”), a bottle of bathroom cleaner (“Cleaner”) and a bottle of wine (“Wine”). At the end of the aisle we had placed the last product on the list: a bottle of shampoo (“Shampoo”).

The robot we used was the Interactive Behavior Operated Trolley (InBOT, see Figure 4). Its obstacle detection uses sensor information acquired from two 270 degree planar laser range finders. It uses an RFID transponder to detect barriers placed on the floor and stores these areas using topological relations. It uses odometry to keep track of its traveled distance when it navigates. The robot’s mecanum drive enables it to perform fully holonomic

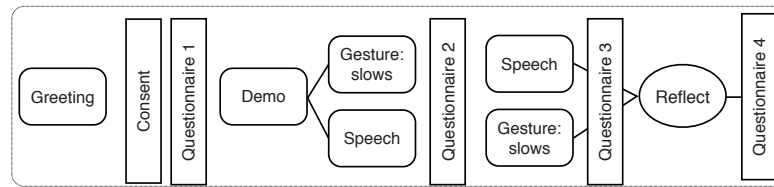


Figure 2. Study design

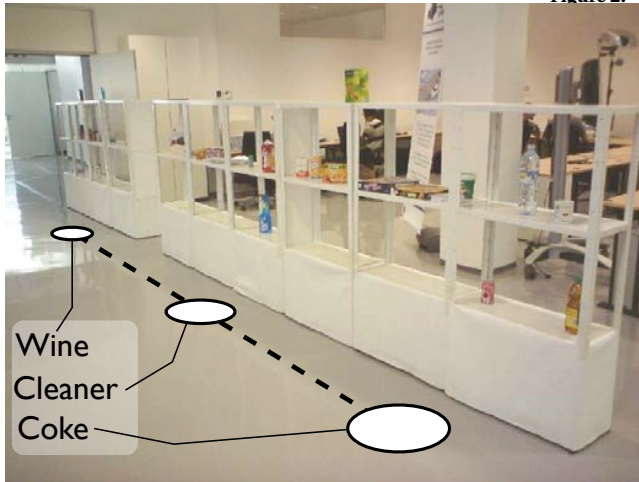


Figure 3. Shop corridor with shelf and products as it were built up in the laboratory used in the study.

movements. For more details on the architecture and control of the robot see [6].

We recruited 20 participants from the University of Karlsruhe campus, giving us a rather homogeneous group of participants both in gender (16 male, 4 female) and in age (average 24.1). Upon initial contact potential trial participants were invited to participate in a study to test “a prototype of a motorised shopping trolley”, avoiding the word “robot” altogether. For their participation in the trial, participants received a small token of our appreciation in the value of about 5 Euro after participating.

Upon arrival the participants were greeted and informed about their task; shopping for multiple products using a prototype “motorised shopping trolley”. They were also informed about their rights as participants and what type of data we recorded. After this they signed a consent form. After this the participants were asked to fill out a questionnaire about socio-demographic data and their shopping habits. Participants then were given a short introduction on how to operate the graphical user interface (GUI), see Figure 5. They were then given the opportunity to enter a shopping list into the system on the screen. They were then introduced to the full system including movements of the robot. The participants were asked to initiate and select a task on the graphical user interface which made the robot start and move to a product placed along the aisle.

The initial familiarisation with the trolley and its interface components was done by the participant but with the assistance of the experiment leader. Each participant was also explicitly asked to demonstrate their ability to use the components that they had been told about by the experiment leader. In this step they were informed about and got to try out the following functionalities of the robot:



Figure 4. The FZI-InBOT used in this study

- driving the trolley to a product,
- stopping the robot on its way by using the “Stop”-button on the touch-screen and/or the scanner, and finally,
- using the bar-code scanner to scan the product on the shopping list once the trolley had arrived at the product’s position.

After this demonstration, the experiment procedure started. We gave the participants a print-out of a four item shopping list consisting of “Cola”, “Cleaner”, “Wine”, and “Shampoo”. The list was entered so that the robot had “Shampoo” as the first target location, meaning that the robot, when it passed all the other products on the list could, suggest them to the participant.

The setup of the study was identical for all participants up to this point. Then the robot used one of these strategies to influence the user during their shopping:

- During the mission the robot displays the full-body gesture “slowing down”. The robot slows down while uttering “Here is (product) from your shopping list”. This happens before the robot passes three of the products on the shopping list.
- The other run was with the same spoken utterances while passing the three products on the list, but without a full body gesture.

A schematic overview of the setup and study design can be seen in Figure 2.

After a participant had used the robot with one type of behaviour in the first run, we shifted behaviour so that the robot was used with the other behaviour in the second run.

After the first run participants were asked to fill out a questionnaire, asking about their impression of the shopping they just had performed. When they had answered the questionnaire, they used the robot with the other robot behaviour activated.

After completing the two different “shopping rounds”, we asked the participants to run the robot again. In this run the full body, slowing down body gesture was displayed again (identical to the one in the trial round). This time the participants were asked to verbalise their experience using the think-aloud method [3]. When the robot carried out its slowing down gestures, the test leader asked them to reflect in detail of the robot’s behaviour. After this reflective shopping round, the participants were asked to answer a questionnaire about the general idea of a “motorised shopping trolley”. Finally they were debriefed and handed the small gift.

6 Results

The results of the study are based on the analysis of two data sources: the questionnaires and the video recordings of the different runs. The videos were annotated according to a pre-determined metric from two annotators; for reasons of brevity only the most relevant aspects are presented here.

The statistical analysis based upon observations in video includes 15 participants that completed both test runs. For each participant we have measured the number of products the participant has taken. A participant could have taken between 0 and 3 products (Cola, Cleaner, and Wine). In Figure 6 the X-axis displays how many products the participants took. The frequencies of the products taken in the first run is shown in this figure. The right-hand side bar in each column shows the condition without gesture display and the left-hand side shows the condition with the full body gesture.

Participants using the robot that performed the full body gesture (slowing down) took more products than participants using the robot without it in the first run. The number of taken products correlates with the behaviour of the robot at 0.821 and is highly significant at the 0.01 level⁴. In the second test run, participants using the robot with the gesture behaviour took approximately the same number of products as participants using the robot without the gesture behaviour activated.



Figure 6. Frequencies of products taken in the first test run.

Based on the data of both runs we calculated a two-factorial ANOVA (analysis of variance) with the robot’s behaviour and the run as independent fixed factors.

The ANOVA results in Table 1 show that participants take significantly ($p=0.022$) more products if the robot uses the full body gesture. Further, it shows that in a second run participants have learnt how to use the robot trolley and take significantly ($p=0.003$) more products than in the first run. When we look at both variables (run and body gesture enabled or not), we see that participants need to learn less between the runs if body gestures are enabled and that they initially take more products with body gestures enabled which drops to the same amount in further runs. Thus, there is a significant ($p=0.001$) interaction effect between these two variables.

Looking at the second run, we see a high learning effect that superimposes all other effects. Thus, we cannot say if an effect of the robot’s behaviour is present starting with the second run.

While the measurement of products taken during the experimental runs are convincingly pointing towards the full body gesture of slowing down as having a clear and measurable effect, the *perception of this gesture* by users seems much more questionable.

One observation we made when analysing the experiments is that very few users reported that they had perceived some kind of gesture performed by the robot. Only 4 out of 20 participants clearly stated that they noticed the full body gesture of the slowing down trolley to indicate products. This is an remarkable outcome given that participants had the chance to experience two, to their nature, quite different trolley driving runs and thus handling experiences between the two runs performed right after one another. Most participants plainly stated that they had not spotted a difference between the runs, some commented on their *performance* or *learning experience*, or talked about other issues, e.g., the spoken utterance of the trolley, or the actual placement of the products.

Using the same approach as in the initial exploratory design study we let users watch the video of themselves interacting with

⁴ The correlation value represents the Pearson correlation of a point-biserial correlation.



Figure 5. The Graphical User Interface used in the robot.

Table 1. Test of between-participants effects.

Dependent Variable:products

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	26,182 ^a	3	8,727	11,059	,000
Intercept	112,634	1	112,634	142,728	,000
behavior	4,715	1	4,715	5,975	,022
run	8,715	1	8,715	11,043	,003
behavior * run	11,834	1	11,834	14,996	,001
Error	20,518	26	,789		
Total	155,000	30			
Corrected Total	46,700	29			

a. R Squared = ,561 (Adjusted R Squared = ,510)

the robot in order to reflect on the user experience, and especially the full-body gesture of slowing-down.

The result of the video-taped and analysed reflective runs is interesting: Users were (again) led to the three products with the robot performing the slowing down gesture. But after every product the experiment leader and the participant took some time to go through a couple of standardised questions and discussed them together: “What did the trolley just do?”, “What did the speech interface utter?”, and “What did you do?”, were questions to let participants express in their own words their experience they just had with the trolley.

Filtered for the full body gesture of slowing down this method could during this third run show more users being (or becoming) aware of this gesture as compared to the questionnaire data collected, however, from the 20 reflective runs performed, only 7 participants clearly expressed that they noticed the slowing down. Another 8 participants never realised even during the reflective runs that the robot trolley is slowing down. The remaining 5 cases are difficult to rate clearly, as participants for example hinted in their used expressions that they took advantage of the slowing down of the trolley, but never expressed this in clear terms in using words such as “braking”, “slowing down”, “reducing speed”, or similar expressions to describe the robot’s behaviour.

7 Discussion

7.1 Motivation for robotic body gestures

Throughout our work of designing movements as full body gestures that would properly accompany other robot communication, we have reflected on the details of how designed movement patterns can make robots more socially interactive.

When a human makes a deictic reference, a spatial relation needs to be established in order for the conversation partner to understand the reference made. We assume that the human behaviour corresponding to what the robot does in the slowing down phase of the study involves a deictic component of interaction that is both temporal and indicative. If a human would not accompany an utterance containing a deictic reference, with adjustments of their movement, i.e., keep the same walking speed and body orientation while referencing something in the surroundings, they may be perceived as non-collaborative (or even rude).

Some of our observations in the study can be better understood using this assumption as a backdrop. Some users in the condition without slowing down body gesture qualified the robot behaviour with negative characterisations such as “unfair”, “I felt like a dog on a leash”, or “I had to pay attention – stressful”.

Users who first received the condition without the slowing down gesture were more “on their watch” in their second run and stopped the robot more than the users who got the presumably more usable slowing down condition first. This suggests that users adapt to overcome flaws in the design.

Perception and execution of gesture in human-human conversation is sometimes carried out in what can be understood as the periphery of awareness, i.e., human language users do not (need to) pay attention to their specific body postures and gestures during communication. We perceive and adapt to body posture and gesture with little effort. In conversation we expect this adaptation to be performed by our conversation partners. It is part of the fabric of our social expectations and we rarely pay attention to them. We do however notice when they are missing.

This can explain what we regard as an interesting finding, which came out both of our exploratory design study and out of our experiment: few of our users noticed the robot’s movement adaptation, even if it was shown to them on video in the pilot study with the mock-up robot. This was also the case in the experimental study, where very few users noticed the robot’s full body gesture during the reconstruction runs.

This phenomenon may be related to information that is in the periphery, i.e., information which is being absorbed and understood without being in focus of attention [25]. Designing movement is then akin to designing calm technology [26], like ambient information displays [12] or other approaches for providing information without distracting the user. Our understanding of robotic body gestures is in accordance with [5], who regards meta-discursive actions crucial for spatially situated interactions. Robotic body gestures should therefore be considered carefully when designing human-robot communication.

7.2 Beyond making the first move

We are currently redesigning the movements and robotic body gestures of our robot and we are considering the issue of movement adaptations, and how they can be performed in a more structured way. This goes beyond the movement(s) accompanying a first, initiating utterance, taking into account the response. The theoretical framework we are using is based on the notion of adjacency pairs [16]. In such a model, each communicative initiative has a preferred response and deviation from this response leads to a state of miscommunication.

Applied to our context we understand the slowing down of the robot as an initiating move for which the preferred response is to slow down and show interest in the indicated product. If the user walks past a robot that slows down to show something, or even attempts to push the robot to move on, can be interpreted as active lack of interest in the product communication and negative response to the communicative action.

In our view such bodily ways of responding to communicative actions need to be taken into account when designing communication for robots. This means that body movements that indicate refusal of offers should lead to adaptation of the robot behaviour, possibly by decreasing the number of offers. Conversely acceptance of the robot offer could lead to an increase in offers made by the platform.

As a consequence of understanding robotic body gestures as the first part of an adjacency pair, we see the need to interpret and detect responses from human users. This requires machine perception that goes beyond the technical scope of this paper.

8 Conclusion

The fine spatial interplay between humans and robots is something that we are only beginning to understand. In this study we have focused our interest in one type of robot gesture, namely slowing down when passing a product. This can be contrasted with another approach taken by Kanda et al. [14] who used a conversational approach to achieve what they call an “Advertisement effect” triggered by verbal recommendations by the robot. This meant that the participants reported that they had been influenced in their shopping behaviour (for instance, buying ice cream because the robot suggested it). It should be noted that the anthropomorphic design of the robot used by Kanda et al. [14] allowed for a

larger repertoire of gestures than the robot system we describe in this paper, involving eye gaze, pointing gestures using an arm manipulator. The automated robot trolley lacks such capacity, but our results show that accompanying spoken utterances with robotic body gestures lead to changes in the shopping behaviour of humans, i.e., there is an increase in the number of products taken during a shopping round. It appears that it is not necessary to employ a full-blown humanoid robot to influence the shopping behaviour of people.

We have yet to establish an explanation to why the full body gesture has an effect on the shopping behaviour. The explanation we suggest is that the slowing down gesture allows for the participants to better perceive the messages spoken by the robot while simultaneously looking around in the environment to find a suitable referent in the context of the robot. This line of reasoning is supported by the circumstance that when the robot passes a product and issues a spoken utterance (using the on-board speech synthesiser) *without slowing down*, it seems to lead to users being frustrated and reporting that they experience a sense of loss of control.

All in all this calls for a careful design of robot movements to accompany its communicative utterances, especially when they refer to objects in the environment. However, as our experience shows, due to the fact that it was hard for participants to notice the movements, designing and evaluating such movements pose interesting challenges. In our view this requires adaptation of user-centred design and evaluation methodology, as gestures that are close to what is being noticed by users affect the reaction, interpretation and communicative behaviour during design and evaluation sessions.

Acknowledgement

This research has partly been carried out in the CommRob project and is partially funded by the EU (contract number IST-045441 under the 6th framework programme).

REFERENCES

- [1] Herbert H. Clark and Meredyth A. Krich, 'Speaking while monitoring addressees for understanding', *Journal of Memory and Language*, (2004).
- [2] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg, 'Wizard of Oz studies - why and how', *Knowledge-Based Systems*, **6**(4), 258–256, (1993).
- [3] K. A. Ericsson and H. A. Simon, 'Verbal Reports as Data', *Psychological Review*, **87**(3), 215–251, (1980).
- [4] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn, 'A survey of socially interactive robots', *Robotics and Autonomous Systems*, **42**(3-4), 143–166, (2003).
- [5] Satinder P. Gill and Jan Borchers, 'Knowledge in co-action: social intelligence in collaborative design activity', *AI & Society*, **17**(3), 322–339, (2003).
- [6] Michael Göller, Florian Steinhardt, Thilo Kerscher, J. Marius Zöllner, and Rudiger Dillmann, 'Robust navigation system based on rfid transponder barriers for the interactive behavior-operated shopping trolley inbot', *The International Journal Industrial Robot*, **36**(4), (2009).
- [7] Anders Green and Helge Hüttenrauch, 'Making a Case for Spatial Prompting in Human-Robot Communication', in *Multimodal Corpora: From Multimodal Behaviour theories to usable models, workshop at the Fifth international conference on Language Resources and Evaluation, LREC2006*, Genova, Italy, (May 22-27 2006).
- [8] Anders Green, Helge Hüttenrauch, Mikael Norman, Lars Oestreicher, and Kerstin Severinson Eklundh, 'User-Centered Design for Intelligent Service Robots', in *Proceedings of 9th IEEE International Workshop on Robot and Human Interactive Communication*, Osaka, Japan, (2000).
- [9] Anders Green, Helge Hüttenrauch, and Kerstin Severinson Eklundh, 'Applying the Wizard-of-Oz framework to Cooperative Service Discovery and Configuration', in *13th IEEE International Workshop on Robot and Human Interactive Communication RO-MAN 2004*, pp. 575–580, (20-22 Sept 2004).
- [10] Edward T. Hall, *The Hidden Dimension: Man's Use of Space in Public and Private*, The Bodley Head Ltd, London, UK, 1966.
- [11] Helge Hüttenrauch, Kerstin Severinson Eklundh, Anders Green, Elin Anna Topp, and Henrik I. Christensen, 'What's in the Gap? Interaction Transitions that make HRI work', in *Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2006)*, Hatfield, UK, (Sep 6–8 2006).
- [12] Hiroshi Ishii, Craig Wisneski, Scott Brave, Andrew Dahley, Matt Gorbett, Brygg Ullmer, and Paul Yarin, 'ambientroom: integrating ambient media with architectural space', in *CHI '98: CHI 98 conference summary on Human factors in computing systems*, pp. 173–174, New York, NY, USA, (1998). ACM.
- [13] Hermann Kaindl, Juergen Falb, and Cristian Bogdan, 'Multimodal Communication Involving Movements of a Robot', in *CHI '08 extended abstracts on Human factors in computing systems*, pp. 3213–3218, New York, NY, USA, (2008). ACM.
- [14] Takayuki Kanda, Masahiro Shiomi, Zenta Miyashita, Hiroshi Ishiguro, and Norihiro Hagita, 'An affective guide robot in a shopping mall', in *HRI '09: Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pp. 173–180, New York, NY, USA, (2009). ACM.
- [15] Adam Kendon, 'Gesture', *Annual Review of Anthropology*, **26**, 1089–128, (1997).
- [16] Stephen C. Levinson, *Pragmatics*, Cambridge Textbooks in Linguistics, Cambridge University Press, 1983.
- [17] W. E. Mackay, A. V. Ratzer, and P. Janacek, 'Video artifacts for design: bridging the gap between abstraction and detail', in *Proceedings of the 3rd Conference on Designing interactive Systems: Processes, Practices, Methods, and Techniques*, eds., D. Boyarski and W. A. Kellogg, pp. 72–82, New York City, New York, United States, (August 17 – 19 2000).
- [18] Marek Michalowski, S. Sabanovic, and Reid Simmons, 'A Spatial Model of Engagement for a Social Robot', in *9th IEEE International Workshop on Advanced Motion Control*, pp. 762 – 767, (March 2006).
- [19] Candace L. Sidner, Cory D. Kidd, Christopher Lee, and Neal Lesh, 'Where to look: a study of human-robot engagement', in *IUT'04: Proceedings of the 9th international conference on Intelligent User Interfaces*, pp. 78–84, New York, NY, USA, (2004). ACM Press.
- [20] Lucy Suchman and Randall Trigg, 'Understanding practice. video as medium for reflection and design', in *Design at Work*, eds., Greenbaum J. and M. Kyng, Laurence Erlbaum Associates, (1991).
- [21] T. Tasaki, K. Komatani, T. Ogata, and H.G. Okuno, 'Spatially Mapping of Friendliness for Human-Robot Interaction', *Intelligent Robots and Systems, 2005. (IROS 2005). 2005 IEEE/RSJ International Conference on*, 1277–1282, (Aug. 2005).
- [22] David R. Traum and James F. Allen, 'Discourse Obligations in Dialogue Processing', in *Proceedings of the Thirty-Second Meeting of the Association for Computational Linguistics*, ed., James Pustejovsky, pp. 1–8, San Francisco, (1994).
- [23] M. L. Walters, K. Dautenhahn, K. L. Koay, C. Kaouri, R. te Boekhorst, C. L. Nehaniv, I. Werry, and D. Lee, 'Close encounters: Spatial distances between people and a robot of mechanistic appearance', in *Proceedings of IEEE-RAS International Conference on Humanoid Robots (Humanoids2005)*, pp. 450–455, Tsukuba, Japan, (December 5-7 2005).
- [24] M. L. Walters, K. Dautenhahn, S. N. Woods, and K. L. Koay, 'Robotic etiquette: Results from user studies involving a fetch and carry task', in *Proceedings of the 2nd ACM SIGCHI/SIGART Conference on Human-Robot Interaction (HRI 07)*, Washington DC, USA., (March 2007).
- [25] Mark Weiser, 'The Computer for the 21st Century', *Scientific American*, **265**, 94–104, (1991).
- [26] Mark Weiser and John Seely Brown, 'The Coming Age of Calm Technology', (1996).
- [27] Fumitaka Yamaoka, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita, 'How close?: model of proximity control for information-presenting robots', in *HRI '08: Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, pp. 137–144, New York, NY, USA, (2008). ACM.

Validating Characterizations of Sociality in HRI: The Case of Interaction Patterns

Peter H. Kahn, Jr.¹, Brian T Gill², Aimee L. Reichert¹,
Takayuki Kanda³, Hiroshi Ishiguro^{3,4}, and Jolina H. Ruckert¹

Abstract. As the HRI field evolves, researchers increasingly seek to provide characterizations of sociality in human-robot interaction. But how does one assess whether the characterizations are valid? Using our research on “interaction patterns” as a case in point, this paper offers 5 approaches toward establishing validity: The psychometric approach, literary approach, modeling approach, philosophical approach, and structural approach. We argue that when it comes to validating characterizations of sociality in HRI, too often people ask for evidence of psychometric validity, without clarity of what that involves, and without awareness of the benefits of these other approaches.

1 INTRODUCTION

As the HRI field evolves, researchers increasingly seek to provide characterizations of sociality in human-robot interaction. These characterizations may be framed broadly, as when researchers seek to show how people establish *friendships* with robots, or *trust* robots, and *care* about robots, or attribute to robots *moral accountability* [6, 13, 17, 24, 34, 37, 46, 48]. The characterizations may also be framed with greater specificity. For example, researchers in HRI have proposed different characterizations of psychological “benchmarks” to measure success in building increasingly humanlike robots. To date, researchers have characterized such benchmarks as *privacy*, *reciprocity*, *imitation*, *intrinsic moral value*, *conventionality*, *creativity*, *authenticity of relation*, and *autonomy* [5, 9, 15, 33, 37, 51]. Characterizations such as these are “foundational” in the sense that they seek to say something substantive, deep, and abiding about human-robot interaction.

When researchers propose foundational characterizations of sociality in HRI, an important question emerges: How does one know whether the characterizations are valid? It is an important question for if one wants to know something about the world, one almost always wants one’s knowledge to be valid. The problem however with this question is that it is often asked as if the answer will and indeed must entail some psychometric assessment of validity. But the psychometric approach to validity is only one of many, and not often the most amenable to validating characterizations of sociality.

Thus in this paper we distinguish between five meanings of validity, and their corresponding methods, using our research on “interaction patterns” as a case in point. Our purpose is to provide some clarity when researchers seek to assess the validity of their proposed characterizations of sociality in HRI.

2 INTERACTION PATTERNS

By interaction patterns we mean foundational features of social interaction between humans and robots, characterized abstractly enough such that many different instantiations of the pattern can be uniquely realized given different types of robots, purposes, and contexts of use, while delimited so as to meaningfully segment interaction [31].



Figure 1. Initial Introduction Interaction Pattern

Consider, for example, the interaction pattern we have called the “Initial Introduction” (Figure 1). When people are introduced to a new person, typically in a Western culture they engage in a largely scripted and conventionally-established interactional repertoire wherein they (a) recognize the other with their eyes, (b) engage in ritualized short exchanges of salutation (“hello”), (c) which are responded to by the other (“hi”), (d) offer

¹Dept. of Psychology, Univ. of Washington, Seattle, Email: {pkahn, aimee3, jhr333}@uw.edu.

²Dept. of Mathematics, Seattle Pacific Univ., Email: bgill@spu.edu.

³Intelligent Robotics and Communication Lab, ATR, Kyoto, Japan, Email: kanda@atr.jp.

⁴Dept. of Systems Innovation, Osaka Univ., Japan, Email: Ishiguro@sys.es.osaka-u.ac.jp

reciprocal statements of initial pleasure (“pleased to meet you”, “pleased to meet you, too”), (e) ask polite questions of concern for the other (“how are you?”), and (f) engage in physical reciprocity (a handshake), (g) with bodies frontally positioned to one another. In non-Western cultures, different instantiations of this pattern occur. For example, people might bow to one another upon meeting, instead of shaking hands, or offer each other a Namaste greeting with hands together. While the Introduction is never enacted exactly the same way twice, the activity is structured, it follows a recognizable pattern. It is likely universal. If we were dropped into an unknown culture, it is likely that people in the culture would enact some form of this pattern, and that when we saw it enacted for the first time we would immediately recognize it. In turn, such a pattern is but one of many that we believe can be used to help structure human-robot interaction.

Our conceptual framework for interaction patterns draws on Christopher Alexander and his colleagues work in the field of architecture. According to Alexander et al. [4], a design “pattern describes a problem which occurs over and over again in our environment, and then describes the core of the solution to that problem, in such a way that you can use this solution a million times over, without ever doing it the same way twice” (p. x). For example, one of their 253 patterns is titled “Light on two sides of every room.” They write: “The importance of this pattern lies partly in the social atmosphere it creates in the room” (p. 748). They also write that this “pattern, perhaps more than any other single pattern, determines the success or failure of a room” because “when they have a choice, people will always gravitate to those rooms which have light on two sides, and leave the rooms which are lit only from one side unused and empty” (p. 747). There is now a burgeoning body of work that extends the idea of design patterns into the fields of ubiquitous computing [10], software engineering [18], interaction design [7, 45], and usability [8, 25].

In 2008 we presented an initial account of how the Interaction Pattern approach can be implemented in a laboratory context, and provided descriptions of eight such patterns [31]. The Introduction is one of those. We now summarize the other seven interaction patterns to provide the reader with a better sense of what they look like.

Didactic Communication. One of the least complicated forms of social communication involves the transmission of information from one to another. Didactic Communication is an interaction pattern for such one-way communication of information, situated in a context where each party has motivation to remain engaged. When teachers lecture to a class they embody this interaction. Potentially robots, even with today’s capabilities, could be very good at generating instantiations of this interaction since it requires minimal responsiveness to the listener.

Prosocial Request. Part of social interaction involves at times requesting assistance from another: for example, asking for the whereabouts of an item, for directions to a location, or for how to use a tool or new software program.

In Motion Together. Being in a social relationship with others can involve aligning one’s physical movement with others, such as often occurs when walking together (Figure 2).

Personal Interests and History. Social life gains much of its character and depth through sharing of one’s personal interests

and history with others, and of being known and knowing others through such sharing.

Recovering from Mistakes. Social life involves at times the making of mistakes. Recovering from Mistakes is an interaction pattern that creates the potential for both parties to maintain a social affiliation following the mistake.

Physical Intimacy. One important social characteristic of humans (and most primates) is that they engage in holding, touching, embracing, and other forms of physical intimacy. Physical Intimacy is an interaction pattern that allows this characteristic to find expression.

Turn-Taking in Game Context. Most social games, no matter their variation, involve taking turns with one another, such as many board games (e.g., Monopoly), many card games (e.g., poker), tag, and baseball, to name just a few. Turn-Taking in a Game Context is an interaction pattern for sociality that may easily set into motion claims of unfairness.

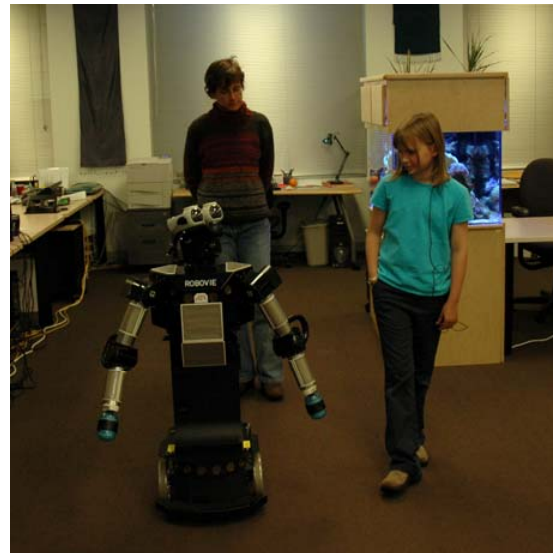


Figure 2. In Motion Together Interaction Pattern

If an interaction patterns program – something of the form we are proposing here – proves successful, it could provide HRI researchers with (a) a new methodology by which to conceptualize and build meaningful social relationships with robots, (b) foundational basic knowledge about what is unique to human-robot interaction, as compared to human-human interaction, and (c) a more systematic approach to help shape decisions about which technical problems in HRI to work on insofar as certain interaction patterns will likely be found to be particularly important in supporting meaningful social relationships, and thus might well merit greater resources (funding and time) to find the means for robots to implement the interaction patterns autonomously in different contexts.

3 FIVE APPROACHES TO VALIDATING CHARACTERIZATIONS OF SOCIALITY

With this handful of interaction patterns as a backdrop, we can now turn to different ways that validity can be understood and assessed.

3.1 The Psychometric Approach

Perhaps the most common approach that people think of in terms of establishing validity is based on psychometrics. But what this approach is about has changed as ideas about psychometric validity have taken shape for over sixty years.

In 1954, the American Psychological Association (APA) published its first *Technical Recommendations for Psychological Tests and Diagnostic Techniques* [1]. This publication described four categories of validity: *predictive validity*, *concurrent validity*, *content validity*, and *construct validity*. In 1966, the APA with the American Educational Research Association and the National Council on Research in Education together published an update [2]. The 1966 update combined predictive and concurrent validity into a single category called *criterion-related validity*. Criterion-related validity involved comparing scores on a test or scale with other variables or measures taken on the same subjects, typically by computing correlations between the different measures. If the different measures were all obtained at approximately the same time, criterion-related validity was known as concurrent validity. Concurrent validity was often studied to determine whether or not one test could substitute for another; or, in other words, whether or not different tests were measuring “the same thing.” When the score on the test was compared to other measures obtained some time after the original test, criterion-related validity became known as predictive validity.

In turn, in this same 1966 landmark publication, content validity was understood as assessing how well the content of a test actually represented the entire range of situations or subject matter about which conclusions were to be drawn. Assessments of content validity were somewhat subjective and not quantified with statistical methods; but some rational justification of content validity was necessary to avoid easy rejection of a researcher’s findings. For example, a researcher might develop a survey consisting of questions relating to issues such as abortion and gay rights. On the basis of this survey, the researcher might develop a score which was claimed to represent how liberal or conservative respondents were. However, this interpretation of the scores would presumably have poor validity since the questions used on the survey only represent highly contentious social issues and not a wider range of subject matter (such as economic issues) presumably implied by the labels “liberal” and “conservative.”

In 1966, construct validity was also discussed extensively, though the term did not mean quite what it sounds like. According to Cronbach [12] “construct validity is evaluated by investigating what psychological qualities a test measures; i.e., by determining the degree to which certain explanatory concepts or constructs account for performance on the test” (p. 144). As an example, consider the IQ test which seeks to provide a quantitative measure of the construct of intelligence. Is intelligence a valid construct? Construct validation does not

actually answer that question. Rather, the focus is on the interpretation of the scores on the IQ test itself, of whether the interpretations are warranted based on empirical evidence.

The emphasis on the different “types” of validity in the 1950’s and 1960’s led to a fragmented view of psychometric validity, and there was eventually a movement to try to unify the theory of validity. Ultimately, the goal of all psychometric validation questions is to assess “the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests” (p.9) [3]. Over time, all test scores came to be viewed as measures of some underlying construct, and the first four forms of validity (content, predictive, concurrent, and criterion-related validity) came to be viewed as aspects of construct validity. Since it is the link between the test scores and the underlying construct that gives the scores their meaning and supports their interpretation, ultimately all psychometric validity questions are construct validity questions. Thus in their most recent Standards [3], the American Psychological Association dropped the term construct validity and reverted back to simply the term validity. For example, the most recent American Psychological Association position [3] states:

Validation logically begins with an explicit statement of the proposed interpretation of test scores, along with a rationale for the relevance of the interpretation to the proposed use. The proposed interpretation refers to the construct or concepts the test is intended to measure. Examples of constructs are mathematics achievement, performance as a computer technician, depression, and self-esteem. To support test development, the proposed interpretation is elaborated by describing its scope and extent and by delineating the aspects of the construct that are to be represented. The detailed description provides a conceptual framework for the test, delineating the knowledge, skills, abilities, processes, or characteristics to be assessed. The framework indicates how this representation of the framework is to be distinguished from other constructs and how it should relate to other variables (p. 9).

With such language, all of the different aspects of validity are being wrapped within the single term. Rather than talking about different “types” of validity, the APA intentionally shifted toward emphasizing different kinds of evidence that might be called upon to support interpretations of test scores for particular uses. Thus, rather than using terms such as “content validity” or “criterion-related validity”, the 1999 Standards discuss “evidence based on test content” and “evidence based on relations to other variables.” All of the available evidence is then combined together to form a single, coherent argument for the validity of the desired interpretations and uses of the test scores.

One key idea stated in the above APA definition is that it is not a sensible question to ask whether a test is valid or not valid. Rather, it is the interpretation of the test scores that are valid or not valid. As the passage states above: “Validation logically begins with an explicit statement of the proposed interpretation of test scores, along with a rationale for the relevance of the interpretation to the proposed use.” Psychometricians have emphasized this idea time and time again since the 1950’s, because people tend to think otherwise. They tend to think that tests themselves are valid or not.

To emphasize this idea – that it is interpretation of tests, and not the test itself, which is the subject of validation – let's imagine that we develop a new standardized test of basic algebra skills. Now let's say that three different schools use our test in three different ways. One school uses our test to assess the effectiveness of instruction in their algebra courses. The second school uses our test as a screening tool to determine whether or not to allow students into a geometry course, which is next in the math sequence. The third school uses our test to plan instruction in a course by giving the test to students at the beginning of the course and using an item-by-item analysis of student errors on the test to select topics to focus on during the course. Each of these different uses of the test involves different interpretations and conclusions based on students' performance on the test. In turn, our test might prove to be an excellent assessment tool for the end of the algebra course. But our test may correlate very poorly with success in the sequenced geometry course, and it would likely enough function very poorly as the basis for an item-by-item restructuring of curriculum. If that was what we found empirically, then it would be valid to use the test in the first way, but not in the second and third.

These strengths and constraints of psychometric validity apply not just to scores on a test but “generically in its broadest sense to mean any coding or summarization of observed consistencies or performance regularities on a test, questionnaire, observation procedure, or other assessment devices such as work samples, portfolios, and realistic problem simulations” (p. 741) [40]. It follows, then, that questions such as “Is the questionnaire valid?” or “is your observation procedure valid?” or “is your realistic problem simulation valid?” are not sensible questions from a psychometric standpoint [39]. No more so than “Is the test valid?”

And that is precisely the point we want to drive home about validity and interaction patterns. From the psychometric approach to validity, it is not sensible to ask whether an interaction pattern is valid or not valid. No more so than “Is the test valid?” Or “Is the questionnaire valid?”

The one caveat is that it is possible to invoke the conception of content validity, described above, and ask in that sense of the term whether an interaction pattern is valid. But as we discussed above, evidence for content validity is not based on quantitative or statistical analysis, and tends to be viewed by the psychometric field as not being rigorous. To make it rigorous, we think some of the other approaches to validity described later in this paper could be better employed.

Before we conclude this section on psychometric validity, a reader might respond: “Well, what I mean by the term psychometric validity, as it applies to an interaction pattern is: Does the interaction pattern work? Is it useful? If it does and is, then it's valid; otherwise it's not.” These are reasonable questions. But they are not yet framed in a way that is tractable from the psychometric approach. To reframe them, we would need to specify what is meant by “work” or “useful” in terms of interpretations or conclusions that might follow from the interaction pattern. For example, let's say that we first hypothesized that the Introduction pattern facilitates the ease of an initial greeting. Let's say we could find ways of measuring the “ease of an initial greeting.” Then let's say we compared the Introduction pattern to other patterns or entities to test which provides the greatest ease of an initial greeting. At this point, the interaction pattern itself becomes a factor in an experimental

design. And, in this sense, depending on the empirical results, the interaction pattern could be said to be valid or not. But notice now that we are right back to the essence of what is meant by psychometric validity: whether interpretations of a test score or comparable entity (in this case interaction patterns) are warranted based on the empirical data. That is fine. But we still have not answered the question of whether the characterization of the phenomenon in terms of the specification of the pattern is in itself valid. For that, we need to draw on other validation approaches.

3.2 The Literary Approach

It might seem that that the fields of psychometrics and literary theory have little in common, especially in terms of what they have to say about validity. But both approaches seek to answer a fundamental question: “Are the conclusions warranted from the data?”

In literature one might, for example, collect “data” from the narrative text of Shakespeare's *King Lear* and argue that the play is substantially about the difficulty old men have in giving up power and of recognizing love in family settings; or argue more nuanced positions. It can then be asked, are those interpretations valid? That is a sensible question, and it requires the literary theorist to provide textual evidence and good reasons for the interpretations. It is in this way – in terms of establishing whether a conclusion is warranted based on the data – that asking “Is one's interpretation of *King Lear* valid?” is of the same form of validity as asking “Is IQ a valid predictor of school performance?”

This point is important to recognize. As mentioned earlier, psychometricians often discount the rigor involved in establishing content validity, but that is likely because they focus on questions that have quantitative answers. But literary theorists know all too well that their positions require detailed, sophisticated, logical, and well-documented arguments, and counterarguments, to hold sway and to be valid. For example, literary theorist Stanley Fish highlights the distinction between interpretation and textual evidence, he writes: “The rhetoric of critical argument, as it is usually conducted in our journals, depends upon a distinction between interpretations on the one hand and the textual and contextual facts that will either support or disconfirm them on the other...” (p. 340) [16]. Herbert Dingle [14], on the scientific approach and its application in literary criticism quotes R. G. Moulton: “The fundamental axiom of inductive literary criticism is that ‘*Interpretation in literature is of the nature of a scientific hypothesis, the truth of which is tested by the degree of completeness with which it explains the details of the literary work as they actually stand*’” (pp. 6-7 italics part of original text). E.D. Hirsch [28] expands upon this approach and presents a model for establishing the validity of an interpretation. He writes:

To establish a reading as probable it is first necessary to show . . . that it is possible. This is the criterion of legitimacy: the reading must be permissible within the public norms of the langue [Saussure's term for the language as a system] in which the text was composed. The second criterion is that of correspondence: the reading must account for each linguistic component in the text. Whenever a reading arbitrarily ignores linguistic components or inadequately accounts for them, the

reading may be presumed improbable. The third criterion is that of generic appropriateness: if the text follows the conventions of the scientific essay, for example, it is inappropriate to construe the kind of allusive meaning found in casual conversation. When these three preliminary criteria have been satisfied, there remains a fourth criterion which gives significance to all the rest, the criterion of plausibility or coherence. The three preliminary norms usually permit several readings, and this is by definition the case when a text is problematical. Faced with alternatives, the interpreter chooses the reading which best meets the criterion of coherence (p. 1406).

Here Hirsch presents the methods by which to establish an interpretations' validity, which includes careful observations and logical inferences. An interpretation that captures the greatest coherence or plausibility among the elements within the text is judged to be the most valid [21]. Over the last four decades, this approach has been extended into the social sciences, often under the rubric of "hermeneutics," wherein human behavior has been treated as a "text" to be interpreted, and validity established along similar lines [19, 29, 38].

Thus the literary approach to validity is just as demanding and rigorous as the psychometric approach, though qualitative and not quantitative in form. Moreover, this approach emerged as its own field, as hermeneutics, in terms of applying it to the social sciences. Accordingly, it is possible that some degree of validation of interaction patterns can occur by drawing on this literary approach, where the "text" is the human-robot interaction being interpreted, and where the interaction pattern then becomes in effect the "conclusion" of an argument.

3.3 The Modeling Approach

A third way of understanding the validation of interaction patterns can be framed in terms of the following question: "How well does the pattern bear up compared to the corresponding human benchmark?" Benchmarks in HRI have been defined as "categories of interaction that capture fundamental aspects of human life" [37]. This form of validation is less common to a psychometric approach to validity, and more akin to the approach of a social science modeler. Such a modeler attempts to represent theory about some complex real-world phenomenon by simulating components of the phenomenon [47]. The plausibility of the model increases – and thus it becomes more valid – when model performance matches real-world findings [49].

As a case in point, consider the embodied learning model of infant gaze-following developed by Trafton et al. [49]. Their model simulated 5 theoretical components thought to be central to the emergence of gaze-following in infants. These components were reactivity, habituation, spatial module, gaze-following, and utility learning. To test the plausibility of their model they matched their model's performance to results from an experiment by Corkum and Moore [11], which assessed gaze-following in 6-11 month old infants. The embodied gaze-following model matched these results, and thus provided evidence for the validity of their model. In this sense, the empirical data served as a human-benchmark for gaze-following.

Similarly, interaction patterns can be used to better understand the complexities of social interaction, and they can be said to be more plausible – and thus more valid – when the patterns incorporate components that maximize the match to human-human interactions. That said, an important distinction between modeling and interaction patterns is that a model's validity increases when a parsimonious understanding of real-world systems is achieved by focusing on the least number of components that can account for the real-world data [49]. For example while Trafton et al.'s [49] gaze-following model initially incorporated 5 components, they found that only three of them were critical to the success of the model. In contrast, one way of understanding interaction patterns is that they seek to maximize the number of components accounted for by the pattern in order to better represent the diversity of social interaction.

Recall, for example, the specific details of the Introduction pattern described earlier. We posited that when people enact this pattern, they typically recognize the other with their eyes, engage in ritualized short exchanges of salutation that are responded to by the other, offer reciprocal statements of initial pleasure, ask polite questions of concern for the other, and engage in physical reciprocity, with bodies frontally positioned to one another. It may be possible to enact this pattern with only a single component, such as the ritualized short exchange ("Hello"), and still fit the pattern to human-human interaction, but the simplification of the pattern in this way fails to capture the rich and diverse social interplay that is possible – and often enacted – between interacting entities. Thus interaction patterns that account for more components, while still falling within the parameters of the pattern itself, are in this sense of the term more valid than interaction patterns that account for fewer components.

3.4. The Philosophical Approach

The fourth form of validity seeks to provide sound reasons for the labels of the patterns themselves. Why, for example, call the Introduction an Introduction? Why not call it a "Greeting"? or a "Meeting"? or a "Gathering" or "a Lunch-Time Break"? To develop our answer, think of the distinction between formal and informal fallacies in logic. Formal logical fallacies occur through errors in deductive reasoning (e.g., if $a < b$, and $b < c$, then it is a fallacy not to agree that it must follow that $a < c$). In contrast, informal logical fallacies occur when bad qualitative reasons are offered (e.g., in "argumentum ad hominem" it is a fallacy to critique someone's argument based on the person's character). It is in this informal sense that interaction patterns, to be valid, need good reasons for being called what they are called. Some of those reasons will follow standard understandings of what words mean in one's established language (e.g., English).

Other reasons can be established based on philosophically grounded arguments for conceptualizing a construct [26]. As a case in point, some moral developmental psychologists have drawn on moral philosophers such as John Rawls [44] to characterize the moral domain in terms of judgments that pertain to human welfare and fairness [30], and which generalize to people cross-culturally [35, 50]. In other words, well-grounded

arguments in philosophy have helped to establish the validity of constructs in psychology.

Because we have now invoked the word “construct,” the reader might suggest that construct validity is best established through a psychometric approach rather than a philosophical approach. But as we broached in Section 3.1, it is important to understand what construct validity is and what it is not from a psychometric standpoint. The 1966 APA publication, which, as noted earlier, is one of the canonical standards on validity, says the following:

[Construct validity] is evaluated by investigating what qualities a test measures, that is by determining the degree to which certain explanatory concepts or constructs account for performance on the test. To examine construct validity requires a combination of logical and empirical attack. Essentially, studies of construct validity check on the theory underlying the test. ... Construct validity is ordinarily studied when the tester wishes to increase his understanding of the psychological qualities being measured by the test. Construct validity is relevant when the tester accepts no existing measure as a definitive criterion of the quality with which he is concerned...or when a test will be used in so many diverse decisions that no single criterion applies (p.13).

We find this key passage, and others like it, difficult to interpret. On the one hand, construct validity, by the sound of its name, should seemly tell us whether it is valid or not valid to conceptualize constructs such as intelligence as a unified whole. On the other hand, the above passage is mostly focused on the test (e.g., whether “constructs account for performance on the test” and “the psychological qualities being measured by the test”). It does not actually say anything specific about the theoretical construct itself. The passage does not show us the means, for example, of whether we should hold out for a single construct called intelligence, or whether, as others have proposed [20] there are multiple intelligences, such as mathematical, musical, linguistic, artistic, and social, or whether all of these constructs are not framed in a valid way. It is for this reason that the philosophical approach to validity offers a complementary and important addition to the range of validation approaches.

3.5 The Structuralist Approach

If, roughly cast, the modeler’s form of validity focuses on theory fitting data, and the philosopher’s form of validity focuses on whether theory is well-grounded in an epistemological and ontological conceptual framework, then the structuralist form of validity can be viewed as iterating between both these forms [41].

As a case in point, consider Piaget’s cognitive-developmental theory wherein he sought to characterize the development of the structure of human knowledge. That ambitious undertaking led Piaget to posit four overarching stages in development from infancy through late adolescence – sensory motor, preoperations, concrete operations, and formal operations – through which Piaget characterized how children’s thinking increased in its capacity for synthesis and abstraction. Characterizations of each stage are comprised of detailed formal properties. For example, one central property of concrete operations is that of

reversibility, that in the movement from one conception to another, the original conception can be simultaneously invoked.

How, then, did Piaget seek to establish the validity of his stage characterizations? In part, he drew on philosophical theory of what counted as domains of knowledge – such as of physics, space, time, logic, and number. For each of these domains, he further drew on philosophical criteria of what comprised those domains; and then he sought to find their origins in childhood. Thus, for example, Piaget drew widely from mathematicians such as Russell and Whitehead to help formulate the ideas for logical groupings [43]. But, for Piaget, the philosophical criteria by themselves did not establish validity. Rather the theory led to investigations of how the child formed concepts about these domains in the world, which in turn led to revisions in the theory of what those domains consisted of, which in turn led to further refined investigations, more empirical data, and so it went, back and forth between theory and data, for the entire span of Piaget’s long career. Piaget called this approach *genetic epistemology* [42]. By this term, he meant that valid knowledge did not exist “out there” to be discovered in a world independent of humans, as proposed by correspondence theories of epistemology, nor is valid knowledge just socially constructed, as in it being anything we want it to be or might imagine it to be. Rather valid knowledge is constructed through interaction with the reality of an external world and the biological affordances and constraints of our species.

Our approach to establishing the validity of our interaction patterns can often be best understood by this structuralist’s approach. The interaction pattern of the Introduction, for example, draws substance from unpacking the form and function of Introductions across time, context, and culture; but it further needs empirical evidence to increase its validity; and in the process of gaining such evidence, more conceptual understanding of the pattern itself becomes revealed – and the iterative process is underway.

One difference in Piaget’s structuralist approach and ours is that Piaget focused on the structure of individual cognition, and its ontogenesis. In contrast, with our focus here on interaction patterns, we are interested in the structure of human interaction between the organism and its environment. Our endeavor with interaction patterns is more akin to ecological psychologists like Gibson [22, 23] who focused on person-environment interaction as the unit of analysis; though we differ from many ecological psychologists in not positing a pragmatist epistemology, where truth is what “works” [27] – but a broader epistemology that is amenable to the different forms of validity distinguished in this paper.

4 NEXT STEPS

In our view, all of the above forms of validity can apply to interaction patterns, though perhaps the structuralist approach is the most relevant. What, then, might next steps look like in generating and validating interaction patterns? First off, we need to build up a large set of possible interaction patterns, perhaps in the range of 50-70. Why not 15 or 250? We need enough interaction patterns to begin to map out the diversity of interactions, so as to make this endeavor compelling. Fifteen is too few. Two hundred and fifty seems too ambitious at this early stage.

In generating this first large group of interaction patterns, why will we choose some over others? Our answer is that some of our patterns – such as the Introduction – seem to us fundamental for motivating many forms of human-robot interaction, and thus warrant particular attention. Other patterns tap into what will become huge societal issues (and concerns) in the immediate future, and thus warrant our particular attention. For example, the interaction pattern of physical intimacy with a robot has been to date sensationalized by such books as *Love and Sex with Robots* [36]. By focusing on less sexual instantiations of physical intimacy, our immediate work will let us speak to these issues on a societal level, and shape healthy directions for this interaction pattern in HRI. Still other patterns (such as physical motion) are opportunistic in that we are building on the existing repertoire that is currently engineered autonomously in the robot we have been working with (ATR's "Robovie"). Finally other patterns will be chosen to fit different categories of interaction, per our theoretical commitments of how to understand the ontological nature of human-robot interaction: that is, it can be like human-human interaction, human-animal interaction, and human-machine interaction, as well as something uniquely itself [32].

Two concerns in particular could be raised about our general approach to interaction patterns. One concern is this: Are there an infinite number of interaction patterns that correspond to the infinite number of ways humans interact with other humans, animals, nature, and artifacts? If so, there's no end to this project. Our answer is that an endeavor like this one might seem to the reader as endlessly open-ended, as one can seemingly always come up with a new interaction pattern in HRI. But at some point, "new" patterns will be seen as providing very little new information to an existing pattern, and thus either subsumed within an existing pattern or recognized as an instantiation of an existing pattern. The validity claim here is somewhat akin to that of parsimony by the modeler (Section 3.3).

A second concern is that we have been speaking of the interaction patterns as independent entities. But is it not the case that many patterns interact with one another, and do not exist in discrete units? Our answer is that our laboratory work in part does aim first to carve out discrete units, but eventually to link and integrate them. By analogy, Alexander articulated his 253 patterns as discrete units but emphasized how they can be combined into a "pattern language." Framed as our long-term goals, we aim for something similar: an interaction pattern language for sociality in HRI.

5 CONCLUSION

The *Oxford English Dictionary* says the following about validity, that it is the "quality of being well-founded on fact, or established on sound principles, and thoroughly applicable to the case or circumstances; soundness and strength (of argument, proof, authority, etc.)." In turn, there are different ways to establish this quality of being well-founded. Psychometrics is just one, though often the most referenced in the social-scientific communities. Thus, in this paper, we have offered five overarching approaches to validation. To portray these approaches, we used interaction patterns as a case in point. As the HRI field evolves, many useful approaches for characterizing social interaction between humans and robots will emerge. Thus

the approaches toward establishing validity as sketched in this paper may well have broad appeal.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant Numbers IIS-0842832 and IIS 0905289. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank Greg Trafton and Daniel Levin for early discussions on these matters.

REFERENCES

- [1] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 'Technical recommendations for psychological tests and diagnostic techniques', *Psychological Bulletin*, **51**, supplement, 1954.
- [2] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for Educational and Psychological Testing*, American Educational Research Association, Washington, D.C., 1966.
- [3] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for Educational and Psychological Testing*, American Educational Research Association, Washington, D.C., 1999.
- [4] C. Alexander, S. Ishikawa, and M. Silverstein, *A Pattern Language: Towns, Buildings, Construction*, Oxford University Press, New York, 1977.
- [5] D. Bernstein, K. Crowley and I. Nourbakhsh, 'Working with a robot: Exploring relationship potential in human-robot systems', *Interaction Studies: Social Behavior and Communication in Biological and Artificial Systems*, **8**, 465-482, (2007).
- [6] C.L. Breazeal, *Designing Sociable Robots: Intelligent Robots and Autonomous Agents*, MIT Press, Cambridge, MA, 2002.
- [7] J. Brochers, *A Pattern Approach to Interaction Design*, John Wiley and Sons, Chichester, England, 2001.
- [8] G. Casaday, *Notes on a Pattern Language for Interactive Usability*, Paper presented at the CHI '97 Extended Abstracts on Human Factors in Computing Systems, Atlanta, Georgia, (1997).
- [9] J. Cassell and A. Tartaro, 'Intersubjectivity in human-agent interaction', *Interaction Studies: Social Behavior and Communication in Biological and Artificial Systems*, **8**, 391-410, (2007).
- [10] E.S. Chung, J.I. Hong, J. Lin, M.K. Prabaker, J.A. Landay, and A.L. Liu, 'Development and evaluation of emerging design patterns for ubiquitous computing', in *Proceedings of the 2004 Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques*, Cambridge, MA, USA, 233-242, (2004).
- [11] V. Corkum and C. Moore, 'The origins of joint visual attention in infants', *Developmental Psychology*, **34**, 28-38 (1998).
- [12] L.J. Cronbach, 'Test validation', in *Educational Measurement*, ed., R.L. Thorndike, American Council on Education, Washington, D.C., 2nd edn., 443-507, (1971).
- [13] K. Dautenhahn and I. Werry, 'Towards interactive robots in autism therapy', *Pragmatics and Cognition*, **12**, 1-35, (2004).
- [14] H. Dingle, *Science and literary criticism*, Nelson, London, 1949.
- [15] D. Feil-Seifer, K. Skinner and M.J. Mataric, 'Benchmarks for evaluating socially assistive robotics', *Interaction Studies: Social*

- Behavior and Communication in Biological and Artificial Systems*, 8, 423-439, (2007).
- [16] S.E. Fish, *Is There a Text in This Class?: The Authority of Interpretive Communities*, Harvard University Press, Cambridge, MA, 1980.
- [17] T. Fong, I. Nourbakhsh, and K. Dautenhahn, 'A survey of socially interactive robots', *Robotics and Autonomous Systems*, **42**, 143-166, (2003).
- [18] R.P. Gabriel, *Patterns of Software: Tales from the Software Community*, Oxford University Press, New York, 1996.
- [19] H.G. Gadamer, *Philosophical Hermeneutics*, trans. D.E. Linge, University of California Press, Berkeley, 1977.
- [20] H. Gardner, *Frames of mind: The theory of multiple intelligences*. Basic Books, New York, 1983.
- [21] K.J. Gergen, *An Invitation to Social Construction*, Sage, London, 1999.
- [22] J.J. Gibson, *The Senses Considered as Perceptual Systems*, Houghton-Mifflin, Boston, 1966.
- [23] J.J. Gibson, *The Ecological Approach to Visual Perception*, Lawrence Erlbaum, Hillsdale, NJ, 1986 (original work published 1979).
- [24] R. Gockley, A. Bruce, J. Forlizzi, M. Michalowski, A. Mundell, S. Rosenthal, et al., 'Designing robots for long-term social interaction, in *Proceedings of IROS*, IEEE, Piscataway, NJ, 305-312, (2005).
- [25] I. Graham, *A Pattern Language for Web Usability*, Addison-Wesley, London, England, 2003.
- [26] J. Habermas, *Communication and the Evolution of Society*, trans., T. McCarthy, Beacon Press, Boston, MA, 1979.
- [27] H. Heft, 'The social constitution of perceiver-environment reciprocity', *Ecological Psychology*, **19**, 85-105, (2007).
- [28] E.D. Hirsch, *Validity in Interpretation*, Yale University Press, New Haven, 1967.
- [29] M.A. Honey, 'The interview as text: Hermeneutics considered as a model for analyzing the clinically informed research interview', *Human Development*, **30**, 69-82, (1987).
- [30] P.H. Kahn, Jr., 'Children's obligatory and discretionary moral judgments', *Child Development*, **63**, 416-430, (1992).
- [31] P.H. Kahn, Jr., N.G. Freier, T. Kanda, H. Ishiguro, J.H. Ruckert, R.L. Severson, and S.K. Kane, 'Design patterns for sociality in human-robot interaction', in *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction (HRI '08)*, Amsterdam, Netherlands, 97-104, (2008).
- [32] P.H. Kahn, Jr., B. Friedman, D.R. Pérez-Granados, and N.G. Freier, 'Robotic pets in the lives of preschool children', *Interaction Studies: Social Behavior and Communication in Biological and Artificial Systems*, **7**, 405-436, (2006).
- [33] P.H. Kahn, Jr., H. Ishiguro, B. Friedman, T. Kanda, N.G. Freier, R.L. Severson, and J. Miller, 'What is human? - Toward psychological benchmarks in the field of human-robot interaction', *Interaction Studies: Social Behavior and Communication in Biological and Artificial Systems*, **8**, 363-390, (2007).
- [34] T. Kanda, T. Hirano, D. Eaton and H. Ishiguro, 'Interactive robots as social partners and peer tutors for children: A field trial', *Human-Computer Interaction*, **19**, 61-84, (2004).
- [35] M. Killen, 'Children's social and moral reasoning about exclusion', *Current Directions in Psychological Science*, **16**, 32-26, (2007).
- [36] D. Levy, *Love and Sex with Robots. The Evolution of Human-Robot Relationships*, Harper Collins, New York, 2007.
- [37] K.F. MacDorman and P.H. Kahn, Jr., 'Introduction to the special issues on psychological benchmarks of human-robot interaction', *Interaction Studies: Social Behavior and Communication in Biological and Artificial Systems*, **8**, 359-362, (2007).
- [38] S. McNamee and K.J. Gergen, 'Therapy as social construction', *Inquiries in Social Construction*, Sage, London, (1992).
- [39] S. Messick, 'Validity', in *Educational Measurement*, ed., R.L. Linn, MacMillan, New York, 3rd edn., 13-103, (1989).
- [40] S. Messick, 'Validity of psychological assessment', *American Psychologist*, **50**, 741-749, (1995).
- [41] J. Piaget, *Structuralism*, Harper and Row, New York, 1970.
- [42] J. Piaget, *Genetic epistemology*. W.W. Norton & Co., New York, 1971.
- [43] J. Piaget, 'Piaget's theory', in *Handbook of Child Psychology: Vol 1. History, Theory, and Methods*, ed, W. Kessen, series ed, P.H. Mussen, 4th edn., 103-128, Wiley, New York, (1983).
- [44] J. Rawls, *A Theory of Justice*, Harvard University Press, Cambridge, 1971.
- [45] N.A. Salingaros, 'The structure of pattern languages', *Architectural Research Quarterly*, **4**, 149-161, (2000).
- [46] M. Shiomi, T. Kanda, H. Ishiguro, and N. Hagita, 'Interactive humanoid robots for a science museum, in *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction (HRI '06)*, Salt Lake City, Utah, USA, 305-312, (2006).
- [47] C.S. Taber and R.J. Timpone, R. J, *Computational modeling*. Sage university papers series, no. 07-113, Sage Publications, Thousand Oaks, Calif, (1996).
- [48] F. Tanaka, A. Cicourel, and J.R. Movellan, 'Socialization between toddlers and robots at an early childhood education center', *Proceedings of the National Academy of Science*, **104**, 17954-17958, (2007).
- [49] J.G. Trafton, A.M. Harrison, B. Fransen, and M. Bugajska, 'An embodied model of infant gaze-following', *International Conference of Cognitive Modeling*, (2009).
- [50] E. Turiel, 'The development of morality', in *Handbook of child psychology: Social, emotional, and personality development*, eds W. Damon and N. Eisenberg, J. Wiley, New York, NY, 863-932, 5th edn, (1998).
- [51] S. Turkle, 'Authenticity in the age of digital companions', *Interaction Studies: Social Behavior and Communication in Biological and Artificial Systems*, **8**, 501-517, (2007).

Smart Action Selection Architecture Taking into account both Goal-orientedness and Proactivity

Woo Young Kwon¹ and Il Hong Suh²

Abstract. Intelligent interaction with humans plays an important role in robotic applications. For such interaction, a robot has to select an action by considering not only the current situation but also the predicted future situation. Predictive abilities facilitate anticipation and smart decision-making by allowing the robot to decide what action has to be performed. In this paper, we propose smart action selection architecture taking into account both goal-orientedness and proactivity. To select actions with respect to a current situation, we use a reactive plan-based action selection mechanism that can deal with sequential goal-oriented behaviors within the built-in plan. To select actions appropriate to a future situation, we propose a probabilistic temporal prediction method, and integrate the prediction method with the reactive plan-based action selection mechanism. To show the validity of our proposed architecture, we present some experimental results.

1 Introduction

To accomplish a given task, a robot has to select an action that is appropriate for the current situation as well as a given goal. This requirement can be summarized as “what to do next” and is one of the most fundamental problems for a robot. To solve this problem, various approaches have been proposed including classical Artificial Intelligence (AI) techniques [1], ethology and biology-based techniques [2] [3] [4], a fuzzy-based technique [5], and a reinforcement learning based technique [6]. The architecture of early systems was based on traditional AI planning methods that consists of a sense-plan-act sequential cycle. These traditional AI planning methods have some limitations because they assume that sensors of a system provide accurate knowledge about the state of the world. This assumption is not valid because of a number of factors such as the changing state of the world, limited processing resources, and noisy unreliable sensory information [7]. As alternatives to classical AI techniques, behavior-based approaches have been proposed. Brooks [8] has suggested an architecture called “subsumption architecture,” which is composed of competence modules with fixed priorities. This approach gives an advantage for a robot to accomplish a goal in a complex environment by selecting situation-adequate actions.

Although many researchers have addressed situation-adequate action selection, there has been little attention to proactivity; proactivity is acting in advance of a future situation, rather than just reacting. For a robot to be smart, it has to make a proactive action with respect

to the predicted future situation. Proactive and prediction ability permits anticipation and smart decision-making by allowing a robot to decide which action to perform to obtain or avoid a predicted future situation [9]. By the proactive execution of actions with respect to predicted events, a robot can maximize the utility of task execution in many robotic applications such as human-robot cooperation.

On the basis of psychological and biological backgrounds, researchers have proposed computational models for the temporal prediction of future events; these models are essential for achieving proactivity. One of the most widely known approaches to prediction is classical conditioning. This is a learning theory that explains how animals learn to predict the emergence of an unconditioned stimulus from the presentation of a conditioned stimulus [10] [11]. This method is useful for explaining how an animal and/or a robot predicts the presentation of an unconditioned stimulus on the basis of conditioned stimuli. However, it is too simple to apply to real robotic applications because it usually consider only the static and temporal relations between two events. In [12], Bruke has integrated a temporal prediction model into a behavior-based architecture for autonomous virtual creatures based on scalar expectancy theory and a rate estimation theory by Gallistel and Gibbon [13] that is a temporal prediction model based on psychological theory. However, the scalar expectancy theory is too simple and the rate estimation theory is computationally ineffective. In [14], a proactive and goal-oriented action selection mechanism was proposed based on temporal prediction, but a deterministic temporal method was used in the paper. Therefore, it may not be suitable for real robotic applications.

In general, robots are required to learn complex relationships among many stimuli; moreover, these relationships can change dynamically and non-deterministically. In real robotic applications, therefore, an effective probabilistic method is required for the prediction of future events. Moreover, to select proactively an action, the temporal prediction method should be incorporated into an action selection mechanism. In this paper, we propose novel architecture for proactive action selection for smart robots. For situation-adequate and goal-oriented action selection, we use a reactive-plan based action selection mechanism [15] that is based on ethology. For proactive action selection, we propose a probabilistic temporal prediction method, and integrate this method with the action selection mechanism. By using the temporal prediction method, a robot can select an action with respect to the current situation as well as the predicted future situation. By the execution of a proactive action using the prediction of future events, the robot can reduce not only the task execution time for the action but also the human’s waiting time.

This paper is organized as follows. In Section II, a proactive based action selection mechanism is presented. In Section III, a probabilistic temporal prediction method is discussed. Section IV presents the

¹ Hanyang University, Korea, email: wykwon@incorl.hanyang.ac.kr

² Hanyang University, Korea, email: ihsuh@hanyang.ac.kr (Corresponding author)

experimental results and concluding remarks follow in Section V.

2 Action Selection Architecture Taking into account both Goal-orientedness and Proactivity

A robot task is usually described as a nominal sequence of state transitions from an initial state s_i to a goal state s_g . However, a state can be accidentally changed without consideration of the initial task plan. It is difficult to consider all such accidental state changes when programming a robot task using classical programming languages. Classical programming assumes that all possible states can be defined for a robot and its environment. This assumption can be made to hold by defining any unexpected state as an exceptional state. Then, the problem becomes: (1) to program a robot task as an irreducible Markov process, in which a state in the robot task program can be migrated to any other state. This problem requires an action selection mechanism to be designed as a fully connected finite state machine as shown in Figure 1.

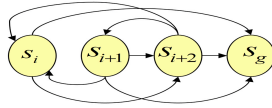


Figure 1. Example of task program described as fully connected finite state machine.

(2) In addition, it is necessary to know how close the current state is to the goal state, since the degree of closeness is an important motivation when selecting actions. To fulfill this requirement, a fully connected finite state machine such as a task program needs to be described as an ordered sequence of state-action pairs. The action selection mechanism has to support the programming of this ordered sequence of state-action pairs. (3) Furthermore, the action selection mechanism has to be structured in the way that some portions of such a sequence of state-action pairs for a robot task need to be learned to adapt to new situations, and/or have to be reused for other tasks.

To solve problems (1), (2), and (3) above, we designed Behavioral Motivation (BM) in [16] as a basic unit to control behavior flow and motivation flow. A BM is designed as a valued connector associated with Perception Filters³ (PF) and Action Pattern (AP)⁴, where a PF is associated with the promotion of motivation. BMs are organized into a flexible hierarchical network, and they have an activation value that depends on the values from a PF, a parent node, and releasers. A releaser is a special type of PF that has a value of 0 or 1. It is used to pass or to block the activation value of a BM. A BM outputs its value to its child node while the relevant stimulus is incoming through the corresponding PF.

However, a robot equipped with a motivation-based action selection mechanism in [16] cannot proactively select an action; it can select only situation-adequate and goal-oriented actions because a BM can receive values only from a PF and the BM on the parent node. For proactivity, we designed a new computational module called a predictor that proactively outputs a value to select an action in advance of an action-triggering event. Figure 2 illustrates an activation-value in a BM. The activation value is accumulated through the path while

the relevant stimulus is presented by the PF. The releasers block the flow of activation values. The value of a BM is given as

$$V_i^{BM} = \left((1 - \varepsilon) V_j^{BM} + V_{PF} + \sum_{k=1}^N V_k^{Pr} \right) \prod_{l=1}^M V_l^{Re}, \quad (1)$$

where V_i^{BM} is the value of BM i ; ε is a small positive value to discount the value of the BM at the parent node; V_j^{BM} is the value of BM j , that is a parent node of BM i ; V_{PF} is the value of a PF that filters an action-triggering stimulus in BM i ; V_k^{Pr} is the value of a predictor that predicts an action-triggering stimulus from stimulus k ; and V_l^{Re} is the value of releaser l . The most appropriate BM to accomplish a task will be selected by choosing the maximum-valued BM.

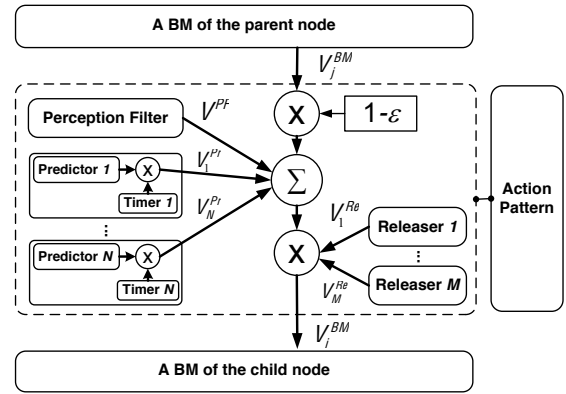


Figure 2. Activation value in BM.

For proactivity, we should consider not only the action-triggering stimulus that define a PF value but also neutral stimuli. A neutral stimulus is not an action-triggering event, but it has a causal and temporal relation with the action-triggering event. Without the predictor, however, the neutral stimulus does not affect the action selection because the PF does not output a value when observing a neutral stimulus. To predict the occurrence of a future action-triggering stimulus from any neutral stimuli, two predictions are necessary: when the stimulus will present and the probability of the stimulus. Two additional computational modules in Figure 2 perform these two predictions: a predictor and an internal timer. The predictor calculates the probability of an action-triggering stimulus from a neutral stimulus, and the internal timer predicts when the action-triggering stimulus will occur from observation of the neutral stimulus.

The value of the predictor V_k^{Pr} can be given by

$$V_k^{Pr} = P(s_i | s_k) \Theta(t_k), \quad (2)$$

where k is a neutral stimulus; $P(s_i | s_k)$ is the conditional probability of an action-triggering stimulus i given the neutral stimulus k that is the output of a predictor; t_k is the elapsed time from occurrence of stimulus k ; and $\Theta(t_k)$ is a temporal function that outputs a value when it is time for the proactive action. A detailed explanation of $P(s_i | s_k)$ and $\Theta(t_k)$ will be given in the next section.

To select proactive and goal-oriented actions, BMs with predictors are hierarchically composed as a tree structure. An example of a BM

³ PF is a function to compute the degree of matching for an action-triggering stimulus from sensory information. The PF value gives the strength of the action-triggering stimulus

⁴ An action pattern is a sequence of primitive actions.

tree in a coffee-delivery task is shown in Figure 3. A robot has to serve coffee or a meal when a human generates the action-triggering event, order-meal or order-coffee. In the right section of Figure 3, labeled Goal-Oriented Action-selection, situation-adequate actions can be generated. For example, when a human orders coffee, the robot will approach a table when a human sits there, then it will receive a coffee and deliver the coffee to the table. During the task execution, any exceptional case also be handled by reactive planning.

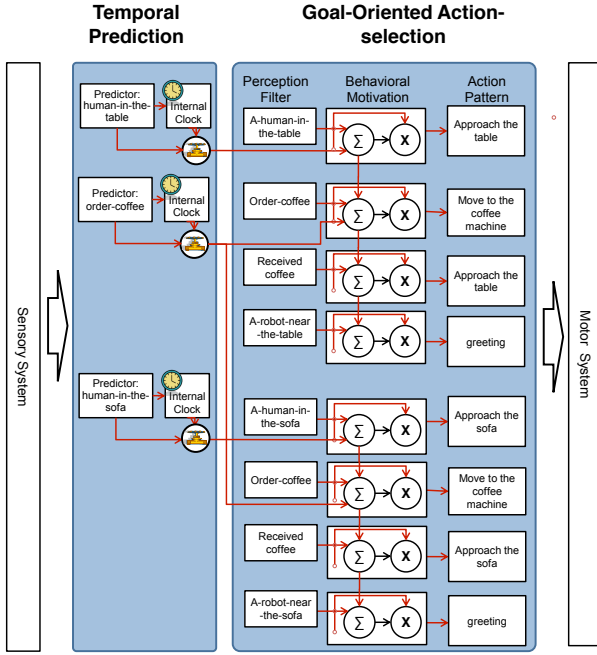


Figure 3. Action selection architecture taking into account both goal-orientedness and proactivity.

As another example, we can assume that the robot will deliver coffee when a human orders coffee immediately after entering the room through the door. Although the human does not sit at a table, the next sequence can be generated by selecting the maximum valued BM. However, these situation-adequate actions keep the human waiting, so the robot seems stupid. By using predictor in the left section of Figure 3, labeled Temporal Prediction, proactive actions can be generated. Before ordering coffee or a meal, the human passes some site such as a kitchen, desk, or table. When the human passes these sites, the robot can receive the neutral stimuli corresponding to the movement of the human. By predicting an action-triggering stimulus from these neutral stimuli, the robot can execute proactive actions to minimize the waiting time.

3 Probabilistic Temporal Prediction of Action-triggering Stimulus for Proactivity

To predict the time and probability of an action-triggering stimulus when observing neutral stimuli, we need to know the a causal probability of $P(s_i|s_k)$ and the best action-time of $\Theta(t_k)$ as shown in

(2). Many robotic applications are complex; the robot is in a multi-stimulus situation, and the temporal relations between an action-triggering stimulus and the neutral stimuli are stochastic. Therefore, a probabilistic prediction method is necessary to learn the causal and temporal relations among multiple stimuli. It is more difficult to find these relations when there are multiple stimuli rather than a single deterministic stimulus. Figure 4 shows the causal relations between an order-coffee stimulus and neutral stimuli in the delivery task of Figure 3.

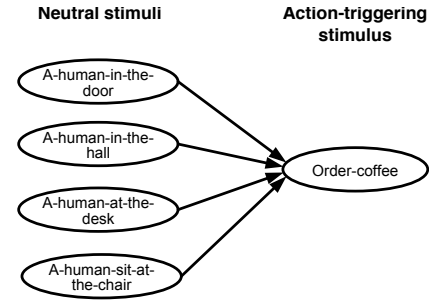


Figure 4. Prediction of an action-triggering stimulus from many neutral stimuli.

Firstly, it is necessary to obtain causal relations between the action-triggering stimulus and neutral stimuli, represented as $P(s_i|s_k)$. These relations are obtained by computing the degree of proximity among those stimuli. In animal timing theories, many researchers have proposed causal relations among many stimuli. In rate estimation theory, Gibbon [17] used a ratio of an overlapping time to the total observation time of the stimulus. If two stimuli are observed at the same time, the ratio will be close to 1. If the overlapping time of two stimuli is small, the ratio will be close to 0. In practical robotic applications, however, there are many relations among stimuli without overlapping time. To compute the degree of proximity among stimuli, we use the concept of a stimulus trace as shown in Figure 5.

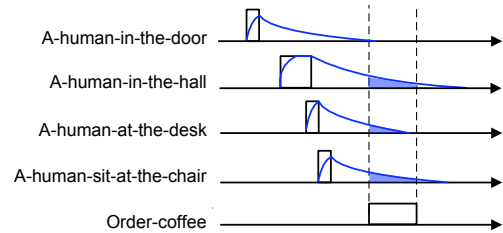


Figure 5. Example of stimulus trace.

The relation between a neutral stimulus and an action-triggering stimulus can be obtained from the degree of temporal contiguity between the two stimuli. If the neutral stimulus disappears before an action-triggering stimulus starts, the two stimuli are not present at the same time. Thus, some form of internal stimulus trace of a preceding stimulus is required as shown in Figure 5. If the difference in temporal contiguity between a preceding stimulus and a target stimulus

is small, the overlapping area between the two stimuli is large. Thus, the relation between a human-sitting-on-the-chair and order-coffee is stronger than the relation between a human-detected-at-the-door and order-coffee. In computational models for classical conditioning, many methods to model a stimulus trace have been introduced [18]. We use

$$P(s_i|s_j) = \frac{\int_{t_j=0}^{\infty} Tr_j(t_j)S_i(t_j)dt_j}{\sum_{all k t_j=0}^{\infty} \int Tr_j(t_j)S_i(t_j)dt_j} \quad (3)$$

as the stimulus trace, where $Tr_i(t)$ is defined as

$$Tr_i(t + \varepsilon) = (1 - \delta)Tr_i(t) + \delta S_i(t), \quad (4)$$

and

$$S_i(t) = \begin{cases} 1 & \text{stimulus i present} \\ 0 & \text{stimulus i absent} \end{cases} \quad (5)$$

Here, δ is a real value between 0 and 1 related to the slope of the trace, ε is an interval of sampling time, S_i is an action-triggering stimulus, and S_j is a neutral stimulus.

The next issue is to find the best time for the proactive action. For temporal prediction of the stimulus, we firstly define a temporal random variable T_A for stimulus A. The T_A is a random variable of the elapsed time after stimulus A was observed. Using this temporal random variable, we use the conditional Probability Density Function (PDF) of an action-triggering stimulus A given a neutral stimulus B as $f_{T_A|T_B}(T_A|T_B)$. The conditional PDF can be defined only when stimulus A will eventually present after B is presented. This conditional PDF can be represented as a PDF with another random variable:

$$f_{T_A|T_B}(T_A|T_B) = f_{T_{AB}}(T_A - T_B), \quad (6)$$

where T_{AB} is a relational temporal random variable of A given B. Here, the Normal distribution is used to model the conditional PDF between an action-triggering stimulus and a neutral stimulus. If an action-triggering stimulus is influenced by two or more neutral stimuli, the conditional probability of the action-triggering stimulus given all combinations of its parents is required. However, modeling temporal distributions which involves modeling PDFs in n dimensions is difficult, and the computational complexity becomes too great. Therefore, we need to use parametric modeling of the temporal conditional probabilities to avoid the complexity of modeling and inference. In this paper, the temporal probability of a conditional event in a specific time interval can be modeled as the weighted sum of two or more PDFs.

Using these modeling methods, the temporal probability of a conditional stimulus can be defined as

$$f_{T_A|T_{B_1}, \dots, T_{B_N}}(T_A|T_{B_1}, \dots, T_{B_N}) = \sum_{i=1}^N \left(P(A|B_i) f_{T_{AB_i}}(T_A - T_{B_i}) \right), \quad (7)$$

where, $P(A|B_i)$ is obtained using (3). By using the PDF of the action-triggering stimulus, we can calculate the best time for proactive action. Here, $f_{T_{Pa}|T_A}$ represents a PDF of proactive action given an action-triggering stimulus to minimize the waiting time, and $f_{T_A|T_B}$ represents a PDF of an action-triggering stimulus given neutral stimuli. If we model these two PDFs, the robot can proactively select an action. In this situation, a human waits for the robot to serve. The waiting time of the human is given by $T_{Hw} = T_{Pa} - T_A$ and the waiting time of the robot is given by $T_{Rw} = T_A - T_{Pa}$. In this case, T_{Rw}

is 0. Thus, T_{Hw} is always larger than T_{Rw} . The expectation of the waiting time of the human is given by

$$E[T_{Hw}] = \int_{-\infty}^{\infty} t_{Hw} f_{T_{Hw}}(t_{Hw}) dt_{Hw}, \quad (8)$$

and the expectation of the waiting time of the robot is given by

$$E[T_{Rw}] = \int_{-\infty}^{\infty} t_{Rw} f_{T_{Rw}}(t_{Rw}) dt_{Rw}. \quad (9)$$

The best time for proactive action is given by minimizing the sum of these two expectations $E[T_{Hw}] + E[T_{Rw}]$. Using only a Gaussian function and a Dirac delta function for PDFs, the expectation for the waiting time is given by

$$\begin{aligned} E[T_{Rw} + T_{Hw}] &= 2 \int_{-\infty}^{\infty} t_{Hw} f_{T_{Hw}}(t_{Hw}) dt_{Hw} \\ &= 2 \int_{-\infty}^{\infty} t_{Hw} f_{T_A} \star f_{T_{Pa}}(t_{Hw}) dt_{Hw} \\ &= 2 \int_{-\infty}^{\infty} t_{Hw} N(t_{Hw}; \mu_A - \mu_{Pa}, \sigma_M^2 + \sigma_{Pa}^2) dt_{Hw}, \end{aligned} \quad (10)$$

where μ_{Pa} is the mean of the PDF of the ending time of the approach action, and $\star$ is a cross-correlation operator⁵. Therefore, the best action time is $t = \mu_A - \mu_{Pa}$. Finally $\Theta(t)$ is given by

$$\Theta(t) = \begin{cases} 1 & t \geq \mu_A - \mu_{Pa} \\ 0 & t < \mu_A - \mu_{Pa} \end{cases} \quad (11)$$

4 Experiment

4.1 Experimental Setup

In this section, we will experimentally evaluate the usefulness of the proposed proactive action selection method using a robot-serving scenario. Figure 6 illustrates an experimental environment used for a serving task by a mobile robot, and Figure 7 illustrates the movements of a human and a robot during the experiment. Figure 7(a) presents a path in which the human selected the action of approaching the table and in Figure 7(b) the human selected the action of approaching the desk.

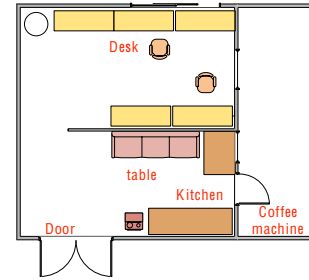


Figure 6. Experimental environment

We used a camera with wide viewpoint to capture the movement of the human, and a Pioneer 3-DX robot to serve coffee or a meal. At the start of the experiment, the robot is located near the desk. Next, the human enters the room through a door. The preferred paths of the human are as follows:

1. door → kitchen → table → table → order a meal
2. door → kitchen → table → table → order coffee
3. door → desk → order coffee

⁵ The cross-correlation of f and g is defined as $(f \star g)(t) = \int_{-\infty}^{\infty} f(\tau) \cdot g(t + \tau) d\tau$

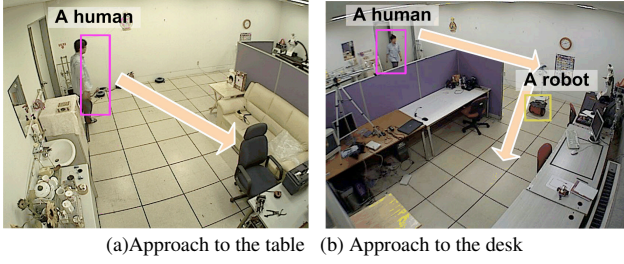


Figure 7. Experimental setups.

4.2 Experimental Results

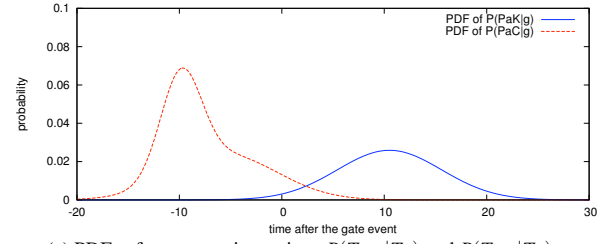
Cause	Effect	μ	σ	P_{AB}
door	kitchen	4.57	1.61	0.5
kitchen	table	7.44	1.73	1.0
table	order a meal	10.8	5.86	0.5
table	order coffee	7.23	4.53	0.5
door	desk	9.26	1.12	0.5
desk	order coffee	5.58	1.65	1.0
order a meal	approaching the kitchen	12.3	1.31	1.0
order coffee	approaching the coffee machine	24.7	2.19	1.0

Table 1. Temporal and static evidence from the experimental data.

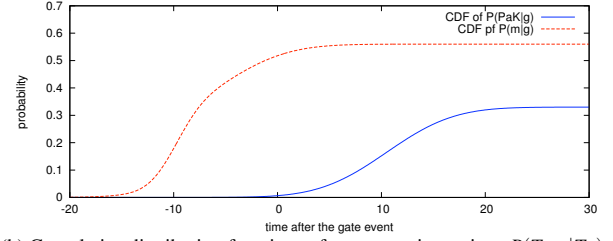
For probabilistic model for the predictor, we must provide CPTs and PDFs for the inference of causal probability as well as temporal probability. In the experiment, we collected the time interval data between events for every trial. Each trial is terminated when the human is successfully served a meal or a cup of coffee by the robot. We made CPTs by counting the frequency of conditional events and learned PDFs for the temporal interval of conditional events for every trials; means and variances of Normal distribution were computed by using the expectation maximization method.

The temporal and causal data is shown in Table 1. From these data, the causal probability and best time for predictors can be obtained. predictor value of an order-coffee given a-human-at-the-door stimulus is 0.57, and for an order-meal given a-human-at-the-door stimulus, it is 0.33. Moreover, their conditional PDFs are given as follows: the PDF of move-to-the-coffee-machine is $0.3N(t; -5.49, 5.56) + 0.26N(t; -9.86, 2.96)$, and the PDF of move-to-the-kitchen is $0.33N(x; 10.51, 6.45)$. The cumulative distribution functions and PDFs of these two proactive actions are shown in Figure 8.

A time-line expression of all stimuli and actions when a predictor does not apply to the action selection mechanism are shown in Figure 9(a). Here, a-human-at-the-door does not have any effect on the action selection. Figure 9(b) is a time-line expression of all stimuli and actions when predictors do apply to the proactive action selection mechanism. After the onset of the a-human-at-the-door stimulus, the predictor of order-coffee outputs a value for proactivity. Given the value from the predictor, a BM having an action of move-to-the-coffee-machine as a proactive action will be selected. Without proactive action selection, the average overall task-execution time is 71s with proactive action takes 60s. This result shows that the predictor

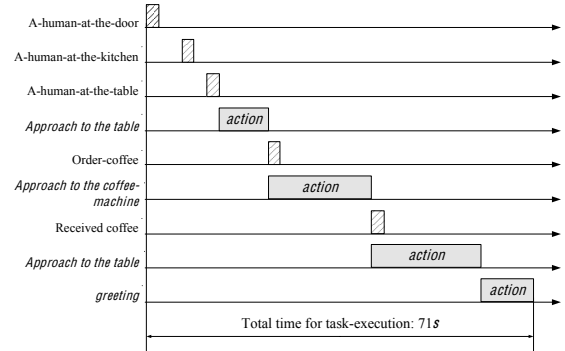


(a) PDFs of two proactive actions; $P(T_{PaK}|T_G)$ and $P(T_{PaC}|T_G)$.

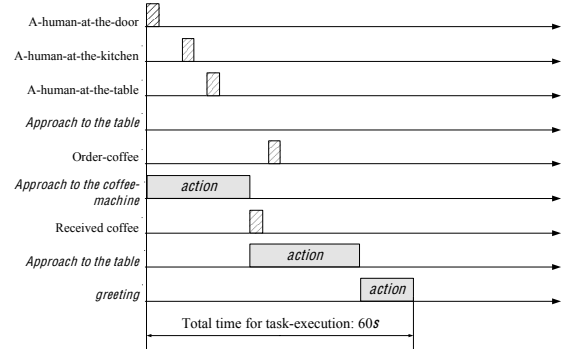


(b) Cumulative distribution functions of two proactive actions; $P(T_{PaK}|T_G)$ and $P(T_{PaC}|T_G)$.

Figure 8. Comparison of two proactive actions given a human-at-the-door event.



(a) Without proactive actions.



(b) With proactive actions.

Figure 9. Temporal expression of stimulus and action-pattern consisting of delivery task.

has a role in reducing the time taken to accomplish a task.

5 Conclusion

For intelligent interactions between a human and a robot, both situation-adequate action and proactive action are necessary. In this paper, we have proposed smart action selection architecture taking into account both goal-orientedness and proactivity that can be used in real robotic applications. By using the proposed method, a robot can predict a future action-triggering stimulus from the currently observed stimuli; by using these predictions, the robot can select a proactive action to minimize both robot and human waiting time. Our experimental results showed that the proposed action selection method reduced the waiting time in a serving task by anticipating a future situation.

ACKNOWLEDGEMENTS

This work was supported by the Technology Innovation Program of MKE/KEIT [KI001849, Development of Context Adaptive Cognition Technology].

REFERENCES

- [1] M. J. Huber, S. Kumar, and D. McGee, "Toward a suite of performatives based upon joint intention theory," *Lecture Notes in Computer Science*, vol. 3396, pp. 226–241, 2005.
- [2] Y. Hoshino, T. Takagi, U. Di Profio, and M. Fujita, "Behavior description and control using behavior module for personal robot," in *Proc. IEEE International Conference on Robotics and Automation ICRA '04*, vol. 4, Apr 26–May 1, 2004, pp. 4165–4171.
- [3] D. Tsankova, "Emotional Intervention on an Action Selection Mechanism Based on Artificial Immune Networks for Navigation of Autonomous Agents," *Adaptive Behavior*, vol. 17, no. 2, p. 135, 2009.
- [4] B. M. Blumberg, P. M. Todd, and P. Maes, "No bad dogs: Ethological lessons for learning in hamsterdam," in *From Animals to Animats, Proceedings of the Fourth International Conference on the Simulation of Adaptive Behavior*, 1996.
- [5] J. Jaafar and E. McKenzie, "A fuzzy action selection method for virtual agent navigation in unknown virtual environments," *Journal of Uncertain Systems*, vol. 2, no. 2, pp. 144–154, 2008.
- [6] E. Hogewoning, J. Broekens, J. Eggermont, and E. Bovenkamp, "Strategies for affect-controlled action-selection in Soar-RL," *Lecture Notes in Computer Science*, vol. 4528, pp. 501–510, 2007.
- [7] J. Bryson, "Mechanisms of action selection: Introduction to the special issue," *Adaptive Behavior*, vol. 15, no. 1, p. 5, 2007.
- [8] R. Brooks, "A robust layered control system for a mobile robot," *IEEE Transactions on Robotics and Automation*, vol. 2, no. 1, pp. 14–23, Mar. 1986.
- [9] P. Capdepuy, D. Polani, and C. Nehaniv, "Construction of an internal predictive model by event anticipation," *Lecture Notes in Computer Science*, vol. 4520, p. 218, 2007.
- [10] C. Balkenius, "Dynamics of a classical conditioning model," *Autonomous robots*, vol. 7, no. 1, pp. 41–56, 1999.
- [11] R. Sutton and A. Barto, "Time-derivative models of Pavlovian reinforcement," *Learning and computational neuroscience: Foundations of adaptive networks*, pp. 497–537, 1990.
- [12] R. Burke, "It's about time: Temporal representations for synthetic characters," Ph.D. dissertation, Massachusetts Institute of Technology, 2001.
- [13] C. Gallistel and J. Gibbon, "Time, rate, and conditioning," *PSYCHOLOGICAL REVIEW-NEW YORK*, vol. 107, no. 2, pp. 289–344, 2000.
- [14] W. Y. Kwon and I. H. Suh, "A Motivation-Based Action-Selection-Mechanism Involving Temporal Prediction," in *International Conference on Service and Interactive Robotics, SIRCon 2009*, 6–7 August 2009.
- [15] R. Sutton and A. Barto, "Time-derivative models of Pavlovian reinforcement," *Learning and computational neuroscience: Foundations of adaptive networks*, pp. 497–537, 1990.
- [16] S. Lee, I. Suh, and W. Kwon, "A Motivation-Based Action-Selection-Mechanism Involving Reinforcement Learning," *International Journal of Control, Automation, and Systems*, vol. 6, no. 6, pp. 904–914, 2008.
- [17] C. Gallistel and J. Gibbon, "Time, rate, and conditioning," *PSYCHOLOGICAL REVIEW-NEW YORK*, vol. 107, no. 2, pp. 289–344, 2000.
- [18] C. Balkenius, J. Morén, and L. Kungshuset, "Computational models of classical conditioning: a comparative study," in *From Animals to Animats 5: Proceedings of the Fifth International Conference on Simulation of Adaptive Behavior*. MIT Press, 1998, pp. 348–353.

Social, functional, and problem-related tasks in HRI - a comparative analysis of body orientation and gaze

Manja Lohse¹

Abstract. This paper analyzes human-robot interaction in a home tour scenario with respect to the tasks that the users had to solve in a study. The analysis shows that the users' behavior regarding body orientation and gaze depends on the task which in the present scenario can be social, functional, or problem-related. The paper argues that knowledge about these variations should be taken into account when designing robots in order to improve their capability for situation awareness and, thus, to improve the interaction.

1 INTRODUCTION

The goal to let robots interact with inexperienced users poses grand challenges in human-robot interaction (HRI) today. This is because the users cannot in general be expected to follow concrete plans or to stick to the way of interaction that was foreseen by the developers. However, the theory of situated actions claims that context shapes actions and provides people with means to interpret and understand actions [1]. It is here assumed that tasks in HRI provide one specific kind of context. Therefore, it can be hypothesized that the users' behavior is somewhat similar for specific task whereas it varies between tasks. With this in mind, the current paper investigates the tasks that arise in a home tour scenario. The home tour focuses on teaching rooms and objects to the robot in a domestic environment in order to enable it to solve, for example, fetch-and-carry tasks. In this scenario, the users are likely to be novices or advanced beginners, meaning that they have little or no experience interacting with a personal service robot for the home. In fact, in the user study presented in the following most of the trials were first-contact situations such that the users had never interacted with a robot before.

The main research question in the paper is how the inexperienced users behave during different tasks. To answer this question, the interaction is analyzed applying task analysis (TA) methods. Within the tasks, the paper focuses on the modalities body orientation and gaze. Both these modalities will show whether a differentiation of tasks based on user behavior is actually possible. If it can be found that the users' behavior depends on the tasks, the knowledge about how the behavior varies could be implemented in the robot in order to improve its ability to determine what the users are trying to do. In other words, knowledge about typical user behaviors can be used to improve the robot's situation awareness.

2 RELATED WORK

Situation awareness (SA) describes the ability of the robot to find out what is going on in an interaction [2]. In other words, the robot needs to perceive the elements of the environment and to know their meaning. The elements can be analyzed on different levels of specificity of the situation [3]. These levels are (from most general to most specific) situations in general; HRI situations that are characterized by the actors (human and robot); scenarios which determine the roles of the actors (in the home tour the human is the teacher and the robot is the learner); the tasks with specific goals and goal structure; and specific situations (one particular situation). The current paper analyzes elements on the task level by subsuming different specific situations observed in a user study in certain tasks. It looks for similarities within the tasks and differences between the tasks of the home tour. As a preparation for this, the tasks need to be identified with the help of task analysis (TA). TA has been used in market research, product testing, and also in human-computer interaction for a long time. User and task analysis focus on understanding how users perform their tasks [4]. They provide a picture of the system from the perspective of the user, which can be used to improve task completion and to integrate the human into the system design. TA helps to generate hypotheses about sources of inadequate performance of system and operator [5]. In recent years, TA has been introduced to HRI; for example, in the context of situational awareness and user roles [6]. TA has been proposed as a means to identify user's work procedures and tasks when interacting with a mobile office robot, the physical design requirements, function allocation and the relation of the work between user and robot, and user's expectations [7]. The authors approached these goals through interviews and focus groups. In a similar vein, TA has been applied for system design in HRI [8]. None of these authors has focused on TA in the evaluation process, and no data-driven approaches based on user studies have been proposed so far. However, TA is also useful in these respects as the following analyses will show.

Once the tasks are derived, the user behaviors within them need to be analyzed in order to find differences and similarities between the tasks. The user behavior consists of several verbal (speech) and non-verbal (gesture, gaze, spatial behaviors, touch, etc.) modalities which could be used for analysis [9]. Speech is not taken into account here because the users received training on utterances that the robot could understand well. From the non-verbal modalities, body orientation and gaze were chosen because both can actively be produced by the robot and are very important for the interaction.

Spatial behavior in general consists of proximity, orientation, territorial behavior, and movement in a physical setting [9]. All these aspects have been in the focus of the social sciences, often with respect to the influence of gender, age, and culture (for an

¹ Applied Informatics, Bielefeld University, Universitätsstrasse 25, Bielefeld, Germany. Email: mlohse@techfak.uni-bielefeld.de

extensive overview see [10], [11]). In HRI so far especially proximity (spatial distance) was taken into account.

Spatial distances were one main research activity in the EU project Cogniron that focused on “observing a correct attitude towards humans, especially with respect to interaction distances and personal spaces” [12]. In this context, [13] differentiated situations with no interaction, verbal interaction, and physical interaction. In a study they found that the participants wanted the robot to keep the greatest distance when they were not interacting with the robot and allowed it to come closest when interacting physically. The authors argue that in the latter situation the robot’s purpose explains the closer approach. In a similar vein, [14] researched situational differences between human preferences for robot approaches in four different scenarios: human seated or standing in the middle of an open space, seated at a table in the middle of an open space, and standing with the back against a wall. All situations are typical for fetch-and-carry tasks. It was found that front left and right approaches were preferred to direct frontal approaches and to rear approaches. Moreover, the participants felt more comfortable with the robot coming closer when they were in open spaces compared to being with their backs against a wall. These results show that contextual and situational factors actually play a role in HRI.

Body orientation also has been of interest for other researchers in HRI. For example, the influence of agent body orientation on task completion was evaluated [15]. However, this work is restricted by the fact that only two orientations were distinguished (facing the user / back towards user). Also Kendon’s F-Formation system [16], which is much more natural in interaction, has been discussed [17]. It is based on the observation that people usually group themselves in certain formations when interacting. F-formation describes the kind of formation when people form a shared space between them which they can all access equally well.

Another important modality in social behavior is gaze [9]. The direction of gaze indicates the direction of attention [18]. Thus, where a person looks is crucial. Even though gaze is an important signal for attention, speakers look intermittently at their listeners [9]. One reason for frequent gaze shifts is that looking at the listener all the time could lead to cognitive overload or the arousal caused by the gaze could infer with the cognitive planning. The amount of gaze also varies with the task [9]. For example, tasks like object-teaching require joint attention towards the object [19]. In other words, the interactors look less at each other when there are other things to look at, especially when there is an object of joint attention. Therefore, gaze is important for the comparison of behaviors between tasks.

Some work has also been conducted on gaze in HRI. An experiment was carried out focusing on the production of robot gaze in a scenario where the robot pointed out objects to the participants either correctly or falsely [20]. The measurement of the reaction time of the human depending on the gaze direction and the speech of the robot showed that in the case that the robot’s gaze or speech behaviors were inappropriate, the interaction slowed down measurably in response time and fixation distribution. Also others found that non-verbal cues such as gaze convey the robot’s internal states which helps to improve human task performance, to prevent and to repair errors [21].

Another study reports how the robot Mel gazed at the participants [22]. In the data, shared looking was coded (mutual

gaze and both look at the same object) and a positive relationship between shared looking and engagement of the interactors was found. With respect to the home tour scenario, gaze has been observed to be one mean to establish contact [23]. Users ‘look for’ feedback in order to find out whether a command has been received correctly by the robot. Therefore, gazing at the robot indicates that some kind of feedback is necessary. It can be concluded that the system should provide continuous information on its status of availability for communication [23]. This status can be provided with the gaze of a camera because just like a human eye the camera behavior of the robot can be used to grab the user’s attention and to make contact. Also the robot BIRON exploits this modality as will be described in the following section.

3 ROBOT SYSTEM

In the home tour scenario a domestic robot has to be employed that can be defined “to be a machine that (a) is designed to work with individuals and groups in their personal and public spaces, (b) has a dynamic spatial presence in those spaces, and (c) can “intelligently” interpret its environment and interact physically with it” [24].

BIRON was developed to have these abilities (see Figure 1). The robot is based on an ActiveMedia™ Pioneer PeopleBot platform. It is equipped with a pan-tilt color camera at a height of 141 cm to acquire images of human faces. The camera is also used to show attention by looking in a certain direction and to actively explore the environment visually.

Below the camera, there is a screen which can be used to display the behavioral states of the robot with an animated character called Mindi. The character resembles BIRON. The Mindi pictures are directly related to the robot state, for example, if the robot is in the follow state, an animation is shown in which the Mindi has a happy face and is also walking; if the user is not perceived well, the face of the Mindi character shows an apologetic expression and a speech bubble appears next to the Mindi in which a picture of a fuzzy stick figure is displayed.

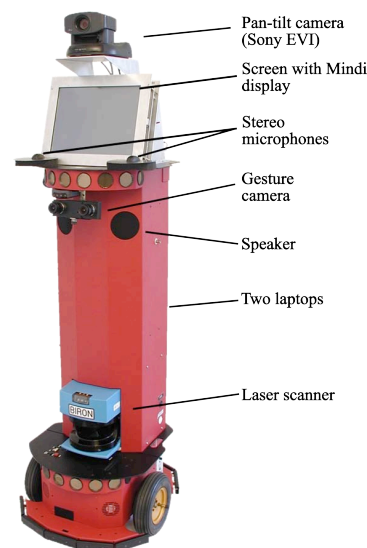


Figure 1. BIRON (Bielefeld Robot companion)

A pair of AKG far-field microphones is located right below the display at a height of approximately 106 cm. With their help, BIRON can localize speakers. For speech recognition itself, the user wears a headset. Moreover, below the microphones there is an extra camera which is used to recognize gestures, and two speakers for speech output. A SICK laser range finder, mounted at a height of 30 cm, measures distances within a scene.

BIRON is able to follow a person and to autonomously move around. Additionally, it can track people and selectively pay attention to humans looking at it. Further technical information about BIRON is provided in [25].

4 USER STUDY

An exploratory user study was conducted in order to answer the question how different tasks are solved by the users in a home tour scenario. The information gained in the study shall shed light on the question of how the context provided by the tasks influenced the users' behavior with respect to body orientation and gaze.

14 participants (9 male, 5 female; between 16 and 71 years old, mean age 47 years) were invited. Their task was to guide the robot from the living room to the dining room (see Figure 2) and to teach it the book shelf in the living room and the floor lamp in the dining room. All actions that were necessary to conduct this task were practiced with the help of a training sheet beforehand. Thus, training effects were caused on purpose because the participants were completely inexperienced with respect to HRI and needed some basic information in order to decrease insecurity and hesitant behaviors. However, after training no systematic changes of user behaviors were noticed throughout the trials. With the basic information that the participants received, 13 of them completed the whole task. One participant could not finish because of a system breakdown. The completed sessions took between 15 and 33 minutes (average 20 minutes including training).

The whole interaction was videotaped by one of the experimenters. The videos were then used to annotate body orientation and gaze behavior of the users manually. The coding schemes are introduced and validated in [3]. The coding of gaze was very simple, only differentiating three gaze directions (at the robot, at the object in question, somewhere else). In contrast, the coding of body orientation was much more complex. It distinguished static body orientations (neither robot nor human move) and dynamic changes of orientation (either human, robot, or both move). These orientations refer to human and robot. Also the orientation towards the objects of interest was coded based on [26] (see Figure 3). This coding scheme includes one more interval to each side in front of the human than Kendon's F-Formation (codes 1 and 7). Moreover, it is more flexible because it allows to code objects which are behind the human.

The static and dynamic orientations with respect to human and robot used the same intervals as the coding of object positions. However, they also took the facing directions of both interactors into account. While this differentiation might play a minor role in HHI, it is assumed to be important for HRI because of the asymmetry between behaviors that are performed by the human and the robot. Moreover, it was found to be necessary to consider the facing direction of both the human and the robot for two reasons. From the point of view of the robot, the facing direction is crucial because the robot loses the users as soon as it

cannot perceive them anymore and the users need to repair this problem before the interaction can continue. Even though the robot is not (yet) programmed to adapt its body orientation to certain tasks based on a social model of the interaction, the teaching of locations requires it to turn around to improve its representation of the environment. From the point of view of the user, the facing direction is important because it indicates the users' expectations about the robot's abilities and their trust in what the system is doing. The users will most probably closely monitor the robot not turning the back towards it if they have doubts about its ability to follow without doing any harm. These considerations resulted in a coding scheme with 23 different codes for static orientations, changes of which were used to code dynamic orientations (see [3]).

The videos were coded with these schemes. If sequences were occluded or unclear they were not taken into account. Reliability of the coding schemes was calculated for two raters to make sure that the video data was suitable to code the modalities and that the coding schemes were usable. The reliability was found to be satisfactory [3].

After having coded body orientation and gaze behavior manually, the annotations were analyzed statistically with the help of SALEM (Statistical Analysis of Elan files in Matlab), an approach to do statistical analysis based on annotations in the software Elan ([3], [27]).

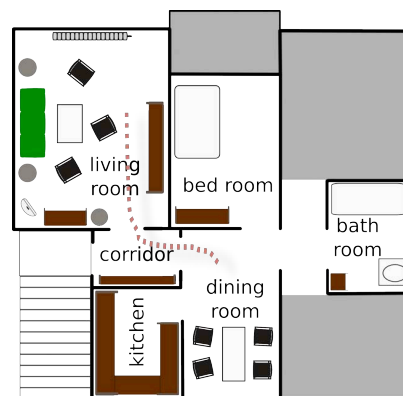


Figure 2. Apartment and path the robot had to be guided

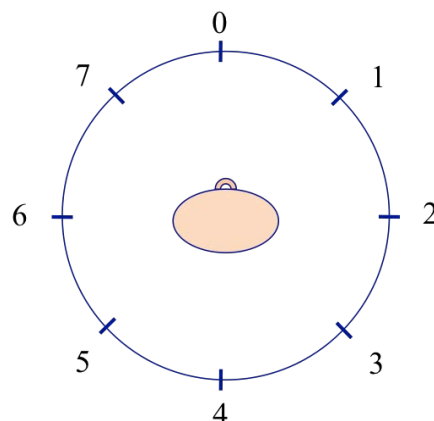


Figure 3. Coding scheme for object position (human in the center; numbers describe object positions)

5 TASK ANALYSIS

In the data of the study, 393 interaction sequences were coded that belong to ten tasks (see Table 1). Each task consisted of certain actions; for example, greeting consisted of the user uttering a greeting which was answered by the robot that also said hello; teaching objects consisted of the user pointing out and naming the object and the robot recognizing the object. Tasks were also coded as such if the user showed a certain intention which the robot did not understand, for example the user greeted the robot which did not reply.

After identifying the single tasks they were grouped according to their function in the interaction. The first group is social tasks. Greeting, intro, and farewell belong to this group. They frame the interaction and are mainly used to manage the robot's attention. On first sight, one could think that greeting and intro could be subsumed into one task. However, they were separated because not every user listened to the intro of the robot, and, more importantly, greetings were also carried out during the interaction to regain the robot's attention while the intro really only occurred after the first greeting. In contrast to the social tasks, guiding, teaching objects and rooms are functional tasks that constitute the main part of the interaction. The functional tasks are mandatory to reach the goal of the interaction. The third group subsumes the problem-related tasks. As the name indicates, these are problems that can take place during the interaction. 'Obstacle' characterizes a situation when the robot cannot move any further because something is in its way. 'Register' describes the state when the robot has lost the person and the user needs to register again by saying hello. The 'reset' task allows the user to reset the robot to a consistent state when it does not react anymore. The 'stop' task is initiated by the robot if the users need to say stop in order to finish a previous task. All these problems, in contrast to mere technical problems, can be solved by the users on the interaction level. The four problem-related tasks could also be seen as part of the other tasks. However, they have been coded separately here for two reasons. Firstly, like the other tasks, they are situations that the user needs to solve using certain behaviors. Secondly, analyzing the problems as tasks allows them to be represented in a transition matrix with all other tasks in order to determine when they occur.

TA revealed that the social tasks took less time than the functional and even as the problem-related tasks. While the introduction of the robot only took place a few times, the greeting task and the farewell task both had very short mean and overall durations. Table 1 also shows that 28.64% of the time was spent on the resolution of problems. The register task alone took up 9.96% of the time of the interaction because it was the most frequent problem. The reset problem did not occur as often; however, it took the longest time to be repaired. This is due to two reasons: firstly, it is the most severe problem and the robot does not tell the user what is wrong and secondly, the robot replies to the reset command that the user has to wait five seconds before saying hello again to reinitiate the interaction. Thus, some time passes without anything happening.

Altogether, most time was spent on the functional tasks (61.26% of the overall duration of the interaction). This is positive because the interaction was designed for these tasks. 34.00% of the time was used for guiding. Guiding is probably the most complex task in the home tour from a technical point of view because it requires the robot to track the the user and the

environment while moving. Guiding also caused the most problem-related tasks. The guiding sequences were followed by showing the room, which is the task that should follow it according to the study design, as often as by obstacle and register (transition probability of about .21 each). The obstacle problem is directly related to the guiding of the robot. Obstacle situations mainly resulted when the participants tried to cross the door between the living room and the narrow hallway. That this problem actually interrupted the guiding of the robot can be shown with the high percentage of guiding tasks that succeeded the obstacle task (.88).

Register situations were caused by different problems. The robot lost the users because they moved too fast and too far away from the robot, the users turned too quickly, or the light changed from one room to the other. On the positive side, it can be noted that the problems were hardly ever directly followed by other or the same problems. This shows that the users were actually able to solve them on the interaction level.

Table 1. Descriptive statistics of the home tour tasks

	count	count of all annotations (%)	min. duration (seconds)	max. duration (seconds)	mean duration (seconds)	standard deviation duration	duration of all annotations (%)
social tasks							
greet	45	11.45	0.99	30.23	8.12	6.921	4.99
intro	5	1.27	19.09	57.50	41.81	18.075	2.86
farewell	18	4.58	2.41	23.15	9.13	5.608	2.24
functional tasks							
guiding	109	27.74	3.28	75.07	22.84	12.748	34.00
teaching object	66	16.79	4.14	61.55	19.50	11.583	17.58
teaching room	40	10.18	3.87	37.36	17.71	9.038	9.68
problem-related tasks							
obstacle	25	6.36	8.35	55.38	22.15	9.959	7.56
register	38	9.67	8.19	55.45	19.19	11.548	9.96
reset	18	4.58	13.26	92.14	28.15	19.394	6.92
stop	29	7.38	3.62	29.98	10.60	6.493	4.20
overall	393	100	0.99	92.14	18.63	12.719	100

6 ANALYSIS OF BODY ORIENTATION

Now that the tasks have been described, it shall be analyzed how the users behaved while conducting the tasks. For this purpose, body orientation was annotated with the coding scheme that has been introduced in Section 4. As explained there, the static and the dynamic body orientation (i.e., movements that change body orientation) were coded separately. First the results of the static orientation will be introduced.

The analysis showed that the participants turned as often to certain angles at the left side as to the same angles at the right side. Therefore, the orientations which correspond to the same angles, will be grouped for further analysis. In Table 2 it becomes very obvious that the users preferred orientations with smaller angles. They chose them more often and stayed longer in the orientations. The mean durations of the sequences between face-to-face and 45°; and 45° and 90° differed by a factor of about 2.5. Face-to-face orientations were most common which is in line with the findings of [17].

It had further been shown that the body orientation depends on the task that is solved [13]. Here the body orientation was analyzed with respect to all home tour tasks (Table 3). It is obvious at first sight that the users chose a very direct orientation in the social tasks. In more than 80% of the time they maintained a vis-à-vis orientation to the robot. In the remaining time they just turned slightly. This was also true for the problem-related tasks with just a few exceptions. Only in the functional tasks did the participants prefer other orientations toward the robot. Thus, with respect to the coding scheme, it seems that there is in fact a functional difference between the 45° and 90° orientations. While turning to 45° the users can still signal that they are very attentive. However, in the guiding task the probability of turning to a 90° orientation is bigger because the users prepare to walk away from the robot in order to trigger it to follow.

Also the orientation of the users towards the objects was evaluated. Table 4 shows the results of the analysis. As for the coding of the body orientation of the human towards the robot, the amount of times that the users turned to certain angles to their left or right agreed. Whether one or the other was chosen was determined by the fact on which side the object was located.

More objects were located to the right of the users, which was true for both the shelf and the floor lamp that most subjects showed to the robot. The angles that the users turned to in order to show these two objects were very similar. Table 4 illustrates that the object was most commonly to the side of the participant's upper body. The object was nearly as often slightly in the back of the user as it was slightly in front of them. This finding indicates that the users did not necessarily navigate the robot until it was in a position where the object was between them as assumed by Kendon's F-Formation system. Figure 4 shows the most common constellations of human, robot, and object.

Table 2. Descriptive statistics of body orientations (overlap (%) = percentage of time the value was annotated in the interaction)

body orientation	count	mean duration (seconds)	SD duration	overlap (%)
face-to-face	169	31.14	43.23	73.71
45°	128	12.21	18.67	21.89
90°	48	5.02	7.26	3.37
135°	12	2.76	1.97	0.46
180°	12	1.59	0.97	0.27
almost side-by-side	1	5.73	0	0.08
side-by-side	1	14.83	0	0.21
overall	371	19.24	33.21	100

Table 3. Body orientation in the tasks (# = count, OL (%) = overlap, percentage of time the value was annotated during the task)

task	body orientation															
	face-to-face		45°		90°		135°		180°		alm. side-by-side		side-by-side		overall	
	#	OL (%)	#	OL (%)	#	OL (%)	#	OL (%)	#	OL (%)	#	OL (%)	#	OL (%)	#	OL (%)
social tasks																
greet	36	87.55	8	8.26											44	95.81
intro	6	99.11	1	0.34											7	99.45
farewell	12	80.21	5	16.64											17	96.85
functional tasks																
guiding	107	57.89	66	13.25	35	3.22	6	0.63	9	0.56	1	0.19	1	0.52	225	76.25
object	36	46.57	30	39.55	7	7.42	2	0.53							75	94.07
room	35	72.43	11	7.16	3	0.52	1	0.16	2	0.59					52	80.86
problem-related tasks																
obstacle	32	60.87	7	3.55	1	0.44									40	64.86
register	26	60.37	21	25.49	5	2.48	1	0.36			1	0.13	1	0.26	55	89.09
reset	21	63.40	14	29.04											35	92.44
stop	16	77.71	7	14.16			1	1.40	1	0.36					25	93.63

Table 4. Orientations of the users towards the objects

orientation of the object	count
between user and robot (0)	1
slightly in front of the user (1, 7)	15
to the side of the user (2, 6)	52
slightly behind the user (3, 5)	13
behind the user (4)	0
sum	81

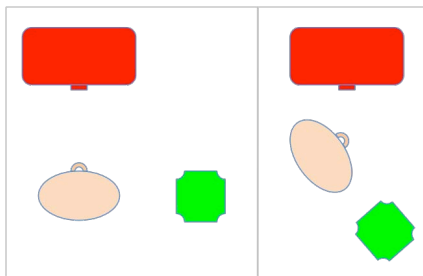


Figure 4. Typical participation frameworks of user, robot, and object in the home tour

Next to the static orientations also changes between different orientations initiated by the users were analyzed. Altogether, 262 movements with a duration of 9 minutes, 36 seconds were coded. The mean duration of the position changes was much shorter than the mean duration of the time that the users spent in a certain position (2.20 seconds vs. 19.11 seconds).

This is underlined by the finding that the users almost always chose the shortest ways between the orientations. But not only the users changed their orientation; the robot also initiated 96 movements that led to a change. This number of changes was significantly lower over all tasks than the number of movements of the users (T-test (two-tailed), $df=10$, $T=2.464$, $p=.036^*$). The robot's movements took 7 minutes, 23 seconds altogether. With a mean duration of 4.62 seconds, the movements of the robot were much slower than the movements of the human. The question now is how changes of body orientation relate to the home tour tasks.

The analysis showed that the users changed their orientation during the guiding task most often by far (3.66 movements/minute). In contrast, the robot most often changed its position in the room-teaching task (1.78 movements/minute), which is in line with the designed robot behavior that makes the robot turn around in order to improve its representation of the room. The task with the second most changes of user (1.97 movements/minute) and robot (1.48 movements/minute) was the register task. Hence, the users tried to position themselves in front of the robot during the task so that it could perceive them well and the robot turned to the user for the same reason. On the other hand, instead of positioning themselves directly in front of the robot, some users walked to the side of it to see whether it followed, which it did as the number of orientation changes of the robot shows. The users did the same during the reset task (2.01 movements/minute). The reset task was different compared to the register task because the robot did not change its orientation towards the user.

Comparing the movements per minute, the social tasks were found to be rather static (greeting 0.66, introduction 0.57, and farewell 1.10 changes of orientation/minute). The farewell task was accompanied by slightly more movements which is probably due to the fact that the users started to turn away before the robot had replied to the farewell command.

To conclude, the tasks could be differentiated based on the orientation of the users. The users most often changed their orientation in the guiding task which was also due to the fact that the robot was moving. Only during guiding it was at times found that the users turned their backs to the robot. But as has been mentioned above, the time they spent in this orientation was very short compared to the time they spent in more direct orientations. The object-teaching task was the only task during which the participants turned as many times to 45° and 90° (both summed up) as to a vis-à-vis orientation and also spent the same amount of time in these orientations. The finding underlines that the participants turned when teaching objects to establish a participation framework that enables the robot to optimally perceive the objects which are usually located on the side of the participants. This was not found for teaching rooms where the users most of the time faced the robot. Therefore, the actions of teaching rooms and objects seem to follow distinct rules and need to be differentiated.

Altogether, the static and dynamic body orientation have been shown to differ between the tasks and can, therefore, be used to distinguish tasks from each other and to improve the SA of the robot. The following analysis will show if gaze further contributes to the distinction of the tasks.

7 ANALYSIS OF GAZE

Altogether, 2 hours, 13 minutes, and 20 seconds of gaze were analyzed. In 90.7% of the time (491 times) the participants looked at the robot, in 2.7% (159 times) at the objects, and in 6.5% (264 times) somewhere else (see Table 5).

The overall looking time at the robot and the mean duration of glances (14.78 seconds) were much longer than in HHI [9]. Even though the mean duration was strongly influenced by some very long glances, the median was also very high (8.02 seconds). The durations of glances at the objects and at places other than object and robots were much shorter (mean 1.37 and 1.98 seconds respectively, median 1.05 and 1.35 seconds).

The question now is why the users looked at the robot for such a long time. One probable explanation is that they were waiting for feedback which occasionally took long in the study because the robot was acting autonomously. Actually, the longest glances were found in situations when the users were

Table 5. Descriptive statistics of gaze direction

gaze direction	minimum duration (seconds)	maximum duration (seconds)	mean duration (seconds)	standard deviation duration	median duration (seconds)	duration (%)
robot	0.11	259.53	14.78	21.58	8.02	90.7
object	0.38	20.72	1.37	1.76	1.05	2.7
else	0.18	17.74	1.98	1.87	1.35	6.5
overall	0.11	259.53	8.69	17.08	2.47	100

waiting for the robot to reply. This behavior is in line with the finding that gaze at the robot signals that the users need some feedback [23]. Moreover, after the interaction many participants stated that they looked at the screen when the robot did not respond. Since a screen is something one is supposed to look at and not have to look away from once in a while in order to not offend the interaction partner, this is the most probable explanation for the long glances. Unfortunately, from the video data that was recorded, it cannot be determined whether the participants actually looked at the screen or at some other part of the robot. An analysis of gazing behavior during the tasks was conducted to shed some more light on this question (see Table 6).

The first observation in Table 6 is that the number of glances and the percentage of looking at the objects were by far highest in the object-teaching task. This finding is in line with the design of the study and the task. The users most often looked somewhere else other than at the robot or the objects during the intro and the guiding task. During the guiding task the participants certainly looked where they wanted to go to. During the intro the frequent glances away from the robot are more surprising because the robot talks about its usage and often refers to pictures on its screen. When having a second look at the data, it was found that in fact only one user looked away from the robot six times. She turned to the side because the robot had called her by a wrong name ('Igor') and she kept laughing about the mistake. Hence, the common gazing direction during the intro really seems to be the robot. But this case also shows that the users might change their behavior crucially if some

irregularities appear in the interaction. With respect to the percentage of the time that the users in general looked at the robot, no significant differences were found. Therefore, it can be concluded that gaze direction alone which is here coded in a very simple fashion is only partially revealing. Hence, the analysis needs to integrate several modalities in order to describe user behaviors that are typical for certain tasks. In this context, in the following coherences between gaze direction and body orientation will be described.

The gaze direction corresponded with certain body orientations in the tasks. In the social tasks and the problem-related tasks the body orientation was very direct and the users looked at the robot most of the time. In the guiding task the body orientation as well as the gazing direction changed most often. The users spent most of the time looking somewhere else other than the robot and the object and keeping indirect orientations. The user behavior differed considerably between the two teaching tasks. During object-teaching the users often slightly turned away from the robot and then back towards it as they gazed at the object and then back at the robot. In contrast, during the room-teaching task they gazed at the robot and kept a direct orientation for most of the time. Accordingly, the robot could identify that something was taught based on the utterance and, furthermore, receive information about what was shown via the gaze direction and the body orientation. This finding points out that in the long run even more modalities such as speech could be included in the description of the user behavior during certain tasks.

Table 6. Gaze direction in the tasks (overlap (%) = percentage of time the value was annotated during the task)

	gaze direction							
	at robot		at object		else		overall	
task	count	overlap (%)	count	overlap (%)	count	overlap (%)	count	overlap (%)
social tasks								
greet	50	94.91	1	1.38	7	2.14	58	98.43
intro	14	90.86	0	0	9	9.13	23	100
farewell	20	94.81	0	0	4	4.48	24	99.29
functional tasks								
guide	211	82.07	19	0.90	105	8.28	335	91.25
object	153	87.81	114	9.85	18	2.34	285	100
room	60	92.40	12	1.10	26	5.27	98	98.77
problem-related tasks								
obstacle	42	64.19	0	0	8	1.78	50	65.96
register	58	91.19	2	0.17	27	6.41	87	97.77
reset	33	93.03	1	0.08	16	5.45	50	98.56
stop	30	86.41	4	4.87	7	5.71	41	96.99
overall	459	85.75	142	2.44	222	5.59	823	93.79

8 CONCLUSION

In the previous sections, TA was applied to the home tour tasks in order to evaluate HRI. It was found that the home tour includes social tasks (greet, intro, farewell), functional tasks (teaching rooms and objects, guiding the robot), and problem-related tasks (obstacle, register, reset, stop). These tasks were hypothesized to serve as a kind of context which shapes the users' actions. In order to show that the users' behaviors indeed differed between the tasks, data were analyzed with respect to the modalities eye gaze and body orientation. Coding schemes were introduced for these modalities.

The analysis of body orientation clearly showed that the behaviors indeed differed between the tasks, which allows to distinguish them from each other, for example, a lot of movement and indirect orientations indicate a guiding task while a direct orientation is more common in social and problem-related tasks. With respect to gaze, the analysis showed that the users most often looked at the robot in the social and the problem-related tasks. However, also during the other tasks they mainly looked the robot. Thus, gaze alone was only partially revealing about what the users intended to do. Therefore, it was concluded that the modalities need to be integrated in order to improve the prediction of the users' intention and, in a next step, to improve the robot's SA based on the human behavior models. Implementing knowledge about these specific behaviors of the users with respect to several modalities will help the robot to interpret and better understand the actions. In other words, knowledge about task-related user behavior can help to improve the robot's SA.

In future work, more modalities that the users employ to interact with the robot shall be integrated in the human behavior models (for example, speech and orientation of the head). Moreover, the gained knowledge cannot only be used to improve the robot's perception of the interaction but it can also be exploited to adapt the robot's own behavior to the human behavior models. In this sense, the results can help to improve HRI in general.

REFERENCES

- [1] L. A. Suchman. *Human-Machine Reconfigurations: Plans and situated actions*. 2nd edition. Cambridge: Cambridge University Press (2007).
- [2] M. R. Endsley. Design and Evaluation for situation awareness enhancement. *Proceedings of the Human Factors Society 32nd Annual Meeting*. Santa Monica, CA: Human Factors Society, 97-101 (1988).
- [3] M. Lohse. *Investigating the influence of situations and expectations on user behavior - empirical analyses in human-robot interaction*. Doctoral Thesis. Bielefeld University, Technical Faculty, Germany (submitted).
- [4] J. T. Hackos and J. C. Redish. *User and task analysis for interface design*. New York: Wiley and Sons (1998).
- [5] A. Shepherd. *Hierarchical Task Analysis*. London and New York: Taylor and Francis (2001).
- [6] J. A. Adams. Human-Robot Interaction Design: Understanding User Needs and Requirements. *Proceedings of the 2005 Human Factors and Ergonomics Society 49th Annual Meeting* (2005).
- [7] K. Severinson-Eklundh, A. Green, H. Hüttenrauch, L. Oestreicher, and M. Norman. Involving Users in the Design of a Mobile Office Robot. *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews* 34(2), 113-124 (2003).
- [8] H. Kim and D. Kwon. Task Modeling for Intelligent Service Robot using Hierarchical Task Analysis. *Proceedings of the 2004 FIRA Robot World Congress*, Busan, Korea (2004).
- [9] M. Argyle. *Bodily Communication*. London: Methuen (1988).
- [10] J. Aiello. Human spatial behavior. In D. Stokols and I. Altman (Eds.). *Handbook of Environmental Psychology 1*. New York: Wiley Interscience, 505-535 (1987).
- [11] L. Hayduk. Personal space: Where we now stand. *Psychological Bulletin*, 94, 293-335 (1983).
- [12] Website of the Cogniron project. www.cogniron.org, 24. September 2009.
- [13] D. S. Syrdal, M. L. Walters, K. L. Koay, and K. Dautenhahn. The role of autonomy and interaction type on spatial comfort in an HRI scenario. *Proceedings of "Robotic Helpers: User Interaction, Interfaces and Companions in Assistive and Therapy Robotics", Workshop at ACM/IEEE HRI 2008*, Amsterdam, the Netherlands, 12 March 2008, Technical Report 470 University of Hertfordshire, UK (2008).
- [14] M. L. Walters, K. L. Koay, S. Woods, D. S. Syrdal, and K. Dautenhahn. Robot to Human Approaches: Preliminary Results on Comfortable Distances and Preferences. *Technical Report of the AAAI Spring Symposium on Multidisciplinary Collaboration for Socially Assistive Robotics, (AAAI SS07-2007)*, Stanford University, Palo Alto, CA, USA (2007).
- [15] S. Suzuki, Y. Morishima, M. Nakamura, N. Tsukidate, and H. Takeda. Influence of body orientation and location of an embodied agent to a user. *Proceedings of the 20th International Conference on Computer Animation and Social Agents (CASA2007)*, Hasselt University, Belgium, 1-10 (2007).
- [16] A. Kendon. *Conducting Interaction – Patterns of Behavior in Focused Encounters*. Cambridge, MA: Cambridge University Press (1990).
- [17] H. Hüttenrauch, K. Severinson-Eklundh, A. Green, and E. A. Topp. Investigating Spatial Relationships in Human-Robot Interaction. *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Beijing, China (2006).
- [18] M. Argyle. *Social Interaction*. London: Methuen & Co Ltd (1969).
- [19] Y. Nagai, M. Asada, and K. Hosoda. Learning for joint attention helped by functional development. *Advanced Robotics*, 20 (10), 1165 – 1181 (2006).
- [20] M. Staudte and M. Crocker. The utility of gaze in spoken human-robot interaction. *Proceedings of Workshop on Metrics for Human-Robot Interaction 2008, March 12th*, Amsterdam, the Netherlands, 53-59 (2008).
- [21] C. Breazeal, C. D. Kidd, A. L. Thomaz, A. L.; G. Hoffman, and M. Berlin. Effects of Nonverbal Communication on Efficiency and Robustness in Human-Robot Teamwork. *Proceedings of International Conference on Intelligent Robots and Systems (IROS)* (2005).
- [22] C. L. Sidner, C. Kidd, C. Lee, and N. Lesh. Where to look: A study of human-robot engagement. In *Intelligent User Interfaces (IUI)*, Funchal, Island of Madeira, Portugal, January 2004, 78-84 (2004).
- [23] A. Green. *Designing and Evaluating Human-Robot Communication. Informing Design through Analysis of User Interaction*. Doctoral Thesis. KTH Computer Science and Communication, Stockholm, Sweden (2009).
- [24] J. E. Young, R. Hawkins, E. Sharlin, and T. Igarashi. Toward Acceptable Domestic Robots: Applying Insights from Social Psychology. *Int. Journal of Social Robotics*, 1, 95-108 (2009).
- [25] M. Hanheide and G. Sagerer. Active Memory-based Interaction Strategies for Learning-enabling Behaviors. *Proc. International Symposium on Robot and Human Interactive Communication (RO-MAN'08)* (2008).
- [26] E. Hall. A System for the Notation of Proxemic Behavior. *American Anthropologist*, 65, 1003-1026 (1963).
- [27] Elan annotation tool. <http://www.lat-mpi.eu/tools/elan>

Towards Intrinsically Learned Skin Models in Robots

Simon McGregor and Daniel Polani¹

Abstract. We describe the use of information-theoretic principles following [4] for sensoritopic map construction on a simulated “skin”. This research constitutes a first step in robot body model learning based on uninterpreted tactile sensors. Results are reasonable, although our findings identify a serious limitation in the information distance metric: the distance is analytically bounded by the number of bins used for discretisation, so that reconstructions tend to “spherise”. Furthermore, the tactile modality is significantly sparser than the visual modality, which introduces further artefacts in the information distance, particularly under maximum entropy binning. We conclude that some method which allows sensoritopic reconstruction directly from mutual information would be preferable.

1 Introduction

In many cultures, tactile contact forms an important part of human social communication. Technology permitting, the tactile modality offers opportunities for enriching and extending human-robot interaction; consequently, the development of full-surface tactile sensors (and algorithms for processing tactile data in robots) are areas of significant scientific interest and active research.

Our approach emphasises a developmental approach to robot intelligence, in which the robot’s mastery of its own sensorimotor contingencies should be integral to its social behaviour. Thus, social behaviour should be built on learned abilities from the ground up, and tactile social behaviour should be based on the robot’s capacity to learn about its tactile embodiment.

This research is being conducted in the context of the EU RoboSKIN project; specifically, as a precursor to tactile gesture recognition within social environments. Our aim is to provide a versatile framework for robot learning in the tactile modality, so that the robot can shape its own interactions with the physical and social world. The target platform is Kaspar, a robot play companion for children with autism, although the experiments described here are in simulation only.

A valuable first step in this direction would be a robot capable of learning its own body model automatically, based purely on information coming from its sensors. This paper describes preliminary research on tactile sensory reconstruction in 3D, extending work by Pierce & Kuipers [7] and Olsson [6, 4]. In our model, the robot’s body surface is covered in tactile sensors and there is no prior information about either the shape of the robot, or the location, topology and response properties of these sensors. [7] describes this scenario as having *uninterpreted sensors and actuators*.

We consider the problem of reconstructing the geometry of a simulated passive rigid body covered in tactile sensors, with no prior information about the body or the sensors. Our experiments involve four distinct body geometries (plane, sphere, cylinder and Y-shape).

No active behaviour occurs on the part of the “robot”. This problem is far more basic than learning through exploration; consequently it is easier for us to conceptualise, although the inability to participate actively makes the problem “harder for the learner”.

To make minimal assumptions about the sensors and the robot’s environment, we use the framework of *information theory*. This framework provides a general way of conceptualising systematic relations between variables, typically at the cost of some computational complexity and data quality. A simple information-theoretic distance measure described by Crutchfield [2] serves to quantify the dissimilarity between different sensors. Using this measure and a standard multidimensional scaling algorithm [3], we can reconstruct a model loosely analogous to sensoritopic maps in the brain. The rationale for this is not necessarily to perform a geometrically faithful reconstruction of the physical placement of the sensors, but to capture their essential interrelation in some meaningful way.

The results were evaluated quantitatively by considering the linear correlation between reconstruction distance and physical sensor distance, as well as qualitatively by visual examination of the output.

2 Previous Work

The concept of learning from uninterpreted sensors and actuators is described in [7], where a variety of learning techniques are used to deduce features of the robot’s sensors and actuators (such as the dimensionality of the sensor field, or the degrees of freedom in the actuators), and to learn simple control behaviours. Our group has extended information-theoretic principles to this domain, discovering visual flow information and reconstructing the visual topology of an AIBO robot (summarised in [6]).

However, previous work on sensoritopic maps from uninterpreted sensors has tended to focus on visual fields. This paper describes the application of the same information-theoretic methods to the tactile modality, which is embedded in a three-dimensional space and will typically carry sparser information than the visual modality.

3 Simulations

A number of simulations were run using the Open Dynamics Engine (ODE) physics library. Sensors were modelled as short-range proximity sensors distributed on the surface of a rigid object. These sensors contributed nothing to the dynamics of the object, and registered a signal when another simulated body came within the sensor’s range. The sensor’s signal was equal to the other body’s maximum penetration into the sensor’s sphere of detection.

We simulated four distinct sensor configurations: 60 sensors distributed uniformly on the surface of a sphere, 60 sensors distributed in 10 columns of 6 on the surface of a cylindrical object, 105 sensors distributed on the surface of a compound y-shape object, and 169

¹ University of Hertfordshire, UK, email: s.mcgregor2@herts.ac.uk

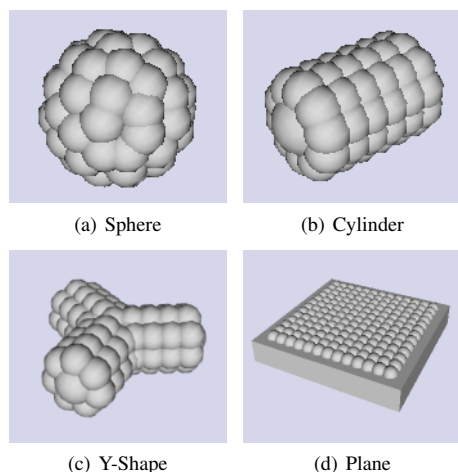


Figure 1. The sphere (60 sensors), cylinder (60 sensors, none on the ends), y-shape (105 sensors), and plane (169 sensors) geometries.

sensors distributed in a grid of 13 by 13 on one surface of a cuboid (see figure 1). The details of each geometry are given in section 3.1.

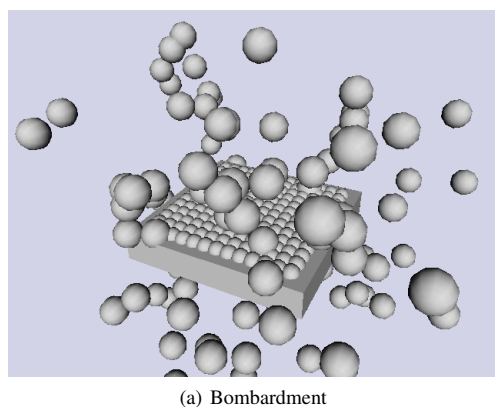


Figure 2. Bombardment simulation scenario.

Each sensor geometry was subjected to bombardment by a “bath” of elastic balls. The simulated tactile sensor data consists simply of the response of the tactile sensors to these balls. Details of the simulation are given in section 3.2.

3.1 Geometries

3.1.1 Sphere Sensor Geometry

For the sphere, sensor positions were calculated by running a multi-dimensional scaling algorithm to embed 60 equidistant points in 3D Euclidean space. This gives a roughly even distribution over the surface of a sphere. These points were normalised to lie on the surface of a sphere of radius 0.5; the sensor range was set to 0.2m.

3.1.2 Cylinder Sensor Geometry

The cylinder object was actually modelled as a rigid cylinder capped by two semi-spheres. The cylindrical part of the object was 1m long with a radius of 0.3m (so the total length of the object was 1.6m). Sensors were placed evenly on the cylindrical portion of the object (i.e. not on the rounded ends) in 10 rows of 6 sensors, each with sensor range 0.2m. (see figure 1(c)).

3.1.3 Plane Sensor Geometry

The plane object was a 2.5m by 0.4m by 2.5m cuboid, with sensors arranged in a regular 13 by 13 planar grid on the central 2m by 2m square of the cuboid’s upper surface. The sensor range was approximately 0.12m.

3.1.4 Y-Shape Sensor Geometry

This was a compound object formed by merging three rounded-end cylinders (as described in section 3.1.2): the three parts are positioned on three evenly distributed radii of a planar circle so as to resemble the letter ‘Y’. Sensors were placed evenly on each cylindrical part in 7 rows of 5 sensors - note that this caused some sensors to overlap with the body of the object. The sensor range was approximately 0.21m.

3.2 Bombardment Simulation

The sensor object was an immobile body inside a large rigid cube in zero gravity. Numerous elastic spheres moved around the interior of the cube colliding with each other, the cube and the sensor object (sphere, y-shape, plane or cylinder).

The object was centred in a 4m by 4m by 4m hollow cube and was not permitted to move. 100 perfectly elastic balls of radius 0.2m and density 1kg/m^3 were initialised at random locations between 0.5m and 1.5m from the cube centre. The velocity of the balls was initialised to a uniform random vector with linear components between -100m/s and 100m/s in each direction.

The ODE simulator has a tendency to “wind down” over time, with numerical integration artefacts meaning that after a certain amount of simulation time, the mean velocity of the balls reduces substantially and eventually grinds almost to a halt. The simulation was modified to prevent this by calculating the total kinetic energy of the simulation at the start, and by every second simulation step rescaling the balls’ velocity uniformly, so as to keep total kinetic energy conserved.

4 Learning The Body Model

Following [4], the algorithm we use for sensoritopic reconstruction consists of two parts:

1. Estimate the distances between each pair of sensors. These estimates are based on information distance (see section 4.1).
2. Convert these distances into a set of 3-dimensional points. This process is known as *multi-dimensional scaling* or MDS (see section 4.2).

4.1 Information Distance

Information distance [2] is a metric defined on equivalence classes of random variables (where direct recordings of variables are equivalent). Briefly, the information distance $D(A, B)$ is equal to

$$2H(A, B) - H(A) - H(B)$$

where $H(X)$ is the Shannon entropy of X :

$$H(X) = \sum_x P(X = x) \log_2 P(X = x)$$

It can be shown [2] that information distance fulfils the metric axioms (symmetry, non-negativity, identity of indistinguishables, and the triangle inequality). This property is what allows us to use it in a standard multidimensional scaling algorithm, and why more well-known information theoretic quantities such as mutual information (a similarity measure) are not directly applicable to this scenario.

For convenience, we discretise the sensor readings into a pre-defined number of bins, which allows a straightforward computation of their individual and joint Shannon entropies according to the standard formulae. Here, we ignore the temporal sequence of the data and attend only to the co-variation in instantaneous readings. The use of information-theoretic measures captures all instantaneous relationships between the sensors, up to the resolution limits imposed by the discretisation. To get the most possible information out of the discretised sensors, we use a maximum entropy binning, which was confirmed empirically in [5] as producing better results in conjunction with the information distance metric.

4.2 Multidimensional Scaling

Multidimensional scaling (MDS) algorithms [1] are those algorithms which, given a matrix of dissimilarity scores between pairs of entities, attempt to assign those entities coordinates in a (typically) Euclidean embedding space so that the distances between the coordinate points correspond to the dissimilarity scores in the matrix. Different algorithms correspond to different ways of measuring how well the reconstructed distances approximate the dissimilarity matrix; we use the SMACOF algorithm [3], which provides a reconstruction by minimising a quantity called stress.

In our case the input to the MDS algorithm is simply the matrix of information distances between sensors, calculated as described in section 4.1. For methodological purposes, it is simplest in the context of this study to interpret this reconstruction as a direct representation of the physical locations of the sensors relative to one another. This is justified by the reasonable assumption that nearby sensors will have high mutual information and hence low information distance.

The two-fold role of the MDS algorithm in our approach should be explained. While the information distance matrix is a mathematically valid matrix of metric distances, it will typically not correspond to a set of Euclidean distances between points in 3-dimensional space. Consequently, in order to produce any coherent 3D reconstruction, the information distance matrix needs to be processed by some algorithm which performs Euclidean embedding - and this is exactly what a MDS algorithm is. However, the MDS algorithm also performs a secondary purpose. In effect, it uses our prior knowledge about the spatial structure of the problem to *improve* the distance estimate provided by the information distance matrix. We can quantify this by considering both the correlation between raw information distance and physical sensor distance, and the correlation between (post-MDS) reconstructed distance and physical sensor distance, as described in section 5.

5 Results

The original and reconstructed geometries can be seen in figures 3 to 4. It should be obvious from these images that the main features of the geometry are reconstructed for the sphere and the cylinder. When the reconstructions are viewed in three dimensions, the reconstructed plane resembles a piece of paper “scrunched up” into a ball, where local relations are reasonably preserved; this is hard to convey in a 2-dimensional image.

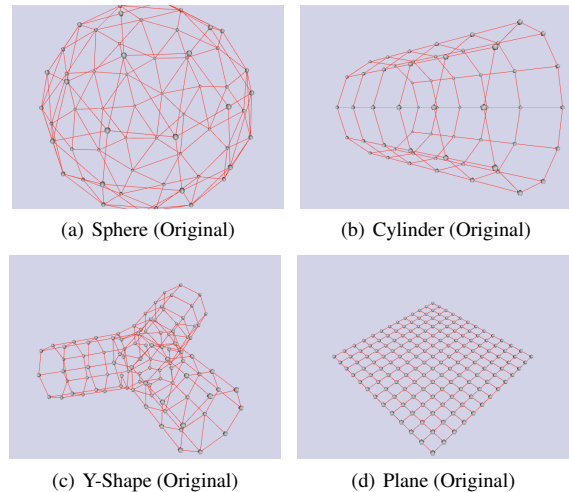


Figure 3. Original sensor geometries for sphere, cylinder, y-shape and plane.

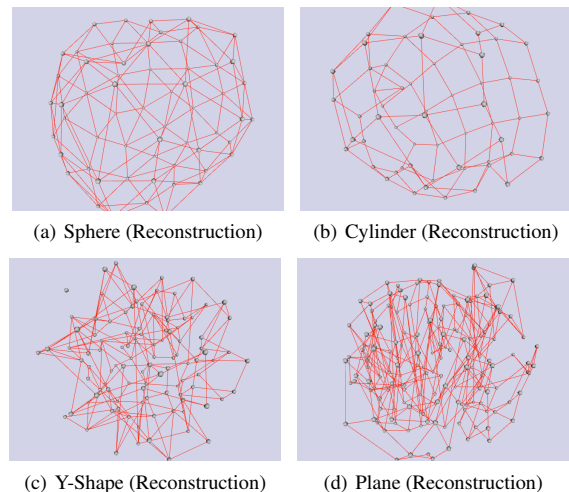


Figure 4. Sensoritopic reconstructions for sphere, cylinder, y-shape and plane.

The Y-shape is evidently less well reconstructed, and this is reflected in our correlation results (figure 5). However, it should be noted that all reconstructed distances, including these this “worse” case, have a positive correlation with the original distances which is statistically significant at the $p = 0.05$ level or better.

Form	r (pre-MDS)	r (post-MDS)
Cylinder	0.57	0.81
Y-Shape	0.24	0.34
Sphere	0.64	0.98
Plane	0.32	0.58

Figure 5. Linear correlation (Pearson’s r) between simulated physical sensor distances and sensoritopic reconstruction distances before and after multi-dimensional scaling in six different experiments. The pre-MDS figures are the information distances.

It can be seen that, in most cases, application of the multi-dimensional scaling algorithm makes a dramatic difference to the quality of the reconstruction. This should not be surprising, since MDS imposes a Euclidean structure on the reconstruction which is not necessarily there in the information distance.

6 Discussion

Our results using this method are generally weaker than the results obtained by Olsson [4]. This has allowed us to analyse the limitations of the technique and identify important differences between the tactile and visual modalities.

The technique’s biggest limitation relates to a hard upper bound on the information distance metric, which causes the majority of reconstructed inter-point distances to saturate at a maximum value. The upper bound arises as follows: the information metric $D(A, B)$ can be written as

$$H(A) + H(B) - 2I(A; B)$$

where $H(X)$ is the Shannon entropy of the random variable X and $I(X; Y)$ is the mutual information of X and Y . Since all these quantities are non-negative, the largest value $D(A, B)$ can take is $H(A) + H(B)$. But our discretisation into n bins means that the maximum entropy for any sensor is simply $\log_2 n$, regardless of the nature of the sensor. Moreover, our maximum entropy binning method means that $H(A) + H(B)$ is approximately $2 \log_2 n$ for most sensor pairs.

This constant-distance phenomenon causes the reconstructions to “warp” such that the reconstructed sensor positions tend to be distributed on the surface of a sphere (see figures 4, although the phenomenon is more evident when viewing in three dimensions). This phenomenon is in fact also visible in Olsson’s reconstructions (see, e.g. figure 2c in [4]). However, our scenario is different from Olsson’s in several ways which we believe significantly exacerbate the problem, reducing the effectiveness of the information distance method:

- Our reconstruction is in 3D, while Olsson’s was in 2D. Consequently there is more hypersphere surface available for sensor points to occupy and non-spherical shapes will experience more distortion.
- Our data is sparse, which reduces the correlation between information distance and physical distance. The maximum entropy binning divides data as evenly as possible amongst bins; if the data is rich, the entropy of each sensor will be the same (equal to the log of the number of bins). In practice, then, only the mutual information between sensors is reflected in the information distance. When data is more than 50% sparse, however, the bin containing 0 will be larger than the other bins, and the entropy of each individual sensor accounts for significant variation in the information distances. This acts as noise, and mutual information is the better indicator of physical distance.

The direct use of mutual information (as opposed to information distance) promises more accurate results. However, mutual information is a similarity measure rather than a dissimilarity measure, so it is not appropriate as an input to classical MDS techniques. Our most recent research focuses on ways to overcome this problem.

ACKNOWLEDGEMENTS

The work described in this paper is conducted within the EU project RoboSKIN and is funded by the European Commission in the 7th Framework Program under grant agreement 231500.

REFERENCES

- [1] I. Borg and P.J.F. Groenen, *Modern Multidimensional Scaling: Theory and Applications*, Springer, 2005.
- [2] J.P. Crutchfield, ‘Information and its metric’, in *Nonlinear Structures in Physical Systems - Pattern Formation, Chaos and Waves*, pp. 119–130. Springer Verlag, (1990).
- [3] Jan de Leeuw, ‘Applications of convex analysis to multidimensional scaling’, in *Recent Developments in Statistics*, eds., J.R. Barra, F. Brodeau, G. Romier, and B. Van Cutsem, 133–146, North Holland Publishing Company, Amsterdam, (1977).
- [4] Lars Olsson, Chrystopher L. Nehaniv, and Daniel Polani, ‘Measuring informational distances between sensors and sensor integration’, in *In Artificial Life X - Proceedings of the Tenth International Conference on the Simulation and Synthesis of Living Systems*, eds., Luis Mateus Rocha, Larry S. Yaeger, Mark A. Bedau, Dario Floreano, Robert L. Goldstone, and Alessandro Vespignani, pp. 316–323. MIT Press, (2005).
- [5] Lars Olsson, Chrystopher L. Nehaniv, and Daniel Polani, ‘Sensor adaptation and development in robots by entropy maximization of sensory data’, in *In Proceedings of the 6th IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA 2005)*. *IEEE Computer*, pp. 587–592. Society Press, (2005).
- [6] Lars Olsson, Chrystopher L. Nehaniv, and Daniel Polani, ‘From unknown sensors and actuators to actions grounded in sensorimotor perceptions’, *Connect. Sci.*, **18**(2), 121–144, (2006).
- [7] David Pierce and Benjamin Kuipers, ‘Map learning with uninterpreted sensors and effectors’, *ARTIFICIAL INTELLIGENCE*, **92**, 169–229, (1997).

Expressive Gestures Displayed by a Humanoid Robot during a Storytelling Application

Catherine Pelachaud¹ and Rodolphe Gelin² and Jean-Claude Martin³ and Quoc Anh Le⁴

Abstract. Our purpose is to have the humanoid robot, NAO, to read a story aloud through expressive verbal and nonverbal behaviors. These behaviors are linked to the story being told and the emotions to be conveyed. To this aim, we are gathering a repertoire of gestures and body-postures for the robot using a video corpus of story-tellers. The robot's behavior is controlled by the Greta platform that drives agents to communicate through verbal and nonverbal means. In this paper we are presenting our research methodology and overall framework.

1 Introduction

Expressivity is an important factor in communication, be human-human, human-virtual agent or human-robot communication. It is linked to emotional states [30]; it also reflects some idiosyncratic features [11]; it has several communicative functions such as calling for attention or showing contrast [6]. Expressivity can be conveyed multimodally, through the voice and the body. Communication involves much more than word. Prosody, gesture, facial expression, gaze, body posture, all participate to conveying a message by a sender (the speaker) to an addressee (the listener). Gesture convey complementary features compared to speech (see Kendon's continuum described by [24]). Furthermore they have been observed to play a role in the expression of emotion [30, 16]. Our idea is to endow a humanoid robot with human communicative capabilities.

Nao is a humanoid robot able to display body movement and gesture. It can talk and interact with humans. In our project, our aim is to have Nao reads expressively a story. This objective implies to enhance the expressivity of the speech synthesizer and the behavior animation technology. When listening to a story being told without expressivity, it is difficult to perceive the various peripeteia of the story and to personalize the different characters of the story. Moreover one gets disinterested rapidly; it is difficult to maintain one's attention. It is even more so when the story is being told by a humanoid remaining still, not displaying any body movements. To have the robot reads expressively, we foresee two main research directions: extract information from the story to be said and define the nonverbal (acoustic and visual) behaviors to said it aloud. Obtaining a thorough analysis of the text of the story is important. One needs to interpret its content: to extract information on its structure, its various semantic and pragmatic elements as well as its emotional content

[3, 28]. Detailed information from the story is crucial to drive expressive speech and behavior synthesizer. This work is part of a French national project named GV-Lex (Gesture and Voice for an Expressive Reading). Within the GV-Lex project our aim is not only to compute an expressive voice but also an expressive animation for a given application: reading tales to children.

Our idea for driving the behavior of the robot is to apply the multimodal behavior generation model implemented in the Greta platform [26]. To this aim we have enhanced the repertoire of gestures of the robot and of the agent. Since they do not have the same motion capacities we are developing a gesture representation schema based on the notion of gesture variant [4]. The resulting animation of the robot and of the virtual agent may not be identical but would convey the same meaning. Thus, at first, we gather information of the type of gestures and body postures that are displayed when telling a story. This is done by conducting a careful analysis of a video corpus of story-tellers. Particular attention has been put on the emotional information body posture conveys. Body posture is defined as a combination of head direction, arm shape and lower body position. The analysis also provides information of the communicative gestures that are used to convey a given intention. On a second stage, we are using the Greta platform to compute the robot's animation. Given an emotional state and communicative intention to transfer, the Greta platform computes which non-verbal behaviors to use and with which expressivity. It is SAIBA compliant [29] and uses two representation languages, one at the intention level FML [13] and one at the behavior level BML [29]. A repertoire of behaviors, specified in BML, contains the behaviors the robot can use. This repertoire is constituted of pairs where one element is the description of the behavior and the second element the meaning attached to it.

In this paper we focus on the nonverbal behavior of the robot. We present our preliminary result and explain our research methodology. After an overview of the state of the art, we present Nao, the humanoid robot we use. We argue how Nao can be accepted as a story teller and which capacities it requires. We then turn our attention to the behavior engine we use. We follow by describing the video corpus of story-tellers and by explaining how multimodal behaviors of actors were recorded when telling the same story. Finally we present a gesture specification for both the robot and the virtual agent. A conclusion ends the paper.

2 State of the art

Several expressive robots are being developed. Behavior expressivity is often driven using puppeteer technique [21, 31]. For example Xing and co-authors propose to compute robot's expressive gestures by combining a set of primitive movements. The four movement primi-

¹ CNRS, LTCI Telecom ParisTech, France, email: catherine.pelachaud@telecom-paristech.fr

² Aldebaran, France email: rgelin@aldebaran-robotics.com

³ CNRS, LIMSI, email: martin@limsi.fr

⁴ CNRS, LTCI Telecom ParisTech, France, email: quoc@telecom-paristech.fr

tives are: walk involving legs movement, swing-arm for keeping the balance in particular while walking, move-arm to reach a point in space and collision-avoid to avoid colliding with wires. A repertoire of gestures is built by combining primitives sequentially or additionally.

Imitation is also used to drive a robot's expressive behaviors [14, 7]. Hiraiwa et al [14] uses EMG signals extracted from a human to drive a robot's arm and hand gesture. The robot replicates the gestures of the human in a quite precise manner. This technique allows communication via the network. The robot can act as an avatar of the human. Another example is Kaspar [7]. Contrary to work aiming at simulating highly realistic human-like robot, the authors are looking for salient behaviors in communication. They define a set of minimal expressive parameters that ensures a rich human-robot interaction. The robot can imitate some of the human's behaviors. It is used in various projects related to developmental studies and interaction games.

In the domain of virtual agents, existing expressivity models either acts as filters over an animation or modulates the gesture specification ahead of time. Emote implements the effort and shape components of the Laban Movement Analysis. These parameters affect the wrist location of the humanoid. They act as a filter on the overall animation of the virtual humanoid. On the other hand, a model of non-verbal behavior expressivity has been defined that acts on the synthesis computation of a behavior [12]. It is based on perceptual studies conducted by Wallbott [30]. Among the large set of variables that are considered in the perceptual studies, six parameters [12] were retained and implemented in the Greta ECA system [27]. Other works are based on motion capture to acquire the expressivity of behaviors during a physical action, a walk or a run [25].

3 Acceptability of domestic humanoid robots

Since 2005, Aldebaran Robotics produces Nao, a 57cm high humanoid robot fully motorized. At the end of 2009, more than 400 Naos have been sold all over the world. Today, the users of Nao are mainly researchers; however Aldebaran is convinced that humanoid robot could become the best robotic assistant at home. There are several objective reasons that motivate our choice of create Nao as a robot with a humanoid appearance and communicative capabilities. Since domestic environments are designed for human beings, the human shape is the most appropriate to intervene in these environments: catching objects on a table or in a closet requires arms, hands and vertical posture; bipedal locomotion is not only necessary to climb stairs but is also more convenient to move in cluttered environments. Among the best researchers are working on these necessary functions [20] [8] which are still only available in laboratories. A lot of work has to be done before the robot can be used outside a controlled environment, into the wild.

From a communication point of view, the humanoid robot is much more attractive than nonhuman-like machines. With its body motions (Nao does not have animated mouth, eyes and eyebrows), the humanoid robot is able to catch attention and often affection of people. By imitating or caricaturing expressive motions of humans, the humanoid robot creates a complicity with its human interlocutors [10]. Movements of a humanoid robot will be considered as communicative gestures if they are natural, encompass dynamic quality and convey appropriate meaning that its human interlocutors can understand. Considering the number of degrees freedom (DOF) of a humanoid robot (Nao has 25 DOF), it appears that programming the motion of each axis independently, even in Cartesian mode, is very time con-

suming when subtle expressive motions are expected.

4 Expressive gesture model in Nao

To ensure Nao can produce communicative gestures, Aldebaran developed a very powerful tool for programming motions on Nao, Choregraphe. This graphical tool has been presented in [9]. It allows one to synchronize the motion of Nao limbs intuitively. Working with actors, ethnologists and character animators coming from movie industry, Aldebaran has created a library of expressive gestures illustrating different concepts: objects (ball, weapon, pencil...), situations (I am lost, I don't understand, my batteries are empty) or even emotions (levels of happiness, fear or sadness). Accompanied with relevant sounds, these gestures give to Nao a real presence. To-day more than 200 meaningful animations are available in the library of Nao.

These animations can be called explicitly within a program. For instance, the robot is looking for an object. When it sees this object, it triggers the "happy dance" animation. This very procedural use of expressive gestures is interesting but not so efficient. Each animation is associated with a semantic meaning. Each of them can be represented by a couple (meaning, intensity) where the first element of the couple, meaning, can take the value of an emotional state (eg happiness, sadness, fear) and the second element represents the intensity (low, normal, high) of a given meaning. According to this very simple model, a program running on Nao can call a function "Expressive_motion" with two arguments (emotion, intensity). So far, this function will trigger randomly one of the expressive animations in the corresponding category.

But before the beginning of the project, Aldebaran Robotics developed a software, called "Narrateur" [2] that proposes to read texts accompanied with expressive animations. If the robot only uses a Text To Speech (TTS) software to read texts, it does not offer much more than a computer. In order to show the advantage of a robot as text reader expressive gestures illustrating the text are necessary. The off-line part of Narrateur proposes a simple interface allowing the inclusion, in the text to be read, of beacons that triggered given animations. A smiley at the end of the sentence will be interpreted by an "happy dance" animation for instance. The input text is parsed at run-time. The sentences are sent to the TTS and the beacons are sent to the animation generator of Nao. When "Narrateur" messages are read by Nao, people are generally seduced by the little robot. This kind of simple functions could convince people to have a humanoid robot at home for fun, in a first time, waiting for a more efficient domestic robot in a diversity of application domains.

If Narrateur is a first efficient step towards the development of an expressive robot, its principle to describe animations is at a too macroscopic level. Its main drawback is the difficulty to synchronize precisely the text and the gestures. For instance, the robot will not be able to synchronise gesture and intonation (on the words "little" and the "big" of the sentence "the little girl meets the big bad wolf"). Precised synchronisation of gesture and intonation are important to convey expressivity.

5 Behavior Engine

We aim to use an existing behavior engine to compute the nonverbal behavior of the robot [27]. The Greta platform is SAIBA compliant [29]. It is composed of three main modules:

Intent Planner determines the agent's communicative intentions and emotional states

Behavior Planner decides and schedules which verbal and nonverbal behaviors to use to convey these intentions and emotional states

Behavior Realizer computes the animation to be visualized by the agent

A gestuary is a repository of nonverbal behaviors. Each element of this repository is a pair where one element corresponds to the meaning carried out by the behavior and the second one to the shape description of the behavior. Communicative behaviors are polysemic. They can be attached to different meanings. Identically, a given meaning can usually be expressed by several behaviors. Two main languages have been defined to encode dataflow over these three modules. FML stands for Function Markup Language. It encodes the communicative intentions and emotional states that are outputted by the Intent Planner. BML, Behavior Markup Language, links the behavior planner to the behavior realizer. It specifies the behaviors to be displayed. BML specification is independent of the animation parameters of the agent. Our aim is to use the Greta platform to control a virtual agent, Greta, and a physical agent, Nao. To this aim we are currently developing a gestuary for Nao. The gestuary is being created using information extracted from a corpus analysis that is explained in the next section. The animation of both agents type is driven by FML and BML tags. In a past study, we have used a similar approach with the Sony Aibo robot [1]. One gestuary of backchannel signals has been defined for the robot and one for the virtual agent. The animation of both agents was driven from the same architecture.

6 A multimodal video corpus of storytelling

Corpus-based models of multimodal behaviors are useful for the design of expressive agents and other multimodal interfaces [17]. Designers of embodied agents and humanoid robots might inspire from a variety of complementary knowledge sources in order to specify naturalistic and context specific multimodal expressions: literature, corpora, and results of perception tests. For example, motion capture corpora can be useful for inspiring the design of postural expressions by humanoid robots [19]. As behaviors of individual human subjects are recorded in video corpora using such a technique, individual gesture profiles can be computed [15]. Such individual gesture profile can be useful for designing storytelling agents and robots with individual styles that can be adapted to the audience at hand. Several experimental and corpus based studies about emotional expression observed discriminative features of emotions categories or dimensions in the global posture, the movement quality [30] and basic gestural form features [16].

The ContAct video corpus [22] was recorded in order to collect data on how storytellers use expressive gestures in a specific storytelling context. One French tale was selected according to several criteria. The duration of the tale was supposed to be long enough to enable the expression of a wide variety of emotional states and gestures (but also not too long so that the audience listening to the story read by a virtual agent or a robot would not get distracted). We selected a French tale called "Three Pieces of Night". It is a thousand words long and relates both positive and negative events. Six actors from a local amateur troupe were videotaped while telling this story so that we would be able to study a variety of styles when relating the same story. Two digital cameras were used (front view and side-view) as postural expression of emotion is expected to be three dimensional [18]. Actors had received the script of the story a few days before the recording session. The text was also displayed on the

wall during the session so that subjects could read it from time to time. Each subject was recorded twice so that the most expressive session would be kept for analysis. The collected corpus contains 80 mn of videos (average 7,5 mn for telling the story).

Gestures observed in the corpus will be annotated to inspire the specification of expressive gestures to be used by the Greta agent and the Nao robot. A scheme that has been already been applied successfully in a similar copy-synthesis approach will be adapted [23].

7 Gestures specification for virtual and physical agents

A difficulty arises from the differences in the agents' body specification, if it is virtual or physics. While Nao can not show facial expressions, Greta does not move its lower body. Nao can also use very few hand shapes but it can keep its equilibrium. Greta can display any hand shapes but has no sense of gravity. These variations result in different animations types.

Even though the bodies of the virtual and physical agents are different, if they are controlled by the same FML inputs, their animation should convey the same meanings. That is the sets of BML tags outputted for Nao and for Greta should be similar in meanings (but not necessarily in behaviors) (see Figure 1). To ensure to maintain the same meaning we are expanding the Gestuary for each agent as follows: first we enhance the gestuary of both agents to have similar behaviors as possible. We are also using the notion of gesture variants that was first used by N. Ech Chafai [5]. Gesture variant was introduced by Geneviève Calbris, a French semiologist that studied communicative gestures [4]. Variant of a gesture encompasses a family of gestures that shares the same meaning (eg to negate something) and a core signal (eg vertical flat hand toward the other). Gestures within a family may differ along the non-core signals they use (eg frowning). Thus in the Greta-Gestuary and the Nao-Gestuary, gesture variant shares similar meaning and signal-core. However they differ along the other signals (eg use of frown vs red leds). When the signal-core of a given gesture can be displayed with Greta but is not possible at all in Nao, the gesture entry appears only in Greta's lexicon. Other entries for the same meaning are nevertheless included in Nao's lexicon, so that the meaning can be expressed using another gesture.

8 Conclusion

In this paper we have presented preliminary work we are conducting to develop an expressive robot capable of reading stories aloud. The expressive behavior lexicon for the robot is based on a video corpus analysis and is described using BML. We are driving the robot's animation using the Greta platform. We need to carry out an evaluation to ensure the robot Nao conveys similar intentions and emotional states as the virtual agent Greta, given a same FML input. We will also perform user tests to validate the robot can read expressively and can maintain human users' interest and engagement. Such perception tests will be conducted by comparing the perception of the story told either by a human (video corpus), the robot with / without expressive gestures or the expressive agent. Both objective (recall of the story and of its emotional content) and subjective measures (preference, perceived expressiveness) will be investigated.

ACKNOWLEDGEMENTS

This work has been partially funded by the French ANR project GV-Lex (ANR-08-CORD-024 (<http://www.gvlex.com/>)).

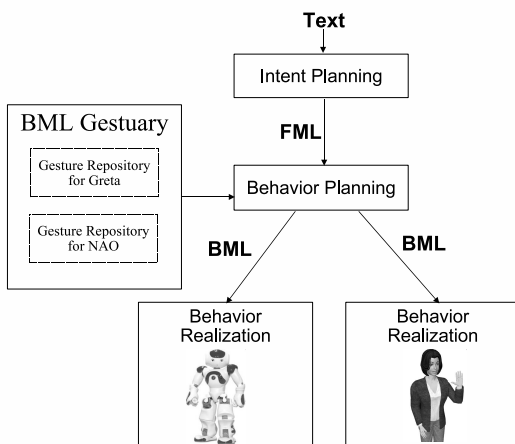


Figure 1. Greta platform driving Nao and Greta

REFERENCES

- [1] S. AL Moubayed, M. Baklouti, M. Chetouani, T. Dutoit, A. Mahdhaoui, J.-C. Martin, S. Ondas, C. Pelachaud, J. Urbain, and M. Yilmaz, 'Generating robot/agent backchannels during a storytelling experiment', in *IEEE International Conference on Robotics and Automation (ICRA'09)*, Japan, 2009., (2009).
- [2] Maisonnier B and Monceaux J. Système et procédé pour générer des comportements contextuels d'un robot mobile. French Patent 66058 FR, 07 2009. Deposit number : 09 54837.
- [3] S. Le Beux, A. Rilliard, and C. d'Alessandro, 'Calliphony: A real-time intonation controller for expressive speech synthesis', in *6th ISCA Workshop on Speech Synthesis (SSW-6)*, Bonn, Germany, (August 22-24 2007).
- [4] G. Calbris, *The semiotics of French gestures*, University Press, Bloomington: Indiana, 1990.
- [5] N.E. Chafai, C. Pelachaud, and D. Pelé, 'Towards the specification of an ECA with variants of gestures', in *Proceedings of Intelligent Virtual Agents, IVA'07*, Paris, (September 2007).
- [6] N.E. Chafai, C. Pelachaud, and D. Pelé, 'A case study of gesture expressivity breaks', *International Journal of Language Resources and Evaluation, Special issue on Multimodal Corpora for Modelling Human Multimodal Behavior*, (2008).
- [7] Kerstin Dautenhahn, Christopher L. Nehani, Michael L. Walters, Ben Robins, Hatice Kose-Bagci, N. Assif Mirza, and Mike Blow, 'KASPAR A minimally expressive humanoid robot for human-robot interaction research', *Applied Bionics and Biomechanics*, (submitted).
- [8] Dimitar Dimitrov, Pierre-Brice Wieber, Olivier Stasse, Joachim Ferrau, and Holger Diedam, 'An Optimized Linear Model Predictive Control Solver for Online Walking Motion Generation', in *IEEE International Conference on Robotics & Automation*, Kobe Japon, (2009).
- [9] E Pot et al., 'Choregraphe: Graphical tool for humanoid robot programming', in *IEEE ROMAN*, (2009).
- [10] J. Monceaux et al., 'First steps in emotional expression of the humanoid robot nao', in *IEEE-RAS Humanoids*, pp. 331–336, Paris France, (2009).
- [11] P.E. Gallagher, 'Individual differences in nonverbal behavior: Dimensions of style', *Journal of Personality and Social Psychology*, **63**(1), 133–145, (1992).
- [12] Björn Hartmann, Maurizio Mancini, and Catherine Pelachaud, 'Implementing expressive gesture synthesis for embodied conversational agents', in *Gesture in Human-Computer Interaction and Simulation, 6th International Gesture Workshop, GW 2005, Berder Island*, pp. 188–199, (2005).
- [13] D. Heylen, S. Kopp, S. Marsella, C. Pelachaud, and H. Vilhjalmsson, eds. *Why Conversational Agents do what they do? Functional Representations for Generating Conversational Agent Behavior The First Functional Markup Language Workshop*, 2008. The Seventh International Conference on Autonomous Agents and Multiagent Systems Estoril, Portugal.
- [14] Akira Hiraiwa, Kouk Hayashi, Hiroyula Manabe, and Toshiaki Sugimura, 'Life size humanoid robot that reproduces gestures as a communication terminal: Appearance considerations', in *Proceedings 2003 IEEE International Symposium on Computational Intelligence in Robotics and Automation*, Kobe, Japan, (July 16-20 2003).
- [15] M. Kipp, *Gesture Generation by Imitation. From Human Behavior to Computer Character Animation*, Boca Raton, Dissertation.com, Florida, 2004.
- [16] M. Kipp and J.-C. Martin, 'Gesture and emotion: Can basic gestural form features discriminate emotions?', in *International Conference on Affective Computing and Intelligent Interaction (ACII-09)*. IEEE Press, (2009).
- [17] M. Kipp, J.C. Martin, P. Paggio, and D. Heylen, *Multimodal Corpora: From Models of Natural Interaction to Systems and Applications*, volume Vol. 5509 of *Lecture Notes on Artificial Intelligence, LNAI 5509*, Springer, 2009.
- [18] A. Kleinsmith and N. Bianchi-Berthouze, 'Recognizing affective dimensions from body posture', in *2nd International Conference on Affective Computing and Intelligent Interaction (ACII 2007)*, eds., A. Paiva, R. Prada, and R. Picard, pp. 48–58, Lisbon, Portugal, (2007). Springer, LNCS, vol. 4738.
- [19] A. Kleinsmith, I. Rebai, N. Berthouze, and J.-C. Martin, 'Postural expressions of emotion in motion captured database and in a humanoid robot', in *International Workshop on Affective-aware Virtual Agents and Social Robots (AFFINE'09) held during the ICMI-MLMI'09 conference*, eds., G. Castellano, J.C. Martin, J. Murray, K. Karpouzis, and C. Peters, Boston, USA, (2009). ACM 978-1-60558-692-2/10/9/11.
- [20] Marin-Urias L, Akin Sisbot E, Pandey A, Tadakuma R, and Alami R, 'Towards shared attention through geometric reasoning for human robot interaction', in *demonstration paper, ICMI MLMI*, Cambridge MA USA, (2009).
- [21] Jun Ki Lee, R.L. Toscano, W.D. Stiehl, and C. Breazeal, 'The design of a semi-autonomous robot avatar for family communication and education', in *Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE International Symposium on*, pp. 166–173, (Aug. 2008).
- [22] J.-C. Martin. The contact video corpus, 2009.
- [23] J.-C. Martin, R. Niewiadomski, L. Devillers, S. Buisine, and C. Pelachaud, 'Multimodal complex emotions: Gesture expressivity and blended facial expressions', *Special issue of the Journal of Humanoid Robotics on "Achieving Human-like Qualities in Interactive Virtual and Physical Humanoids"*. Eds: C. Pelachaud, L. Canamero., **3**(3), 269–291, (2006).
- [24] D. McNeill, *Gesture and Thought*, University of Chicago Press, Chicago, 2005.
- [25] M. Neff and E. Fiume, 'Methods for exploring expressive stance', in *Proceedings of ACM, SIGGRAPH/EUROGRAPHICS Symposium of Computer Animation*, pp. 49–58, Grenoble, (August 2004).
- [26] R. Niewiadomski, E. Bevacqua, M. Mancini, and C. Pelachaud, 'Greta: an interactive expressive ECA system', in *Demo session at AAMAS'09*, Budapest, (May 2009).
- [27] C. Pelachaud, 'Modelling multimodal expression of emotion in a virtual agent', *Philosophical Transactions of Royal Society B Biological Science*, **B 2009 364**, 3539–3548, (2009).
- [28] Sophie Rosset, Delphine Tribout, and Lori Lamel, 'Multi-level information and automatic dialog act detection in human-human spoken dialogs', *Speech Communication*, **50**(1), (2008).
- [29] H. Vilhjalmsson, N. Cantelmo, J. Cassell, N. E. Chafai, M. Kipp, S. Kopp, M. Mancini, S. Marsella, A. N. Marshall, C. Pelachaud, Z. Ruttkay, K. R. Thórisson, H. van Welbergen, and R. van der Werf, 'The behavior markup language: Recent developments and challenges', in *Intelligent Virtual Agents, IVA'07*, Paris, (September 2007).
- [30] H.G. Wallbott, 'Bodily expression of emotion', *European Journal of Social Psychology*, **28**, 879–896, (1998).
- [31] Shusong Xing and I-Ming Chen, 'Design expressive behaviors for robotic puppet', in *Control, Automation, Robotics and Vision, 2002. ICARCV 2002. 7th International Conference on*, volume 1, pp. 378–383 vol.1, (Dec. 2002).

How infants perceive the toy robot Pleo. An exploratory case study on infant-robot-interaction

Karola Pitsch^{1,2} and Benjamin Koch²

Abstract. In order to develop robot systems that can interact with untrained, naïve human users, it is important to understand how people – from different age groups – perceive a given robot system and which features might be relevant for this. While existing studies generally use questionnaires/interviews and/or coding schemes rendering either abstract categories or single features of individual behaviour towards a robot, we suggest to use a different methodological approach: to use the concepts and methodological tools from Ethnomethodological Conversation Analysis (EM/CA). Investigating video data from a study in which users – here: infants 3 to 8 years old – play with the toy robot Pleo, we show that and how (1) a user’s perception, categorization and re-interpretation of a robot system emerges step by step during and from the interaction with the system, and (2) how the user’s attempts to establish coordinated ‘sequences of action’ play a central role in this. The results of our initial exploratory case analysis are discussed in the light of studies which suggest that, in infants, robotic pets seem to blur foundational ontological categories, such as animate vs. inanimate.

1 INTRODUCTION

Over the last years, an increasing body of research has aimed towards developing sociable robot companions that might assist people with their daily routines in domestic environments or provide various sorts of services (such as guidance, entertainment etc.) to the public. As such systems are geared towards untrained, naïve users, it is important to design the interaction between robot and human in a way that users can intuitively interact with a system. In order to do so, it is crucial to understand how people – from different age groups – perceive a given robot system and which features of the system might be relevant for this.

While existing studies have mostly investigated the users’ general attitudes towards and perceptions of robot systems using questionnaires and/or interviews [1, 2, 3], Kahn et al. [4] demonstrate that children’s *reasoning about* robots (e.g. the robotic dog AIBO) and their *behavioural interaction with* these systems differ.³ Most importantly, Kahn et al.’s study suggests

that robotic pets seem to blur foundational ontological categories, such as animate vs. inanimate. They conclude that “the [...] question for the future may not be, “Do young children treat such new technologies as either X or Y?” [...] Rather, what may be needed is a more nuanced psychology of human-robotic interaction that can uncover emergent categories in children’s understanding of and relationships with this potentially new technological genre” [4: 1451f.].

In this paper, we take up Kahn et al.’s call for methodological innovation and suggest to use the ideas and methodological tools from Ethnomethodological Conversation Analysis (EM/CA) to investigate – on the basis of videotaped data – the concrete ways in which users interact with a robot system in a first contact situation. This qualitative approach allows us to uncover (1) how a user’s perception and categorization of a robot system emerges step by step during and from the interaction with the system and (2) to focus thereby on the user’s attempts to establish contingent interaction with the system as a central basis for his/her perception of the system. Our initial case analyses suggest that users might not only (or primarily) consider the robot’s physical appearance as grounds for their perception of a system [2, 3], but rather (or importantly) orient to systematic features of their interactional responsive conduct [6, 7]. To uncover those relevant interactional features is a central challenge for HRI.

In what follows, we will present some preliminary, explorative analysis from a study in which users of different age groups were invited to play with the toy robot Pleo. We will focus on two cases from this corpus – infants 3 to 8 years of age – and address the following questions: How do users perceive a robot system? How might their perception and categorization change over time while interacting with the system? How is this related to forms of contingent interaction they are able to undertake with the system? – Analysis will provide some first insights into the ways in which a qualitative-interactional approach is able to shed light on the question of perception and categorization of robot systems and provide the basis for further, systematic analysis of the entire data set.

2 BACKGROUND AND MOTIVATION

Over the last years, research in HRI has begun to explore people’s attitudes towards and perceptions of robot systems: which kinds of applications and tasks they might be useful for [1], the attribution of competences on the basis of their physical appearance [2, 3], the relationship between the robot’s physical appearance and its behaviour, or the effect of its human-like-ness [8]. On a methodological level, these studies have generally investigated the users’ categorizations using questionnaires and/or interviews. However, Kahn et al. [4] suggest – analysing infant-robot-interaction – that participants’ reasoning about

¹Research Institute for Cognition and Robotics (CoR-Lab) & Applied Informatics, Faculty of Technology, Bielefeld University, 33605 Bielefeld, Germany. Email: karola.pitsch@uni-bielefeld.de.

² Emergentist Semantics Group, Bielefeld University, 33605 Bielefeld, Germany.

³ Similar differences in people ‘talking about their actions’ and ‘doing’ them have often been observed for adults and have lead in the Social Sciences to empirical methods that qualitatively reconstruct the participants’ interactional practices [5].

robots and their behavioural interaction with them differ: On the one hand, infants' explicit evaluation of the robotic dog AIBO and a stuffed dog are similar with regard to animacy, biological properties, social rapport, mental states and moral standing. But their behaviour towards these objects shows e.g. that they engaged more often in exploratory behaviour and attempts of reciprocity with the robot and anxiously flinched away from it after it initiated an action. A follow-up study [9] comparing their conception of AIBO vs. a real life Sheppard dog showed that infants classified it as being more like a robot than a real dog and at the same time treated AIBO in ways that were dog-like. Based on this, the authors suggest that "a new technological genre" might be developing that challenges traditional ontological categories, e.g. between animate and inanimate.

Observations like these call for more detailed exploration of the concrete ways in which participants – both adults and infants – perceive and categorize robot systems. While questionnaires/interviews and coding of videotaped HRI (of measures such as physical distance, contact or body position) are able to reveal a *general attitude* towards these systems, but they won't be able to take into account the *interactional practices* which participants use to explore the system, whether/how their behaviour and their perception might change over time and which features of the robot's conduct they might treat as relevant for their categorizations.

This is where our study starts: We aim at exploring how the users' perception is instantiated within and through the interaction with the robot system, how their categorizations are shaped and re-shaped over time and which features of the systems conduct they orient to.

3 METHODOLOGY: CONVERSATION ANALYSIS AND CATEGORIZATION

In order to investigate the ways in which users perceive a robot in situ, i.e. in and through the interaction with the system, we use the basic concepts and methodological tools from Ethnomethodological Conversation Analysis (EM/CA) [5]. CA constitutes an ethnographic approach that aims at investigating the procedures and methods by which social interaction is organized. Its core, is to investigate the sequential relationship between the different co-participants' actions, the precise timing of events and how they are related to each other. In doing this, it is able – on a structural level – to show that and how some participant's multimodal action makes another set of actions contingently relevant (and provides a basis for detecting missing actions) and to reveal recurring interactional patterns that participants use to solve their concrete, practical problems in and through the interaction.

In this line, Conversation Analysis has suggested an interactional approach to "categorization", which is interested in reconstructing the practices and resources which participants use to refer to persons, objects etc [10]. An analytical apparatus has been developed – the Membership Categorization Device (MCD) – which consists of a set of categories and rules for their applications. Important to us is its idea of 'category-bound activities', i.e. the "kinds of activities or actions or forms of conduct taken by the common-sense or vernacular culture to be specially characteristic of a category's members" [10: 470]. A typical example for such category-bound activities is the often cited "the baby cried – the mommy picked it up" [11], in which

the activity of 'crying' is tied to the category 'baby' in the MCD 'stages of life' and which introduces a scene, in which "the mommy" is understandable as the mommy of that baby. For our investigation of the users' perception of a robot system this is relevant on two levels: (i) through their own pro-active actions (first turn) and (ii) through their reactions upon the robot's actions (second turn) users attribute certain properties to the robot which define it as a token of a type X.

CA's way of proceeding is strictly empirical and qualitative. Analysis begins with a single case analysis in order to reveal – from the data themselves – the relevant analytical issues and categories. In doing this, it attempts to reconstruct the member's perspective – as opposed to applying an analyst's external interpretations or coding schemes to the data. At this stage, it consists of manual analysis, i.e. repeated inspection of videotaped data, transcribing the interaction and by this to uncover the precise timing and relationship of the events of all interaction partners. Its goal is to uncover the structural organization of the event and how one action makes another action contingently relevant next [5, 12, 13]. In a second step (which is beyond the scope of this paper), on the basis of a larger data corpus, it brings together a set of single cases analysis in order to reveal generalizable issues and patterns. A recent and methodologically highly interesting question is whether and, if so, how to best link results from CA to quantitative approaches [14, 15].

4 ROBOT SYSTEM: PLEO

As we wanted to explore how users of different age groups perceive, explore and interact with a given robot system, we chose to use a robotic pet that would be suitable both for infants and adults alike [17, 18]. We needed a system that is able to show 'social behaviour', to engage in sequences of action with the user and to be usable by the participants without prior knowledge or explanations. Also, we wanted it to be easily available and at a reasonable price, so that we would not need to worry having children experiment and play with it. Given these criteria, we choose for our study the robotic baby dinosaur Pleo⁴ developed by Ugobe.

Pleo is about 19 cm tall and 50 cm wide, and disposes of sensors for sight, touch, sound and tilt/shake underneath a painted rubber skin. It has got loud speakers, infrared sensors/transmitters and 14 motors that allow for extensive body movement (Fig. 1).

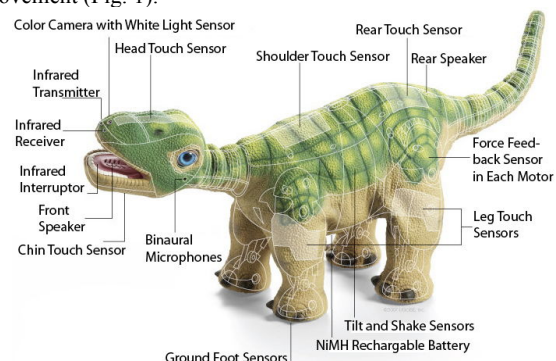


Figure 1. Pleo with its sensors

⁴ For further details see <http://www.pleoworld.org>.

The system disposes of a set of pre-defined actions displaying various emotions and attentional awareness (e.g. swaying its tail, hooting) and, at the same time, is able to learn over time depending on the input it receives from the interaction with the user. In its original state, Pleo behaves like a one week old dinosaur, and – from the interaction with the user – develops to display more and more actions over time. Further capabilities and features can be added to the system via a programming interface.

The Pleo robots used in our study featured the original software and had been intensively ‘played’ with for about 5 hours each, so that they were able to display (at least) the following behaviours: reacting upon caresses, walking, falling asleep and being waked up, ‘eating’ a green plastic leaf (which is part of the Pleo package) and displaying playful behaviour with it.

5 USER STUDY

In order to investigate the ways in which users would interact with a robot system in a ‘first contact situation’, we have undertaken a field experiment videotaping pairs of users who freely played and interacted with Pleo (10-12/2008). We choose to investigate pairs of users (‘dyades’) as this would require the participants to talk to each other, to make their actions accountable for the co-participant and to display their interpretations of the co-participant’s actions. This way, their ‘practical reasoning’ would be available for analysis and reconstruction by the researcher.

The corpus comprises of three different subgroups: (I) Children (3 to 8 years), (II) Teenagers (16 to 18 years) and (III) Adults (21 to 59 years) who all had some general interest in robots, but had no previous experience in interacting with them. For the study, each pair of users was lead to a room where a Pleo was waiting, which was then switched on by the experiment leader. Users did not receive any information about what the system was able to do or any guidance on how to interact with it. They were only instructed as to freely explore and play with the robot as long (or short) as they would like to, yet handle it with care.

	(I) Kids	(II) Teenagers	(III) Adults
Age (years)	3 to 8	16 to 18	21 to 59
Sessions	8	3	2
Duration/Min.	12’52’’	23’44’’	48’51’’
Duration/Max.	42’26’’	33’32’’	54’20’’
Duration/Avg.	28’05’’	28’55’’	51’53’’

Figure 2. Corpus

For the exploratory case analysis presented in this paper, we consider two cases from group (I): a pair of two male siblings (aged 4 and 6 years) and a single girl (aged 3 years).⁵ For group

⁵ Originally, the girl was expected to participate with her brother, but at the play session he was averted at short notice. As this part of the data collection took place during a science festival at Bielefeld University (<http://www.geniale-bielefeld.de>), we decided to not decline the girl’s participation, but rather decided to include it as a special case in our study. In fact, during the session, the girl does not play with Pleo on her own, but – from early on – involves the experiment leader as co-participant.

(I), the infants’ parents or care-takers were generally present in the room being invited to take a seat next to the ‘playzone’, have a snack and observe their offspring playing (Fig. 3). During the first minute of the session, the experiment leader sat close by the children on the floor as to intervene if some damage to the robot would have been foreseeable, and then retreated to control the video cameras.

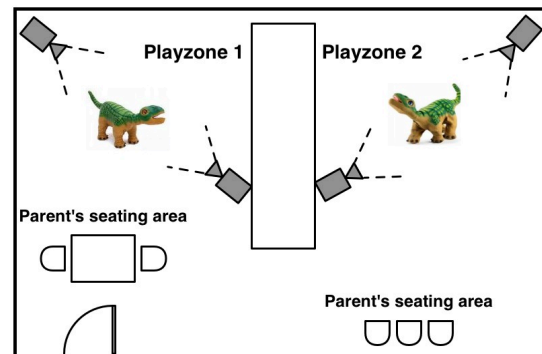


Figure 3. Set up: Infants playing with toy robot Pleo

6 CATEGORY-BOUND ACTIVITIES: STAGES OF INTERACTING WITH PLEO

In a first analytical move, we will be interested in the ways in which a user’s perception and categorization of a robot emerges step by step during the interaction with the system. We will undertake a detailed case analysis in order to reconstruct the subsequent stages of the infant’s conduct and her interaction with Pleo. Analysis will unveil the ‘category-bound activities’ that the infant exhibits, her methods of approaching the system and subsequent interpretations and re-interpretations of the new object during the course of the interaction.

Stage 1 (00:00 to 00:15): First intuitive approach – Handling an inanimate object. When the child – we call her Bea (3 years old, for details see section 4 and footnote 5) – enters the playzone, the experimenter lifts Pleo and switches it on. At this first moment of getting in contact with the robot, Bea touches Pleo at its head/mouth, i.e. that part of the robot that is closest to her and that she could reach instantly (Fig. 4a). The experimenter then offers her Pleo’s leaf, which she takes but immediately puts to the side and, instead, clutches the robot’s neck – making a fist and holding it tightly – and pulls it up (Fig. 4b). Thus, the infant’s first, intuitive approach during the first 15 seconds exhibits no hesitation or fear, and the way in which she grabs it (i.e. rawly clutching) categorizes Pleo as some kind of inanimate object.



Figure 4. (a) Bea touches Pleo at its head while the experimenter switches it on; (b) Bea clutches Pleo’s neck and lifts it up.

Stage 2 (00:15 to 01:20): Socialization into appropriate treatment of Pleo – Experiencing Pleo as an animate object.

When Bea clutches Pleo at its neck and lifts it up, the experimenter intervenes (in her responsibility to ensure the robot's safety): "not so strong- it might feel hurt" ("nicht so doll-dem tut das sonst weh"), and puts Pleo back on its feet. From this, a triadic situation emerges, during which the infant is shown that her handling of the robot might not be appropriate and is offered alternative ways of treating Pleo: (1) Padding gently, and (2) feeding Pleo with the leave. During this second stage (which takes about 65 seconds) the child begins to experience Pleo as an animate object that has its own way of acting and might potentially be threatening.

We will show this in detail at the basis of an interactional fragment⁶, in which the experimenter (E) suggests 'gently padding Pleo' as an appropriate conduct towards the robot. After rejecting Bea's (B) grabbing-round-the-neck (l. 01-03), the experimenter instructs Bea "you need to be very gentle" (04) while beginning to pad Pleo's back. The child immediately copies this conduct (l. 04, 21.10).

Fragment 1 (S005, 00:18 – 00:28)

```
01 B: grabs P's neck | ...
    E: | ah- nicht so ↑DOLL; (.) dem tut
        | ah- not so ↑STRONG; it might

02 E: das doch | ↑WEH; |
    feel | hurt
        | pads P |
    B: | lifts P's front legs |

03 E: | puts P on its feet |
    | nicht dass du dem WEH tust; |
    | not that you would hurt it |

04 E: | mußt du | ga::nz VORsichtig *sein; |
    | you need to be ve::ry careful; |
    | pads P's back |
    B: | pads P's back ...../
        | *21.10
```



But soon – after the experimenter herself stops to do so – Bea again clutches Pleo's (P) neck (l. 05). Immediately after this, the robot turns its head slightly towards the girl and then produces a growling sound (l. 05: "gr::", 24.20), which makes Bea abruptly retract her hand (24.72). Pleo then upgrades its sound as if crying "A::RRR" looking to the other side and then back towards Bea. Once it arrives again at Bea's she shrieks ("AH") and backs off an instance from the robot (26.32, 27.32). At this stage, Bea

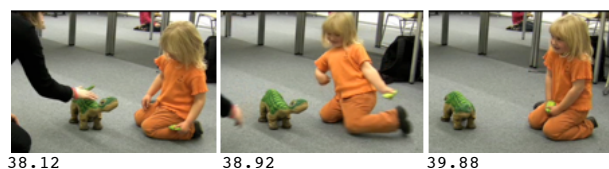
displays her understanding of Pleo as animate, i.e. as an object that reacts upon her actions – here: that exhibits its dislike of a certain kind of treatment (clutching the neck) which the experimenter previously has rejected as inappropriate.

```
05 B: | grabs P's neck | | shrieks, retracts hand |
    P: | head to C | gr:: |
        | *24.20 | *24.72

06 P: ↑AR::: |
    head to E, head to C |
    B: | .hh=AH! |
    B: | retracts hand | moves back
    E: | pads P |
        | *26.32 | *27.32
```



The experimenter reacts upon the child's withdrawing by again padding Pleo (27.32) which Bea observes. However, once Pleo again turns its head towards her (38.12), she instantly backs off even further (38.92, 39.88).



Thus, not only does the robot appear animated to Bea producing its own motions, but also is the concrete interactional timing of this behaviour relevant: following onto what has been considered by the experimenter as being 'inappropriate' conduct, the robot seems to react to this by facing the source of this conduct and producing a growling sound – as if to immediately display its dislike of this conduct by itself. Immediately, the infant displays a new quality of conduct towards the system: by retracting her hand, shrieking and backing off in two subsequent moves, she displays her new interpretation of Pleo as an animated object that is potentially threatening.

One methodological note is on order here: Our analysis takes into account the participants' observable and hearable conduct, i.e. it reconstructs their conduct at the *interactional surface*. This said, we do not make any claims as to whether Pleo – internally – indeed reacts upon the infant's way of clutching its neck or whether its actions simply occur by chance at that moment in time. Instead, important for the analysis presented here⁷ are the participant's interpretations of this conduct in the context in which it occurs.

Stage 3 (01:20 to 06:20): Exploring and Experimenting Pleo – Developing interactional patterns.

In a next stage of approaching, Bea gradually moves from being an observer (of both Pleo and Pleo in interaction with the experimenter) to becoming an actor. After about 1 minute and 20 seconds in the interaction, she begins to explore herself the suggested ways of

⁶ The fragment is rendered using the following transcription conventions: The participants' talk is transcribed as it occurs, i.e. including pauses, hesitations, etc. [16]. Utterances are generally noted in lower case; upper case is used to mark stressed syllables (e.g. not so STRONG) and punctuation is used to signal specific prosodic features (e.g. in ↑STRONG; the prosodic contour starts with a high onset and falls at the end). The participants' visual conduct is annotated as short glosses, written in italics and generally accompanied by a frame grab of the video. The sequential relationship between events is rendered by their positioning on a virtual timeline.

⁷ In general, to link this kind of surface level analysis with logging data showing the system's internal states are a relevant methodological issue. See e.g. the suggestions in [17].

treating Pleo (padding, feeding – see the example in section 7) and to uncover new forms of behaviour (walking, making Pleo fall asleep and wake it up).

This change in Bea's participation status and role in the interaction transpires in her physical position in the room: Contrary to her previously backing off the robot, she physically re-approaches step by step (01:26.28; 03:52.88; 04:28.56). Interestingly, the experimenter – who is setting opposite Bea on the floor (the black person in the video stills) – shadows this conduct and thereby shows her analysis of the emergent interaction: After Bea's first steps of approaching the robot, she begins to physically move backwards and leave the infant to the (originally intended) dyadic situation with Pleo.



This gradual diminishing of the physical distance between herself and the robot goes in pair with a change in her interactional conduct: While, at the beginning, she had been shrieking and withdrew her hand once Pleo's head approached, this reaction disappears step by step. In and through her interaction with Pleo, she uncovers its conduct as being – in part – systematic and begins to develop re-occurring interactional patterns with the system. For instance, when Pleo holds the leave in its mouth for a certain time, it begins to quickly shake its head and grumble. Over several instances, we can observe Bea beginning to treat this as a request to take the leave off its mouth. This way, Pleo becomes an animated object that is able to engage in systematic sequences of actions with her and that is, in this sense, predictable and loses its initial threatening character.

Stage 4 (06:20 to 12:52): Interacting with Pleo in different worlds – A polyfunctional object. After these subsequent stages of discovering and experiencing Pleo, Bea begins to treat the robot as a polyfunctional object with which she can interact in different worlds, real and symbolic ones. Depending on the concrete ways in which she momentarily defines and redefines the situation, Pleo is endowed with different qualities and properties. In addition to continuing her repertoire of the previously developed activities, Bea involves Pleo in a range of different activities, which people – in general – do with babies, puppets, or their pets. She explicitly categorizes Pleo as “baby”, makes it sit on her laps while padding it carefully (11.09.24), picks the robot up and walks away with it from the playzone to show it to her mother (12.12.36, “show mummy” and “baby go with me”) or bids farewell to it when she finally leaves (12.39.44, “bye, robot”).



In sum, this reconstruction of the different stages the infant's conduct and her interaction with Pleo shows how her perception and categorisation emerge and change step by step within and from the interaction. The robot subsequently is considered as (1) an inanimate object, (2) an animate object that is potentially threatening, (3) an animate object that is able to engage with her in re-occurring interactional patterns, and finally becomes (4) a polyfunctional object with which she can interact in different (real and symbolic) worlds.

Presented here as case analysis of one particular user, similar methods of approaching an unknown robot system and its subsequent practical interpretations and re-interpretations during the course of the interaction can be observed with other (both infant and adult) users. To undertake a series of such case-studies focusing on the *interactional practices of categorization* and their development over time, will enable us to get a better understanding of the users' perception of robot systems *in situ*, how these emerge over time, and which features of their interactional conduct might be relevant.

7 EXPERIMENTING SEQUENTIAL ORGANIZATION: FEEDING PLEO

In the course of the previous case analysis detailed inspection of the video-taped data has brought to light that little sequential phenomena of the robot's timely conduct in relation to the infant's actions play an important role in the process of the infant's categorization and re-interpretation: e.g. the robot's head turning and growling which follow directly the infant clutching its neck and thus *appear* to the child as a consequential reaction to her own action (see esp. stage 2). In what follows, we will take a closer look at how users – in their interaction with the robot – are oriented to such sequential relationships and experiment with establishing contingent ‘sequences of action’ with the robot.

We will explore this issue by analysing two instances taken from different user groups in which the users attempt to feed Pleo. In fact, ‘feeding Pleo’ is an ordered sequence of actions which involves the following three consecutive steps: (1) Pleo opens its mouth; (2) User places the green leave in the robot's mouth; (3) Pleo closes its mouth. As a rule, the robot is programmed to open its mouth once it detects – with its camera located at the tip of its nose – the leave's green colour. However, in our experiment the users do not receive this information, and thus are faced with the practical problem of (i) either finding by themselves a systematic way of making Pleo open its mouth or (ii) use those occasional moments for inserting the leave when Pleo randomly opens its mouth by itself.

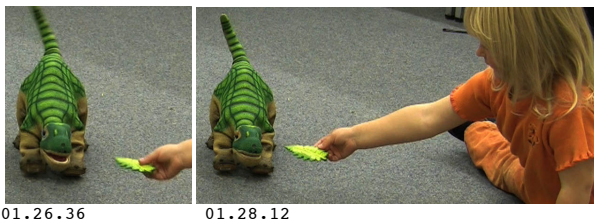
Case 1: Experimenting and the role of a translating mediator. When Bea, the 3 year old girl from our previous case analysis, attempts to feed Pleo for the first time (cf. section 6, stage 3), she encounters some difficulties and needs several attempts to achieve this. In doing this, she experiments with the contingencies that the system's conduct offers.

(I) While Pleo's mouth is open, Bea firstly approaches the leave to about 5 cm and waits (01.26.36). However, after two seconds, Pleo shuts its mouth (01.28.12) and her first attempt fails.

Fragment 2 (S005, 01:24 – 01:42):

01 E: probier DU doch mal; (2.0) | vielleicht *nimmt er's
you try it perhaps it might take
P: mouth open
B: | present-leave-1
*01.26.36

02 E: ja; |
it
P: | mouth shut |
B: | *01.28.12

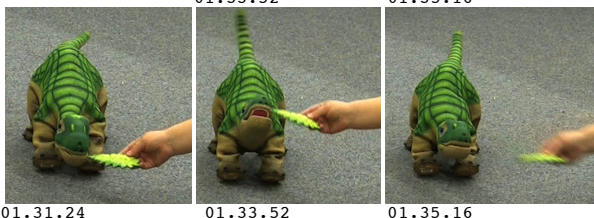


Structurally speaking, the infant treats a certain conduct exhibited by the robot (open mouth) as a first move in a potential 'feeding sequence' which she responds to by offering the leave at distance. This move would make contingently relevant next some reaction by Pleo either indicating that it would like to accept the leave or directly grasping it. Closing its mouth, thus does not meet the relevancies established and closes this attempt of a 'feeding sequence'. It appears that the robot performs a set of actions which could be treated as moves in a potential 'feeding sequence', but – in fact – appear to be part of its own set of action scripts.

(II) Bea then undertakes a second attempt and experiments with another strategy. While Pleo's mouth is closed, she approaches the leave and tries to insert it (01.31.24). Pleo appears to respond to this first turn by a relevant next action: opening its mouth. However, at the same time, the robot lifts and turns its head (i.e. withdraws), which hinders Bea to insert the leave (l. 03, 01.33.52). When Pleo again closes its mouth, she retracts the leave and comments it with a puzzled "↑HM;" (l.03, 01.35.16).

02 E: ja; |
it
P: | mouth shut | mouth open
B: | present-leave-2 |
*01.31.24

03 P: | mouth shut |
B: insert-leave-1 | retract-leave
B: | ↑HM;
*01.33.52 *01.35.16



In this second instance, the robot appears, on the one hand, to engage in a response to Bea's first turn, but then, on the other hand, again continues to follow its own set of action scripts. The

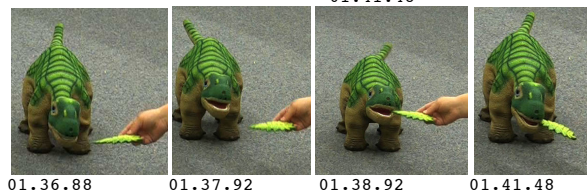
girl thus attempts to arrange with the structures and logic of action provided by the system and tries to find within these a slot that allows her to perform the action that she aims at. This means: Instead of being able to initiate a collaborative sequence of actions in coordination with the robot, Bea organizes her own actions with regard to the contingency exhibited by the system.

(III) Bea then undertakes a third attempt: She holds the leave at about 2 cm off Pleo's head (01.36.88), Pleo then opens its mouth (01.37.92), she moves the leave nearer to the mouth (01.38.92) and the experimenter comments "NOW insert the leaf". Bea finally introduces the leaf, the experimenter acknowledges "alright", Pleo closes its mouth and when Bea retracts her hand, she sees Pleo remaining with the leave in its mouth (01.41.48), which she comments with a happy exclamation "HE!".

04 B: hier; (.) HM; | approach leave |
here; HM;
P: | mouth open |
*01.36.88 *01.37.92 *01.38.92

05 E: und JETZT mußt du's dem REINTun; | genau;
and NOW you have to insert it alright
B: | insert-leave-2

06 P: mouth shut | head down |
B: | ↑↑HE! | laughs
E: | *01.41.48



It is only in this third attempt, in which the experimenter becomes involved as a collaborator, that the infant manages to engage in a successful 'sequence of actions' with the robot. Interestingly, the experimenter operates on a structural level helping the infant to do the appropriate next action at the right moment in time: she observes both the robot's and the infant's actions and helps to 'translate' the robot's actions in terms of which next actions they make relevant from the user. Thus, the experimenter links between the infant's and the robot's actions and thereby enables these two participants to engage in a coherent and collaborative 'sequence of actions'.

Case 2: Observation and reasoning. In another user group, two brothers – we call them Tim and Jon (4 and 8 years old) – try as well to engage in a 'feeding sequence' with Pleo. When they approach Pleo, which is standing in the middle of the playzone and next to its leave, the 4 year old boy Tim immediately grabs the leave and attempts to feed the robot, but fails (similar patterns occur as in Case 1 above, (I) and (II)). Then his older brother Jon grabs the leaf and gives it a try. The robot's mouth is shut when Jon moves the leaf towards it. He presents the leave for about one second holding it horizontally (00.51.64). As no change in Pleo's conduct occurs, he repairs his action by turning the leave into a vertical position (00.52.64). Again, no change in Pleo's conduct occurs and after another second, Jon brings the vertical leave even closer to Pleo's face (00.53.16). About 0.5 seconds later, Pleo opens its mouth and Jon introduces the leave (00.54.32), Pleo closes its mouth and the leave remains inside (00.54.32).

Fragment 3 (S003, 00.48.00 – 01.00.00):

```

01 P: mouth closed ..
P: hackling
J: |present-leaf-1 |present-leaf-2
      *00.51.64 *00.52.64

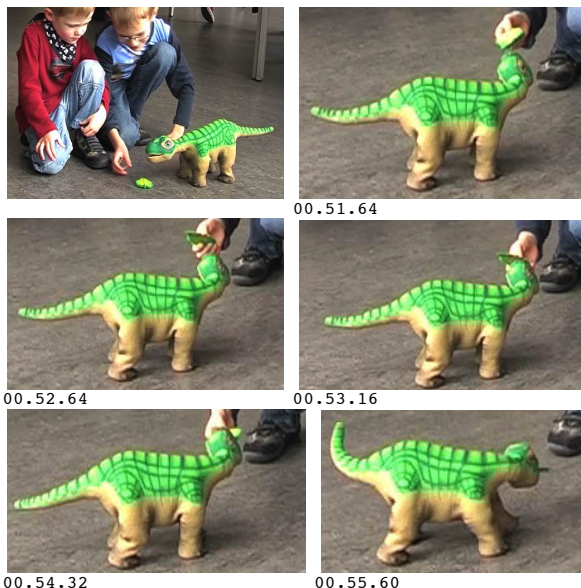
02 A: ich muß das |vor die KA|mera halten; |
      i have to hold it inFRONT of the camera
A: .. |pr-leaf-3 |
J: ..
P: .. |mouth open |insert-leaf
      *00.53.16

03 P: |mouth shut |
J: |turns to M |
M: |du muß das dAVOR halten; |
  |you need to hold it inFRONT;
  *00.54.32

04 J: ja; ich weiß wo die KAMERA ist; (.) darum hab ich
      yes; i know where the cam is that's why i

05 J: das BLATT auch erst dAVOR gehalten;
      did hold the leave inFRONT of it;

```



Jon's success in engaging in a 'feeding sequence' with Pleo results from a basically different approach than that shown by Bea in Case 1: He reaches his goal by observation and reasoning. Jon step by step changes small aspects of his way of presenting the leave (horizontal vs. vertical orientation; distance). In fact, after his second attempt, Jon also begins to explicitly comment his actions: "I need to hold it in front of the camera" and "I know where the camera is (.) that's why I have positioned the leaf right in FRONT of it". In comparison to Bea's attempt of feeding Pleo, Jon (who is considerably older than Bea) not only displays a more technical view on the system, but also he also assumes the approach to evaluate and test the robot's capabilities: Which possibilities does the system offer that allow me to successfully interact with it?

In more general terms, the comparative analysis of these two cases of ‘feeding Pleo’ reveal to which extent and how users – in their interaction with the robot – are oriented to the sequential relationship between actions and experiment with establishing contingent ‘sequences of action’ with the robot. In order to

understand how it comes that users conceive of robots in a certain way, categorize and re-categorize the systems during the course of the interaction, we need to get down for analysis to the level of sequential organisation. Also, we have been able to detect different ways of the users proceeding when attempting to understand the system's conduct and to collaborate with it. In these two cases they are, surely, linked to the infants' different age (3 years vs. 8 years) and it will be relevant whether we can detect – on the basis of the entire corpus – different user strategies depending on their age.

8 SUMMARY & DISCUSSION

Starting point for our investigation has been the aim to investigate how users perceive a given robot system and which features might be relevant for this. Differently from the methods used in existing studies – questionnaires/interviews and coding rendering either abstract categories or single features of individual behaviour towards a robot – in this paper, we have suggested a different methodological approach: to use the concepts and methodological tools from qualitative Ethnomethodological Conversation Analysis (EM/CA), in particular the ideas of ‘category-bound activities’ and sequential organization. Using a corpus of video-taped data from a study in which a set of dyadic user groups freely interacted with the robotic dinosaur Pleo, we wanted to know how a user’s perception and categorization of a robot system emerges step by step during and from the interaction with the system and which interactional features might be relevant for this. We have presented some preliminary, explorative analysis on the basis of two cases, which allowed us to present the methodology and the kind of findings it is able to generate.

Comparing our results to existing studies, we can begin to discuss further the question raised by Kahn et al [4] whether – for infants – robotic pets seem to blur foundational ontological categories, such as animate vs. inanimate, and to which extend a “hybrid object” and/or a “new technological genre” might emerge. Our analysis reveals – using the example of a 3 year old girl – how a user’s perception and categorisation of a robot emerges and changes step by step within and from the interaction. The robot was subsequently considered as (1) an inanimate object, (2) an animate object that is potentially threatening, (3) an animate object that is able to engage with her in re-occurring interactional patterns, and finally becomes (4) a polyfunctional object with which she can interact in different (real and symbolic) worlds. While a single case analysis cannot claim to derive any generalizable conclusion, we are able, however, to uncover those emergent categories. It will be matter of further analysis on the basis of the entire corpus whether the categories derived here might be relevant for other infants as well.

At the same time, our detailed inspection of the video data has brought to light that small sequential phenomena of the robot's timely conduct in relation to the infant's actions play an important role for the infant's categorization and re-interpretation. In a second step, then, we have investigated how users attempt to establish coordinated 'sequences of action' with the robot and have been able to reveal different strategies: (1) Experimenting and organizing their own actions with regard to the contingency exhibited by the system; (2) Making use of a 'mediator' who observes both the robot's and the infant's actions

and helps to 'translate' the robot's actions in terms of which next actions they make relevant from the user; (3) Observation and technical reasoning. Also here, further analysis is required on the basis of the corpus to investigate both the users' strategies and the features they are reacting to.

When turning the single case analyses presented here into a corpus study, we will need to take into consideration further issues: We will need (1) to differentiate between the age groups of our participants – infants, teenagers, adults – and (2) include, for the infants, the different cognitive abilities and ways of conceiving the world between 3 and 8 years of age and how these might be linked to their ways of interacting with the robot. (3) As it transpires from the third fragment, in which two infants play together with the robot, it will be interesting to uncover in which ways this 'collaborative' component impedes on the interaction. In doing this, we might not primarily learn about human-robot-interaction, but about the ways in which siblings interact with each other. (4) If we want to derive guidelines for the design of human-robot-interaction from such a corpus study, we will also need to link the qualitative analyses with quantitative approaches. While EM/CA is able to reveal in a very fine-grained manner features and patterns of the interactional conduct and micro-coordination, we will need to take this further by providing information about the precise timing of events, about the frequency with which certain patterns occur and the probability of some 'interactional paths' rather than others. For this, systematic annotation of the features uncovered in the qualitative analysis will be required using timeline-based annotation tools. On these grounds, we will be able to undertake computational investigation of the corpus (e.g. with MATLAB), in which timestamps and annotation values need to be parsed and algorithms developed for testing the hypotheses generated in the qualitative analysis [15].

ACKNOWLEDGMENT

Karla Pitsch gratefully acknowledges the financial support from the FP7 European Project ITALK (ICT-214668).

REFERENCES

- [1] Dautenhahn, K., Woods, P., Kaouri, C., Walters, M., Koay, K., & Werry, I. (2005). What is a robot companion - friend, assistant or butler? In IEEE/RSJ Int. Conf. on intelligent robots and systems (IROS) 2005, 397-428.
- [2] Hegel, F., Lohse, M., Swadzba, A., Wachsmuth, S., Rohlffing, K., & Wrede, B. (2007). Classes of applications for social robots: A user study. In IEEE International symposium on robot and human interactive communication (jeju island, korea, 26/08/2007).
- [3] Hegel, Frank, Manja Lohse, and Britta Wrede. "Effects of Visual Appearance on the Attribution of Applications in Social Robotics." In RO-MAN 2009. 2009, 64-71.
- [4] Kahn Jr, P. H., Friedman, B., Perez-Granados, D. R., & Freier, N. G. (2004). Robotic pets in the lives of preschool children. In Conference on human factors in computing systems, 1449-1452.
- [5] Sacks, H. (1992). *Lectures on conversation*. Malden; Oxford: Blackwell.
- [6] Pitsch, K., Kuzuoka, H., Suzuki, Y., Süssenbach, L., Luff, P., & Heath, C. (2009). "The first five seconds". Contingent stepwise entry into an interaction as a means to secure sustained engagement. In RO-MAN 2009, 985-991.
- [7] Kuzuoka, H., Pitsch, K., Suzuki, Y., Kawaguchi, I., Yamazaki, K., Kuno, Y., et al. (2008). Effects of restarts and pauses on achieving a state of mutual gaze between a human and a robot. In CSCW 2008, 201-204.
- [8] Mori, M. (1970). The uncanny valley. *Energy*, 7(4), 33-35 (translated by Karl F. MacDorman & Takashi Minato).
- [9] Melson, G. F., Kahn, P. H. J., Beck, A. M., Friedman, B., Roberts, T., & Garret, E. (2005). Robots as dogs? - children's interaction with the robotic dog AIBO and a live australian shepherd. In CHI 2005 (april 2-7, 2005, portland, oregon, USA). ACM Press, 1649-1652.
- [10] Schegloff, E. A. (2007). A tutorial on membership categorization. *Journal of Pragmatics*, 39(3), 462-482.
- [11] Sacks, H. (1972). On the analyzability of stories by children. In: Gumperz, J.J., Hymes, D. (Eds.). *Directions in Sociolinguistics: The Ethnography of Communication*. New York, 325-345.
- [12] ten Have, P. (1999). *Doing Conversation Analysis. A Practical Guide*. London.
- [13] Heath, C.; Hindmarsh, J. & Luff, P. (2010). *Video in Qualitative Research. Analysing Social Interaction in Everyday Life*. London.
- [14] Schegloff, E. A. (1993). Reflections on quantification in the study of conversation. *Research on Language and Social Interaction*, 26(1), 88-128.
- [15] Pitsch, K., Vollmer, A. -L., Fritsch, J., Wrede, B., Rohlffing, K., & Sagerer, G. (2009). On the loop of action modification and the recipient's gaze in adult-child-interaction. In *Gesture and Speech in Interaction (GESPIN) 2009*. Poznan, Poland.
- [16] Seltling, M. et al (1998). Gesprächsanalytisches Transkriptionssystem (GAT). In: *Linguistische Berichte* 173, 91-122.
- [17] Lohse, M., Hanheide, M., Pitsch, K., Rohlffing, K., & Sagerer, G. (2009). Improving HRI design applying systemic interaction analysis (SInA). *Interaction Studies. Special Issue "Robots in the Wild: Exploring Human-Robot Interaction in Naturalistic Environments"*, 298-323.
- [18] Weiss, A., Wurhofer, D., & Tscheligi, M. (2009). "I love this dog"—children's emotional attachment to the robotic dog AIBO. *International Journal of Social Robotics*, 1(3), 243-248.
- [19] Jacobsson, M. (2009). Play, belief and stories about robots: A case study of a pleo blogging community. In *RO-MAN 2009*, 232-237.

Ibn Sina Steps Out: Exploring Arabic Attitudes Toward Humanoid Robots

Laurel D. Riek¹ and Nikolaos Mavridis² and Shammah Antali² and Noura Darmaki² and
Zeeshan Ahmed³ and Maitha Al-Neyadi² and Amina Alketheri²

Abstract. As humanoid robot technology makes its way into the global economy, it is important for robot designers to try to understand the variety of cultural attitudes that exist throughout the world. However, to our knowledge, no one has yet surveyed Middle Eastern attitudes on this topic. We aim to try to bridge this gap. In particular, we wanted to understand how Arabic people might view having humanoid robots in their daily lives, and developed a questionnaire to understand these views. We also built an end-to-end Arabic conversational system on our humanoid robot Ibn Sina, and brought the robot to a mall in the United Arab Emirates. At the mall we encouraged people to interact with the robot as well as complete our questionnaire. With 131 respondents from 21 countries, our findings show significant regional differences in overall attitudes (people from the Gulf viewed humanoid robots more favorably than people from Africa), as well as an effect from college education (people with college degrees viewed humanoid robots less favorably than people without college degrees).

Category: Exploratory studies (*E*)

1 Introduction

As humanoid robot technology starts to make a foray into the global economy, it is important for robot designers to try to understand the variety of cultural attitudes that exist throughout the world. To date, a number of studies have examined cultural attitudes toward humanoids, but they focused mainly on western (USA, Mexico, Western Europe) and far eastern (Japan, Korea, and China) populations [3, 7, 12]. However, we are not aware of any studies examining Middle Eastern attitudes regarding humanoid robots.

People from the Middle East typically hold varied cultural and religious beliefs surrounding depictions of human-likeness. Islam traditionally espouses an anti-iconic doctrine, particularly with regard to three dimensional figures. The main reason for this is that when an artist forms an image of a living being, he or she is trying to adopt the role of creator, which is intended to be reserved for only God. This idea is even ingrained in Qur'anic Arabic; the word for "to fashion or form" (*sawwara*) is synonymous with the word "to create" (*bara'a*), and God is often referred to as a creator (*al-Bari'*) [15]. While we do not wish to delve into a deep analysis of Islamic doctrine and linguistic tendencies, we merely wish to mention this as a potentially strong influence on how people might view humanoid robots.

For other types of technology, some researchers have looked at how Arabic cultural attitudes might affect technology acceptance. Straub et. al noted that a lot of information technology to date has been developed with an ethnocentricity biased toward the cultural systems of the countries that built them. Such oversights can create many problems with technology acceptance [16]. Albirini found similar problems when studying how EFL teachers in Syria view information technology in the classroom. While most of the teachers viewed computers as being a very viable way to improve Syrian standard of living and education as well as being pertinent to Syrian schools and society, they felt the computers first needed to better suit Arabic identity and culture. The teachers felt there were many outstanding social issues to be addressed before using computers in education, and that computers were proliferating too quickly [2].

Thomas [19] and Rogers [14] stress how important it is to understand the social and cultural norms of a country in order to help ensure technology acceptance by its people. If a technological tool does not fit within the macro or micro-culture of a society, potential adopters may resist it [2].

Thus, we wanted to understand how Arabic people might view having humanoid robots in their daily lives, and developed a questionnaire to try to probe some of these ideas. While we considered administering the NARS or RAQ questionnaires developed by Nomura et al. [13], recent work by Syrdal et al. suggests that the internal consistency of NARS may be threatened in cross-cultural/cross-lingual studies [17]. Furthermore, we wanted an instrument that better captured important areas of Arabic life: community, domestic life, and education. Thus, we developed a new questionnaire which we've



Figure 1. A man gives Ibn Sina robot a traditional Emarati greeting.

¹ Computer Laboratory, University of Cambridge, UK, lr314@cam.ac.uk

² Interactive Robots and Media Laboratory, United Arab Emirates University, UAE

³ Mathematics and Informatics, University of Nancy, France

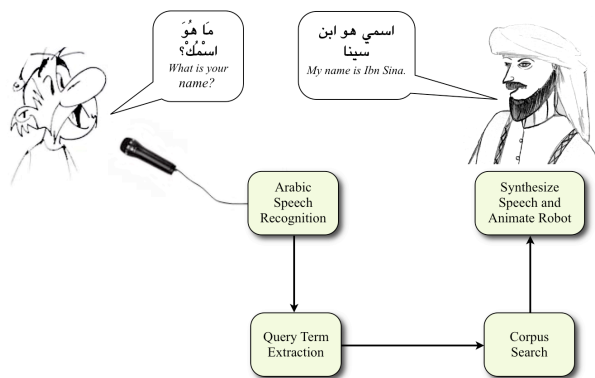


Figure 2. An overview of the system. First a person speaks in Arabic to the robot, next their speech is recognized, then appropriate query terms are extracted, next the corpus is queried, and finally the robot is animated to make facial expressions and speak.

called the Culture Education and Domestic Attitudes toward Robots (CEDAR) scale.

Rather than distribute the CEDAR scale without context, we thought it would be better to run a pilot study in a public setting where people could see and interact with a humanoid robot to get an idea about its capabilities. Thus, we brought our robot, Ibn Sina, to a local mall in the United Arab Emirates.

Overall, we wanted to know if peoples' attitudes were influenced by their region of origin, college educational level, gender, or age. Our major findings, presented in Section 4, are that people from the Gulf had more favorable views toward humanoid robots than people from Africa, and people with college degrees had less favorable views than those without one.

2 System Description

The robot used in our experiment was developed by Hanson Robotics made to resemble Ibn Sina, a well respected Islamic philosopher, doctor, and polymath who lived from 980-1037 A.D. The robot is a central part of the IbnSina Theatre, an augmented reality theater with intelligent robotic and virtual characters [9].

Ibn Sina robot has 19 degrees-of-freedom in its face, and each degree of freedom is intended to represent the human musculature. Its facial movements and expressions are very life-like and natural. The robot also has two degrees-of-freedom in its arms and is able to move them up and down. Figure 1 shows a person interacting with Ibn Sina.

We built an end-to-end system that allowed people to talk to Ibn Sina in Arabic and have the robot generate an appropriate response. The system consisted of three main components: Speech Recognition, Corpus Searching, and Robot Expression Generation. (See Figure 2)

In addition to the system described in this paper, other cognitive engines are currently being developed and transferred to the robot from other embodiments. These include the FaceBots engine [10], which utilizes online social information for more effective dialogues, and the Grounded-situation-models engine [11], which enables situated language capabilities with reference to objects and events the robot is perceiving or imagining.

2.1 Speech Recognition

The speech recognition component of our system is based on the Acapela speech recognition engine [1]. To facilitate speech with the robot, we modeled a number of different sentences that are common to daily life. Figure 3 shows a subset of the Arabic sentences our system can recognize, as well as the corresponding phoneme realization of the Arabic speech. Phoneme realization was necessary because Acapela is based on phoneme based speech recognition technique. These phoneme realizations were generated by a fluent Arabic speaker using the Lexical Editor tool provided with Acapela system.

The Acapela engine utilizes both an acoustic model as well as a language model. The acoustic model contains statistical representations of the sounds that make up each acoustic unit, while the language model contains the probabilities of sequences of words.

Acapela provides options for two modes of recognition: isolated words and continuous speech. In our speech recognition system, we modeled the system to recognize continuous speech.

At the Language model layer, we developed an artificial grammar that restricts the recognition to a list of sentences that were modeled in grammar. Therefore, the effective speech recognition accuracy for our task was significantly improved, as compared to the accuracy that would have been obtained by having used a generic grammar.

2.2 Ibn Sina Corpus and Search

In order to help facilitate a meaningful dialogue, as well as to give Ibn Sina a bit of personality, we developed a corpus of phrases for Ibn Sina to say. All the phrases in the Ibn Sina Corpus (IBC) were written in the first person, and contained standard greetings (e.g., "Nice to meet you"), interesting anecdotes about Ibn Sina's life (e.g., "I developed the physics equations that Newton used when developing his laws of motion."), as well as a few humorous phrases (e.g., "I'm glad you learned something from me, I suggest you go read my book too.").

All items in the IBC were encoded as UTF-8 text files. Further, we added related keyword synonyms to each file in order to ensure an appropriate response would be generated by the robot when someone spoke to it. For example, if someone asked "What is your name?", we might have a file that contained keyword phrases such as "Name", "Who are you", etc.

In order to find phrases in the IBC, we used Google Desktop to index the directory containing all the text files, and wrote a C++ wrapper to send query terms and retrieve documents from it.

Each phrase in the IBC was also converted into an MP3 file that contained a text-to-speech reading of the phrase. We used the Acapela Arabic text-to-speech engine to do this.

2.3 Robot Expression Generation

Finally, after a user's speech had been recognized and an appropriate response was found in the IBC, we animated Ibn Sina to speak the phrase, move its lips in synchronization, and make appropriate facial expressions. For example, smiling while telling a joke, looking concerned when giving advice, etc. The robot's animations were done manually using Brookshire Software's Visual Show Automation [4].

3 Methodology

3.1 Sampling

We decided to run our study at the Al Ain Mall in Al Ain, United Arab Emirates. We chose this location for our study because it is

Phrase	Phoneme Realization
السلام عليكم ورحمة الله وبركاته	assalamualkumwarahmatuallahiwabarakatuh
ما هو اسمك	mahwaasmak
كيف حالك	kaifahaluk
متى ولدت	matawulidta

Figure 3. A subset of the Arabic phrases our system can recognize and their phonetic transliterations.

very egalitarian; frequented by people from a wide variety of cultures, races, income levels, and job classes.

Ibn Sina robot was set up at the entrance to the food court, a wide open space with many visitors. The robot was there for 12 hours, from 10:30AM to 10:30PM. Anyone who came to see the robot or talk to the experimenters were politely offered questionnaires to complete in the language of their choice (Arabic, English, or Urdu). Several people declined to complete the questionnaire for various reasons, such as a lack of time, concern that it was a test about robotics, etc. However, we believe we still had a good sample that well-represented the diverse population of the Emirates.

3.2 Interacting with Ibn Sina

At the mall, the robot was seated in a chair, and its position was stationary throughout the day. The experimenters asked people passing by if they wished to talk to the robot, and if they agreed the received a headset to wear and some simple instructions in Arabic or English to say hello and greet the robot. Also, people, children especially, were encouraged to shake the robot's hand (See Fig. 4).

In terms of conversation content, the robot was limited to the phrases described in Section 2.2. Many people were fairly shy when interacting with the robot, so the experimenters would prompt them with phrases to say or questions to ask the robot. In the evening, when the mall was extremely crowded and noisy, we provided people talking to the robot a list of phrases and discussion topics to try. (See Fig. 5).



Figure 4. Children being encouraged by the experimenters to shake hands with Ibn Sina.



Figure 5. When the mall became very busy and noisy, people talking to Ibn Sina received a sheet of phrases the robot could recognize.

3.3 Instrument

3.3.1 Construction

Our primary instrument was an 11-item, 4-point attitudinal questionnaire that probed Arabic cultural attitudes toward humanoid robots called the CEDAR Scale. (See Appendix). We developed the questionnaire to capture three main ideas that are of high importance in Arabic society: domestic life, community, and education. The first four questions asked respondents to consider their personal feelings about humanoid robots in their daily lives (e.g., "I wouldn't mind if a human-like robot cleaned my house."). The next four questions asked respondents to consider the emotional feelings of others in their community (e.g., "Many people from my home country would feel comfortable if they saw Ibn Sina robot." The final three questions probed ideas about using robotics for educational purposes (e.g., "Children would enjoy learning from a robot like Ibn Sina.").

The scale was intentionally kept short because most Arabic people do not like to take the time to complete written surveys, particularly in public settings. By keeping the survey brief we helped ensure a high rate of completion and participation.

We prepared the CEDAR scale in three languages: Arabic, English, and Urdu, because these are the three most commonly spoken languages in UAE. The translation to Arabic was done iteratively by four bilingual Arabic/English speakers, and the translation to Urdu was done by two bilingual Urdu/English speakers.

All the questions were phrased to be positively biased to ensure that the questionnaire would make sense in the three languages, though two of the community questions contained negative attitudes ("People from my home country would feel [angry, afraid] if they

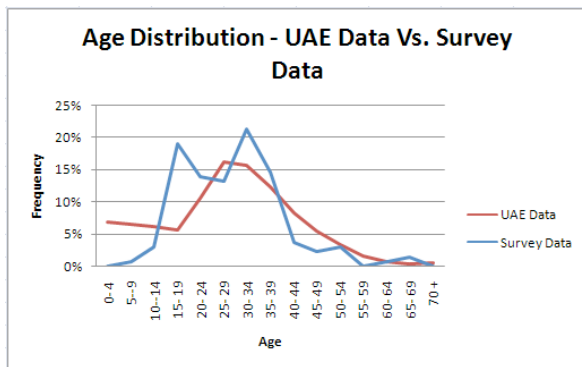


Figure 7. Distribution of ages of UAE nationals and distribution of survey respondents. Source: Cooperation Council for the Arab States of the Gulf.

saw Ibn Sina robot.”). The answers to these items were reversed for the factor analysis described in Section 3.3.2 and for all subsequent results analysis in Section 4.

3.3.2 Reliability and Factor Analysis

Cronbach’s α for the 11-item survey was .675, which gives us confidence in the reliability of using the score as a summed measure.

We also performed Principle Component Analysis (PCA) on the 11-items using an orthogonal rotation. The overall Kaiser-Meyer-Olkin (KMO) was 0.62, and all KMO values for all individual items were greater than 0.5 which is the acceptable limit. Bartlett’s test was $X^2(55) = 285.83$, $p < .001$, which indicated that inter-item correlation was sufficiently large for PCA. Three components with eigenvalues over Kaiser’s criterion of 1 explained 52.82% of the variance. See Figure 6 for the rotated factor loadings. The first component seems to represent the use of humanoid robots for education, the second the use of humanoid robots in domestic life, and the third community feelings toward humanoid robots⁴.

4 Results

4.1 Measures

Our dependent variable was the summed score on the questionnaire. Based on a Shapiro-Wilk normality test we found the score data to be normally distributed ($W=0.97$, $p\text{-value} = 0.006$). Thus, we used parametric statistics in our analysis of this data.

Two manipulations were made before the score was calculated. First, the two questions that asked “Many people from my home country would feel (angry / afraid) if they saw Ibn Sina Robot” were reversed to be positively biased. Second, questions that were not completed were assigned a score of 2.5 so as not bias the means. (Only a few questions were incomplete across the entire sample.)

4.2 Demographics

136 people completed our survey, though 5 people were removed from the sample because they clearly did not read the questions (picked all “1”s). The following demographics and subsequent analysis refer to the remaining 131 respondents.

94 respondents were male, 34 were female, and 3 did not identify their gender. 109 respondents completed the survey in Arabic, 19 in English, and 3 in Urdu. 76 respondents had a college degree, 49 did not, 6 did not answer.

Respondents’ ages ranged from 8-68 years old, the mean age of respondents was 29.15 years old ($s.d = 11.02$). This age distribution corresponded quite well with the overall age distribution in UAE. (See Figure 7). Four people did not report their age.

116 respondents were Muslim (89%); the remaining respondents wrote they were Orthodox: 1, Catholic: 3, Christian: 2, Hindu: 2, and 5 did not identify their religion.

Respondents came from a wide range of countries around the Middle East and Southeast Asia: UAE: 43, Egypt: 21, Oman: 12, Syria: 8, India: 5, Jordan: 5, Pakistan: 4, Palestine: 4, Philippines: 4, Iraq: 3, Saudi Arabia: 3, Sudan: 3, Iran: 2, Morocco: 2, Tunisia: 2, Bangladesh: 1, Lebanon: 1, Libya: 1, Qatar: 1, and Yemen: 1. Also, one person was a UK national, one was from Moldova, and three did not answer.

The questionnaire also asked “Did you talk one-on-one with Ibn Sina robot today? Y/N”; 63 respondents replied “Yes”, 56 replied “No”, and 12 did not answer.

4.3 Quantitative Results

Overall scores on the summed questionnaire ranged from 19 - 42, out of a possible 44. The mean score was 32.4, $s.d.= 4.6$.

4.3.1 Regional Analysis

Given the diverse nationalities of our respondents, we decided to group them by geographic region. The regions we selected were: African (Egypt, Morocco, Tunisia, Libya, and Sudan; $n = 29$), Southeast Asian (Bangladesh, India, Pakistan, and Philippines; $n = 14$), The Gulf (Iran, Iraq, Oman, Qatar, Saudi Arabia, United Arab Emirates, and Yemen; $n = 65$), and Shaam (Jordan, Lebanon, Palestine, and Syria; $n = 18$). Figure 8 depicts the score distributions by region.

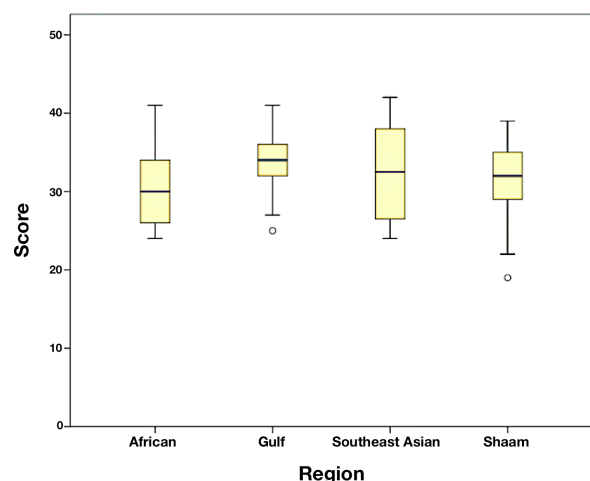


Figure 8. Score distributions on the CEDAR questionnaire by the four regions: African, Gulf, Southeast Asian, and Shaam.

⁴ Factor analysis reporting paraphrased from Field [5])

Item	Rotated Factor Loadings		
	Education	Domestic Life	Community
I wouldn't mind if a human-like robot treated me at the hospital	0.24	0.71	-0.13
I wouldn't mind if a human-like robot cleaned my house	0.04	0.74	0.02
I wouldn't mind if a human-like robot was in my workplace	-0.05	0.68	0.10
I wouldn't mind if my child was instructed by a human-like robot	0.28	0.31	-0.40
Many people from my home country would feel happy if they saw Ibn Sina Robot	0.75	0.22	0.20
Many people from my home country would feel comfortable if they saw Ibn Sina Robot	0.64	0.13	0.42
Many people from my home country would feel angry if they saw Ibn Sina Robot	0.14	-0.04	0.80
Many people from my home country would feel afraid if they saw Ibn Sina Robot	0.04	0.10	0.80
Elderly people would learn a lot from robots like Ibn Sina	0.41	0.51	-0.06
It would be easier to learn about history from a robot like Ibn Sina than from a textbook	0.61	0.05	-0.32
Children would enjoy learning from a robot like Ibn Sina	0.66	0.01	0.02
Eigenvalues	2.10	1.94	1.78
Percentage of Variance	19.05	17.63	16.13

Figure 6. The rotated factor loadings for all 11 items on the questionnaire. Factor loadings above .40 are in bold.

Two of the respondents from Europe (Moldova and UK) were excluded from this analysis because of their small sample size, and the three people who did not identify their region were also excluded.

We performed a one-way ANOVA on the scores from the four regions, and there was a significant effect between them, $F(3, 122) = 3.192$, $p < .05$, $\omega^2 = .05$. Based on Kirk [8], this indicates a medium effect size between the four groups. A Hochberg posthoc analysis test revealed a significant difference in scores between people from the Gulf and African regions (2.87 , $p < .05$), however, the differences in scores between the other regions was not significant.

4.3.2 Age

After distributing our data into age groups as suggested by the brackets in Figure 7, a one-way ANOVA did not find a significant effect of age on score. The four people who did not identify their age were excluded from this analysis.

4.3.3 Gender

We did not find a significant effect between male and female respondents' scores. Three respondents who did not identify their gender were excluded from this analysis.

4.3.4 College Education

Next, we analyzed if having a college education influenced respondents' scores. We found that having a college degree resulted in a significantly lower score ($M = 31.7$, $s.d. = 6.0$) than not having one ($M = 33.8$, $SE = 6.1$), $t(117) = 2.73$, $p < .05$, $r = .06$. However, because our effect size was small this does not represent a substantive finding. The six respondents who did not specify their educational status were excluded from this analysis.

4.3.5 Conversation with Ibn Sina

Speaking to Ibn Sina before completing the questionnaire did not have a significant effect on respondents' scores. The 12 respondents who did not specify if they spoke to Ibn Sina were excluded from this analysis.

4.4 Qualitative Results

In addition to the quantitative data, we also received some qualitative data from the "Comments" section of the questionnaire as well as from verbal comments made by respondents. (English translations of Arabic comments are italicized.)

Some respondents said that they liked the idea of Ibn Sina robot and wanted to help improve it to be even more human-like:

« زيادة كمية المعلومات في الرجل الآلي، و تنمية قدرته على التعامل مع البشر »

"Add more variety to the types of things Ibn Sina can talk about and improve its ability to interact with humans."

On the other hand, a few people opined that they would rather prefer interacting with real humans than with robots. One person wrote:

« أعتقد أن الإنسان البشري أفضل بكثير من الإنسان الآلي وذلك إذا كان الإنسان متعلماً ناضجاً لأن الأحساس والمشاعر أفضل بكثير من أي شيء مع التعليم أيضاً والخبرة »

"I believe humans are better than robots. Humans are educated, mature, have feelings and emotions, and have experiences. Robots don't have this."

Other positive views were that some people expect that robots in the next generation will be more acceptable and popular in Emirates society. Many young people, especially young children and teenagers, love the idea of robots, and are willing to learn more about them and deal with them more often in their daily lives. For example, one person wrote:

« فكرة جيدة وجديدة وربما تلقى قبولاً أكثر في الأجيال القادمة »

"This is a good and new idea. Perhaps it will be acceptable among future generations."

A few people expressed views on the human-like form of the robot. Most of them were positive, but a few were negative. One negative comment was:

« الآلات هي وسيلة مساعدة ولا يجب أن تكون مستخدمة في حياتنا اليومية عوضاً عن خدمات الإنسان لأنه لا يوجد أعظم وأحسن صورة من خلق الله عز و جل »

"The main job of machines is to help humans, not replace them. No matter how human-like you make a robot, it will never be as good as one of God's creations." However, a number of people expressed positive comments with regards to robot appearance, such as:

« الهيكل فهو ممتاز »

"The body of the robot is excellent", and "It's nice. It's like a real human."

5 Discussion

Our pilot study found that Arabic attitudes toward humanoid robots are fairly positive overall. Respondents from the Gulf region (Iran, Iraq, Oman, Qatar, Saudi Arabia, United Arab Emirates, and Yemen)

had significantly more favorable views toward humanoid robots than respondents from African regions (Egypt, Morocco, Tunisia, Libiya, and Sudan) with a medium effect size. We also found that having a college degree was significantly associated with less favorable views toward humanoid robots, though with a small effect size.

In terms of sampling, we believe our study was representative of the Emirates population in terms of age, ethnic, religious, and cultural background. However, a main limitation of this study is that our sample was probably biased toward people who felt more positive and interested in robots in general, simply because people who were uninterested or felt strongly negative were less likely to stop by to see the robot and complete the questionnaire. On the other hand, the diverse range of regional backgrounds of respondents helped to provide good external validity. In the future we hope to replicate this study in other equally diverse regions and compare our results.

The questionnaire did not directly probe respondents' religious attitudes for two reasons. First, this kind of query is extremely difficult to do properly, and at this stage we merely wished to perform an exploratory study. Secondly, like other Asian cultures, Arabic culture is high-context, meaning people are less verbally direct, and communicate concepts in more subtle ways, using non-verbal behavior, indirect speech, and paralinguistic communication [6, 21]. Thus, asking people directly about their religious views on robots seemed unsuitable at this time. However, this is a topic we wish to delve deeper into in future work.

The qualitative data were interesting in that some of the respondents' views echoed the types of views expressed by subjects in other studies and other countries. For example, people preferred robots not to take the place of humans, but rather to serve as a means of helping improve their lives. This echoes findings by Takayama et al. [18] and Weiss et al. [20], and warrants further investigation.

We have recently collected a second sample of several hundred questionnaires from people who interacted with Ibn Sina at a trade show in Dubai and will analyze this data in the coming months. Using this data, we plan to refine the CEDAR questionnaire by further testing its validity, consistency, and test-retest reliability. We also plan to perform more in-depth data analysis, such as exploring the instrument's subscales across different participant groups.

Acknowledgment

This work is sponsored in part by the Qualcomm Research Studentship in Computing.

We would like to thank Aziz Al Kayoumi, Wajahat Kazmi, Saif Al Ketbi, Panos Toulis, and Astrid Weiss.

REFERENCES

- [1] AcapelaGroup. Acapela arabic speech recognizer and text-to-speech engine. <http://www.acapela-group.com>, Accessed 8 August 2009.
- [2] A. Albirini, 'Teachers' attitudes toward information and communication technologies: the case of syrian efl teachers', *Computers & Education*, 47(4), 373 – 398, (2006).
- [3] C. Bartneck, T. Suzuki, T. Kanda, and T. Nomura, 'The influence of people's culture and prior experiences with aibo on their attitude towards robots', *AI & Society*, 21(1), (2006).
- [4] BrookshireSoftware. Visual studio automation software. <http://www.brookshiresoftware.com>, Accessed 8 August 2009.
- [5] A. Field, *Discovering Statistics Using SPSS*, SAGE Publications, London, 3rd edn., 2009.
- [6] G.J. Hofstede, *Cultures and organizations: Software of the mind*, McGraw-Hill Publishing Co., 2005.

- [7] F. Kaplan, 'Who is afraid of the humanoid? investigating cultural differences in the acceptance of robots', *International Journal of Humanoid Robotics*, **1**(3), (2004).
- [8] R. E. Kirk, 'Practical Significance: A Concept Whose Time Has Come', *Educational and Psychological Measurement*, **56**(5), (1996).
- [9] N. Mavridis and D. Hanson, 'The ibnsina center: An augmented reality theater with intelligent robotic and virtual characters', in *Proc. of IEEE Robots and Humans (RO-MAN)*, (September 2009).
- [10] N. Mavridis, W. Kazmi, P. Toulis, and C. Ben-Abdelkader, 'On the synergies between online social networking, face recognition and interactive robotics', in *Proc. of IEEE Conf. on Computational Aspects of Social Networks (CASON)*, (2009).
- [11] N. Mavridis and D. Roy, 'Grounded situation models for robots: Where words and percepts meet', in *Proc. of IEEE Conf. on Robots and Systems (IROS)*, (2006).
- [12] T. Nomura, T. Kanda, T. Suzuki, J. Han, N. Shin, J. Burke, and K. Kato, 'Implications on humanoid robots in pedagogical applications from cross-cultural analysis between japan, korea, and the usa', in *Proc. of IEEE Robots and Humans (RO-MAN)*, (August 2007).
- [13] T. Nomura, T. Kanda, T. Suzuki, and K. Kato, 'Prediction of human behavior in human-robot interaction using psychological scales for anxiety and negative attitudes toward robots', *IEEE Transactions on Robotics*, **24**(2), 442-451, (April 2008).
- [14] E. M. Rogers, *Diffusion of innovations*, The Free Press, New York, 4 edn., 1995.
- [15] M. Saif, 'The evolution of persian thought regarding art and figural representation in secular and religious life after the coming of islam', *Macalester Islam Journal*, **1**(2), (2006).
- [16] D. W. Straub, K. D. Loch, and C. E. Hill, 'Transfer of information technology to the arab world: a test of cultural influence modeling', in *Advanced topics in global information management*, ed., Felix B. Tan, 141-172, IGI Publishing, Hershey, PA, USA, (2003).
- [17] Dag S. Syrdal, K. Dautenhahn, K.L. Koay, and M. L. Walters, 'The negative attitudes towards robots scale and reactions to robot behaviour in a live human-robot interaction study', in *In Proceedings of the AISB Symposium on New Frontiers in Human-Robot Interaction*, (2009).
- [18] L. Takayama, W. Ju, and C. Nass, 'Beyond dirty, dangerous and dull: what everyday people think robots should do', in *HRI '08: Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, (2008).
- [19] R. M. Thomas, 'Computer technology: an example of decision-making in technology transfer', in *Educational technology - its' creation, development and cross-cultural transfer*, eds., R. M. Thomas and V. N. Kobayashi, 25-34, Pergamon Press, Oxford, (1987).
- [20] A. Weiss, R. Bernhaupt, M. Tscheligi, and E. Yoshida, 'Addressing user experience and societal impact in a user study with a humanoid robot', in *In Proceedings of the AISB Symposium on New Frontiers in Human-Robot Interaction*, Edinburgh, UK, (2009).
- [21] E. Wurtz, 'Intercultural communication on web sites: A cross-cultural analysis of web sites from high-context cultures and low-context cultures', *Journal of Computer-Mediated Communication*, **11**(26), (November 2005).

6 Appendix

	Strongly Disagree	Slightly Disagree	Slightly Agree	Strongly Agree
I wouldn't mind if a human-like robot treated me at the hospital	1	2	3	4
I wouldn't mind if a human-like robot cleaned my house	1	2	3	4
I wouldn't mind if a human-like robot was in my workplace	1	2	3	4
I wouldn't mind if my child was instructed by a human-like robot	1	2	3	4
Many people from my home country would feel happy if they saw Ibn Sina robot.	1	2	3	4
Many people from my home country would feel comfortable if they saw Ibn Sina robot.	1	2	3	4
Many people from my home country would feel angry if they saw Ibn Sina robot.	1	2	3	4
Many people from my home country would feel afraid if they saw Ibn Sina robot.	1	2	3	4
Elderly people would learn a lot from robots like Ibn Sina.	1	2	3	4
It would be easier to learn about history from a robot like Ibn Sina than from a textbook.	1	2	3	4
Children would enjoy learning from a robot like Ibn Sina.	1	2	3	4

غير موافق بدرجة	غير موافق	موافق	موافق بدرجة
1	2	3	4
أنا لا أمانع أن يقوم إنسان آلي بالأعتناء بي في المستشفى			
1	2	3	4
أنا لا أمانع أن يقوم إنسان آلي بتنظيف منزلي.			
1	2	3	4
أنا لا أمانع إذا تواجد إنسان آلي في مكان عملي			
1	2	3	4
أنا لا أمانع إذا تلقى ابني تعليمه من الإنسان آلي			
1	2	3	4
معظم الناس في بلدي سوف يتبعون بالسعادة إذا راوا رجل الآلي "كاين سينا"			
1	2	3	4
كثير من الناس في بلدي سوف يتبعون بإتراءه إذا راوا رجل الآلي "كاين سينا"			
1	2	3	4
كثير من الناس في بلدي سوف يتبعون بالغضب إذا راوا رجل الآلي "كاين سينا"			
1	2	3	4
كثير من الناس في بلدي سوف يتبعون بالخوف إذا راوا رجل الآلي "كاين سينا"			
1	2	3	4
كبار السن قد يتعلمون الكثير من رجل الآلي "ابن سينا".			
1	2	3	4
سيمكن تعلم التاريخ من خلال الرجل الآلي "ابن سينا".			
1	2	3	4
أسهل من تعلمه عن طريق الكتب.			
1	2	3	4
أسهل الأطفال للتعلم من رجل آلي مثل ابن سينا.			

Figure 9. The English and Arabic versions of the Culture, Education, and Domestic Attitudes Toward Robots Scale (CEDAR).

Robot learning of lexical semantics from sensorimotor interaction and the unrestricted speech of human tutors

Joe Saunders, Chrystopher L. Nehaniv and Caroline Lyon¹

Abstract. This paper describes a HRI case study which demonstrates how a humanoid robot can use simple heuristics to acquire and use vocabulary in the context of being shown a series of shapes presented to it by a human and how the interaction style of the human changes as the robot learns and expresses its learning through speech. The case study is based on findings on how adults use child-directed speech when socially interacting with infants. The results indicate that humans are generally willing to engage with a robot in a similar manner to their engagement with a human infant and use similar styles of interaction varying as the shared understanding between them becomes more apparent. The case study also demonstrates that a rudimentary form of shared intentional reference can sufficiently bias the learning procedure. As a result, the robot associates human-taught lexical items for a series of presented shapes with its own sensorimotor experience, and is able to utter these words, acquired from the particular tutor, appropriately in an interactive, embodied context exhibiting apparent reference and discrimination.

1 INTRODUCTION

Humans acquire language quite effortlessly during their early years and continue to refine and develop their language ability throughout their lives. There are many alternate views on how language is acquired varying, for example, from the Chomskyan [4] idea of existing innate pre-cursors of syntactic structures and mechanisms which support the acquisition process to alternatives suggesting that that no ‘specialist’ functions exist and where more general learning mechanisms, including learning from social interaction, are used [24]. In our work we explore this latter view and consider that language learning is a social and interactive process. We review research on how language is acquired in human infants and suggest that similar mechanisms might be applied to robots in order for them to acquire linguistic capabilities. In fact we believe that it should be possible for a robot to use what it learns individually and from others socially through grounded sensorimotor interaction to bootstrap the acquisition of language which in turn will create a positive feedback cycle between using language and developing other cognitive abilities.

In this paper we report on an initial case study where, through unrestricted speech and interaction with a human partner, a robot can attach meaning to a series of individual shapes. To do this we firstly consider how human infants acquire such meanings and then carry out experiments to see whether broadly equivalent mechanisms will allow a robot to do the same. Two research questions are explored, firstly, will a human interaction partner employ similar and effective

interaction styles (such as child directed speech) with a robot partner?, and secondly, could a robot be able to extract sufficient information from such interaction episodes in order to ground the meaning² of the lexical items (words) used when taught by particular humans about a set of physical shapes?

2 BACKGROUND

In this research we take a ‘usage-based’ view on language acquisition following Tomasello [24] and Bloom [3] and take inspiration from the constructivist work of researchers such as Steels on the emergence of various linguistic capacities in agent communities [21], and also Roy and Pentland [18], Roy [17], and Yu and Ballard [27] in attempting to link perceived speech with object and action perception. We differ from the above in that the focus of this research considers the effect of the robot feedback in modifying the interaction style of the human tutor and how this change enhances (or worsens) the learning experience of the robot.

Social interaction appears to be essential for effective language acquisition in children [19]. Adults do not specifically attempt to ‘teach’ language to children but rather aim for joint understanding with the child. This often comes about by the adult using a form of language called Child Directed Speech (CDS) which is specifically tailored to the perceived level of linguistic skill of the child [5]. CDS differs from adult directed speech in that adults speak more slowly, often repeat what is said and may correct mistakes by the child during the interaction as well as prominently drawing attention to topics via changes in pitch, pause duration, word placement, word duration and intonation. As joint understanding between the child and adult is perceived by the adult to grow, fewer aspects of CDS are employed. However, although infants are exposed to a high number of ambient speech events, being exposed to as many as 7,000 utterances per day [3], they actually learn new words with very little exposure. According to Bloom [3] the natural co-occurrence of seeing an object and hearing the object name does not occur regularly enough to allow an associative mechanism to differentiate the object. In order to achieve learning in few episodes the infant must have its learning experiences biased in some way. This is thought to occur via intentional reference. Thus for example 9-month old babies will follow the gaze of its mother, follow her pointing gestures and monitor her emotional states. By 1 year old, the infant points on its own and observes the adult’s gaze whilst checking if it has changed the adult’s attention. If the child fails to capture the adult’s attention it will alternate between gazing at the object and the adult until it succeeds in getting the

¹ Adaptive Systems Research Group, University of Hertfordshire, College Lane, Hatfield, Herts, AL10 9AB, UK, email: j.1.saunders@herts.ac.uk

² Following Wittgenstein [26], any derivation of ‘meaning’ must ultimately be evidenced by appropriate embodied action in language games (cf. [12]).

attention of the adult onto the object. The utterances of the adult together with reinforcement via forms of affective feedback (prosodic features of CDS) and success/failure of shared intentional reference for the infant whilst situated in context allow enough bias for fast learning to take place. Within this shared context the action modalities of the child are associated with simple verbal speech patterns of the adult. In our experiments learning relies on the association of perceived speech and action modalities of the robot together with the robot's ability to share context via a form of rudimentary referential 'intent' with a human teacher.

In order for our robot to be attentive to the topic spoken by the human we currently only consider *pause duration* and *word placement and word duration*, although in future research we will also consider prosody via changes in pitch and intonation. In terms of word placement and duration we make the assumption that regularities can be extracted from the human's speech stream. In English (and other languages) attention is drawn to new nouns by placing them at the end of an utterance (reflecting the subject-verb-object structure of the language) [5, pages 62-66]. There is also evidence that a similar phenomenon occurs in CDS in other languages (e.g. Turkish) even when to do so would appear to be ungrammatical or uncommon in adult speech for that language [2]. As well as indicating salience via utterance-final word placement, often the word is also spoken with a longer duration than average during the utterance. We therefore use both the duration and placement of a word in the perceived speech matched with the sensorimotor stream experienced by the robot in order to bias the robot's learning ability.

Clearly it would be trivial to provide the robot with a set of pre-programmed rules attaching a label to a given perceived object, however our research attempts to allow the robot to acquire lexical semantics ('meaning') of words used by a human tutor for the objects (in terms of the set of sensory-motor experiences over time) dynamically through the interaction process. We believe this is an important step not only towards the robot understanding objects in its world (the focus of this study) but also understanding actions and states (e.g. 'push the star', 'the star is next to the square'). This latter step forms part of our future research and which we hope will provide entry into basic grammatical constructions [24].

3 RESEARCH QUESTIONS

In this study we considered two issues, the first from a human-robot interaction perspective and the second from the perspective of learning efficiency (in terms of classification performance) of the robot during the sessions. Exploring the first issue, we expected that the human interaction partner would employ a language style similar to that employed in child directed speech and that with each subsequent interaction session the feedback from the robot (the robot in effect expressing what it had learnt during the previous sessions) would motivate the human partner to modify the interaction, supplying greater or lesser focus as the abilities of the robot improved. Note that although the participants were specifically asked to treat the robot as a 1-2 year old child (see 4.1 below) and the robot had the appearance of a young child, it was by no means certain that the interaction style would 'mimic' CDS as other factors which may provoke CDS with human children (such as emotional attachment and richness of interaction) may be missing from the interaction with the robot. Thus the first two research questions were:

1. Does a human interaction partner engage with a humanoid robot in a verbal style similar to that of Child Directed Speech?

2. Does a human partner modify aspects of their robot-directed speech as the robot learns and changes its interaction capability?

The second issue explored was that of the robot's capacity to learn, given the interaction style of the human. We expected that the robot would become more skilled as the human partner interacted with it which may in turn motivate the human. This skill level of the robot indicating that the multi-dimensional set of sensorimotor attributes had become refined such that relevant meaning emerged. This latter aspect being measured by a number of factors, including shape classification performance and increasing mutual information of particular sensorimotor attributes. Thus the final research questions explored were:

3. Does a robot's ability to recognize and correctly classify shapes by its utterances improve during the interaction sessions?
4. Can a robot select those parts of the sensorimotor stream that are relevant to classifying presented shapes by acquired vocabulary?

4 CASE STUDY DESCRIPTION

In this section we describe both in summary and in subsequent detail the experimental procedures in the case study.

4.1 Experimental Summary

In this experiment we asked participants (see 4.3) to explain/teach (without any constraints on their language) a series of shapes (see 4.5) to a humanoid robot during a series of interactions *as if* the robot were a 1-2 year old child (see figure 2). The robot was pre-programmed to track and habituate for a given period on these shapes. Following each interaction the speech stream of the human was converted into phoneme strings marked with word boundaries. These phoneme strings were subsequently aligned with the sensorimotor modalities experienced by the robot during the interaction session. The aligned speech and sensory modalities were then processed to highlight words of long duration and that appeared at the end of utterances. This processed modality stream became the basis for the robot's learnt experiences for the next interaction session with the human. In these subsequent sessions (from sessions 2 onwards) the robot was allowed to match its current sensorimotor input (that it was experiencing during the interaction) against that learnt in the previous session(s). This allowed the robot to react to the human by expressing (via its own speech) what it had learnt during the previous session(s). Thus, for example, when presenting the 'moon' shape and depending on how the human had previously described it, the robot might say 'moon'. Alternatively if the human had used different words (or different pronunciations) the robot might say for example 'mooone', 'crescent' or 'smile' etc.

4.2 Software Platform and Architecture

In carrying out these studies we have used an existing social learning architecture (ROSSUM [20]). The architecture has been used to learn scaffolded behaviours via directed learning from a human teacher (but not previously linguistic behaviour). ROSSUM allows experiments to be carried out where, via the processes of self-imitation and observational imitation, a robot learns through experiences grounded³ on its own visual, auditory and sensorimotor feedback and the relevant interaction modalities presented by a human

³ The idea of 'grounding' refers to relating the meaning of symbols to embodied sensorimotor experience [9].

tutor. In this experiment *speaking* is regarded as a particular type of *gesturing*, i.e. motor activity for manipulating the physical or social environment (see [13, 15]). Actions which can be taught to the robot (see figure 1 below) now include speech acts (in this case uttering words represented as phonemic strings in a learnt experiential context). These modalities are in addition to existing object detection and proprioceptive feedback modalities already present in the ROSSUM architecture (see 4.4).

4.3 Participants

Eight adult participants took part in the case study. All participants were between the ages of 27 and 58 comprising 5 females and 3 males. The backgrounds of the participants were either administrative (6) or research related (2), the latter not connected with robotic language research. Four of the female participants cared for either children or grandchildren under the age of 5.

Each of the eight participants took part in 5 interaction sessions of approximately two minutes with the robot (in total 40 robotic interaction sessions) and all of the sessions were videotaped for later analysis. Prior to the interaction sessions each participant also had 4 training sessions with an automatic speech recognition system. Unfortunately initial results from using this system were very poor and an alternative was subsequently used (see 4.6 below).

The experiment was carried out over a three month period between March and June 2009 based on availability of the participants. Participants were paid a small stipend of £20 if they completed all 9 sessions (4 training + 5 interaction) and all participants did complete all sessions.

4.4 Robot

The experiment used the Kaspar2 robot (see figure 1). This is a minimally expressive small humanoid robot designed for HRI studies [7] with 8 degrees of freedom in the head, and 5 degrees of freedom in each arm (although the arms were not used in this experiment).



Figure 1. Learning via Social Interaction. Here (in a previous experiment) the humanoid robot Kaspar2 is taught how to hold a patterned box via interaction (manipulation of its joints) with a human teacher.

Kaspar2 employs two video cameras for its eyes. These are used to obtain images which are processed [14] to yield an additional set of modalities including object/pattern recognition and face detection. In this experiment the following set of sensory-motor modalities were recorded by the robot at approximately 30ms intervals - face detection (binary), object id, head pan, head tilt, head roll, position of object in visual field (x,y position of centre of object) and distance to object.

4.5 Shape recognition

Six graphic shapes were used and pasted onto the sides of a box (a square, circles, triangle, sun, moon and star - note however that no verbal prompting or description of these shapes was given to the participant). These objects are detected using the ARToolKit system [1]. The objects are pre-learned using ARToolKit, thus Kaspar2 can detect these objects and recognize that they are individual entities in the world (effectively encoded by the integers 1-6 in a sensory stream), however *no other 'meaning'* is attached to them. We justify the use of this simplifying step in these experiments as, firstly, it eliminates the need for a complex vision processing modality and secondly, and more importantly, it reflects a 'whole object' bias found both in children and adults (see Bloom, Chapter 4 [3] and Mervis [10]).



Figure 2. Sharing Reference with a Teacher in Context. Kaspar2 interacts with the teacher, where both the robot and the human share attention on the shapes on the box, a case of rudimentary shared intentional reference.

4.6 Speech Processing

Initially a standard speech recognition system [11] was used to process the speech stream of the human. However it was found that the accuracy of such a system rapidly decreased when dealing with the unrestricted speech stream used when interacting and teaching the robot. An alternative approach was taken which used speech alignment software (see [22]). This software takes as input a transcript of the human's speech during the interaction session together with the recorded speech stream (as a .wav file). It outputs timings, phonemes and word boundaries for the recorded speech.

4.7 Rudimentary Shared Intentional Reference

In the experiment it was important that the robot and human share attention and focus on the same objects whilst simultaneously provoking actions which would motivate the human teacher to explain the shapes. We make a key assumption - that the robot is motivated by novelty. This was achieved by having Kaspar2 fixate and track shapes as they entered its camera images. When a shape was not present Kaspar2 would search the environment semi-randomly (randomly within fixed limits) by varying the control signal sent to its pan/tilt/roll head motors. If a shape was found Kaspar2 would 'smile' and fixate on that object. A set of linear 'boredom' filters, one for each shape and for the human's face, would saturate if the same shape (including the face) remained in focus for more than ~20 seconds. A saturated filter would cause the robot to begin searching for a new object (and Kaspar2 would stop smiling). The filters would linearly

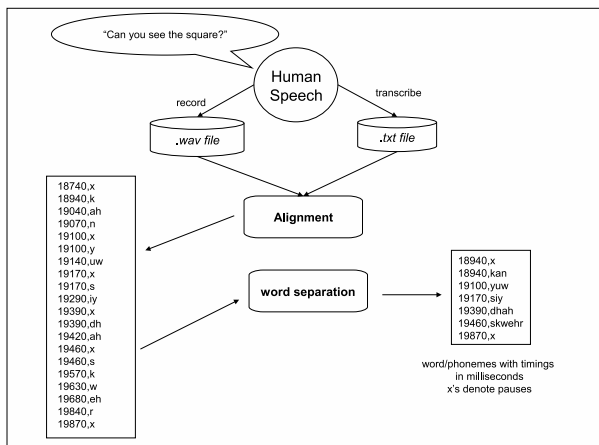


Figure 3. Segment of Words/Phonemes output. The human's speech stream is converted to a file of timed phoneme/word outputs (x's denote word boundaries). Note that different speech patterns can produce different phoneme strings for the same word e.g. different vs. diffrent (i.e. middle 'e' not expressed).

de-saturate the longer the object was not in focus. The net effect of these measures was for Kaspar2 to appear to be interested in the objects/face but become rapidly bored if viewed for too long. This had the effect of provoking the human teacher to try to gain the attention of the robot with new shapes and subsequently explain them.

4.8 Attaching Meaning to Words for Shapes

In this context we consider that 'meaning' of a communicatively successful utterance is grounded in its usage based on the robots sensorimotor history from acting and interacting in the world. These grounded meanings can then be scaffolded via regularities in the recognized phoneme/sensory-motor stream of the robot. The first step in this process is to merge the speech stream of the human (represented as a set of phonemes with word boundaries) with the robot's sensory-motor stream. This is achieved by matching the two modalities based on time (see figure 4).

This merger effectively associates what was said to the robot with what was experienced by the robot at that time. Within this study the set of sensorimotor attributes is limited to the robot's head and vision proprioception together with the recognition of pre-trained shapes. However although the categorization of the shapes is already available (as an integer) the robot has not associated this attribute with either the words in the human's speech stream or its own proprioception. In this study sensorimotor attributes which do not affect the primary association of object and word (such as the head and image proprioception) should become less relevant over time (i.e. a 'moon' shape is still a moon shape no matter where or how it is seen). To achieve such associations we face a number of difficult challenges. Firstly is that of associating what was said to the appropriate parts of the sensorimotor stream. Thus the human tutor may show the robot a shape (e.g. the 'triangle'), but only say the word 'triangle' within the utterance before or after the shape has appeared/disappeared from the view of the robot (e.g. 'here's a triangle' and then show the triangle or 'that was a triangle' after having shown the triangle). Sec-

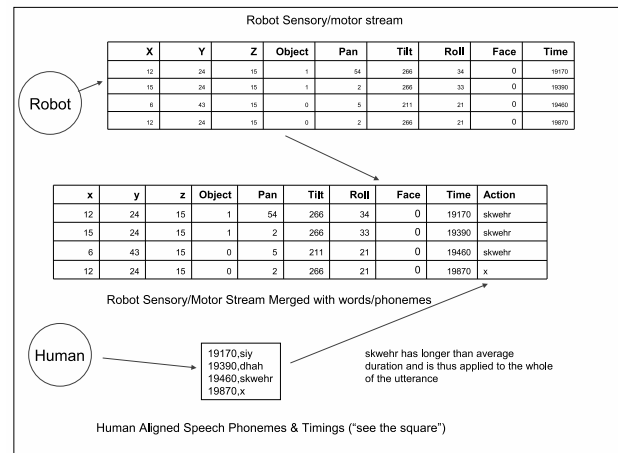


Figure 4. Aligning Human Speech Words/Phonemes with Robot Sensorimotor Streams. The timed word/phoneme strings are aligned with the robot's sensory-motor stream based on time. Here the word 'square' is an end of utterance word with higher than average duration and so is propagated over the whole of the preceding utterance and concomitant sensorimotor experience. This results in a table subsequently used for sensorimotor similarity matching (using k-Nearest Neighbour) in the robot's execution phase. The 'action' executed in this case would be a robot speech output (i.e. if similar to the current sensorimotor state the word 'square' would be expressed by the robot).

ondly, which of the set of sensorimotor attributes and at which points in time are such attributes relevant to the speech act? We make no pre-programmed choices as to what is relevant for the robot however in the simple scenario presented here the most important will be the object identifier returned from the ARTToolKit system. But we have not only the object identifier but also the modalities from the camera image (x,y,z dimensions of the object in space), the encoder readings from the robot's pan/tilt/roll head unit and the binary value of the face detection algorithm. What makes any of these less (or more) relevant than any other? In order to cope with these issues we apply two heuristics. Firstly, that the relevant part of the human speech stream is the word expressed at the end of an utterance and with a higher than average duration (following the ideas discussed in section 2 above). We extract a word if it occurs at the end of an utterance (the end of an utterance defined as a pause longer than average word duration). This word is then remapped back onto each element of the time sensorimotor stream of the previous utterance. In effect this makes the word chosen relevant to the whole of that utterance and thus relevant to any sensorimotor inputs which arose during that time. This means that most (though not all) of the timing effects mentioned above can be avoided as the relevant word will occur throughout the utterance. Secondly, having associated the relevant word with the utterance/sensorimotor stream we then compute the *information gain* [16] between the sensory attributes and the chosen word (effectively a measure of *mutual information* indicating the expected amount of information (in bits) that discriminates the given word by the given attribute). This measure is used during the next interaction sessions to weight the similarity measure of current vs. stored experiences. This is described in section 4.9 below. Note, in our experiments, the robot learnt lexical semantics separately from each participant so that learning occurred in effect as if each partic-

participant had their own robot that had learnt only from them. This allowed us to analyse the diversity of learning trajectories, dependent on interactions with particular tutors (who could in principle be using completely different lexicons or even different languages).

4.9 Execution and Action Selection

We consider robot speech acts (including ‘saying’ nothing) to be no different from any other ‘action’ (such as primitive head or arm movements) that may be selected for execution. Therefore an existing selection method, based on matching the current sensory stream against a set of pre-learned/pre-taught behaviours, is used [20]. All previously learnt action/state sequences are held in *memory models*. These are tables in which each individual row is a state/action pair. Execution starts at the topmost memory model and recurses through the possible set of models until a primitive (or marked goal state) is found. For the language experiment discussed here only one memory model is used. This equivalently contains the set of state/action pairs learnt by the robot (see 4.8 above). Here the ‘action’ is uttering a phoneme string representing a word. The phoneme strings are output by the robot using the eSpeak speech synthesizer [8].

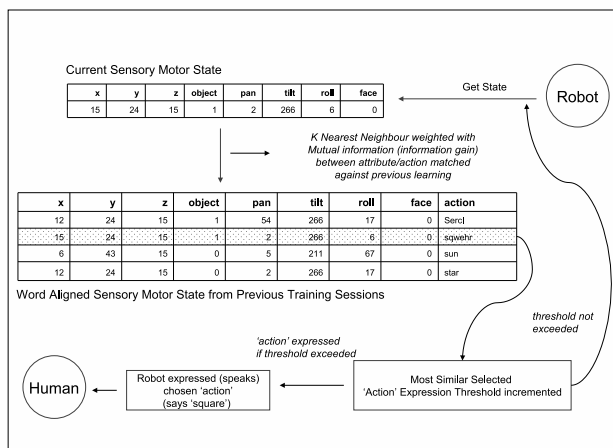


Figure 5. Interaction Execution Loop. The robot continuously monitors its state. This is matched against the table previously learnt using the k-Nearest Neighbour algorithm. Dimensional attributes are weighted using information gain. The nearest neighbour is selected. The ‘action’ part of the selected previous learning will only be expressed (spoken) by the robot if it has been chosen a given number of times (exceeds a predefined threshold). Note that the execution loop and the learning loop (where the robot’s proprioceptive state from the current session is recorded) happen simultaneously.

The robot continually polls its current state (see figure 5). This state is matched against the memory model using the k-Nearest Neighbour (kNN) algorithm with dimensional attributes weighted using information gain. We use the TimBL memory based-processing system to carry this out [6, 23]. In this experiment the k value was set to 1 and similarity metric used is a 1-norm Manhattan distance. The kNN algorithm always returns a classification, in this case an ‘action’. This is effectively a prediction based on matching what the robot experiences and what the robot should then ‘say’. The

robot state changes rapidly during the interaction and therefore rather than simply express each matched state the action (a phoneme/word string) will only be expressed once it exceeds a given threshold. The selected action is incremented if matched and all other actions penalized. If the same action is predicted successively it receives a bonus which serves to reward consistent prediction (for complete details of this scheme see [20]). The net effect is that by lowering the threshold more matches may be expressed by the robot (and thus similar words matching the same state may be output (e.g. participants might have described the ‘square’ as both a ‘square’ and a ‘box’, the robot may choose either). Raising the threshold means that fewer words are expressed and the similarity criteria to the current set of states is higher (e.g. only ‘square’ would ever be expressed). In this experiment words are output (including silence) with a threshold, reward, penalty and bonus set during pre-experimental tests providing a balance between an overly talkative robot and one who said very little.

5 RESULTS AND ANALYSIS

As discussed in section 2 above, CDS is characterized by slower speech, repetition, mistake correction, prosody, duration of pauses and words and word placement. In our analysis we analysed speech rate, repetition, word placement and word duration. The latter four items based on an analysis of transcripts and recordings as well as considering the number of nouns extracted using the heuristics discussed in section 4.8 above. In order to assess learning performance of the robot we studied how well it classified words and how well items in the sensorimotor stream separated to focus meaning on the correct attribute. These are discussed below.

5.1 Speech Rate

Considering first the speech rate, this was calculated in words per minute (WPM) based on the number of words used by each participant in each session. The WPM varied on average between 80 and 132. However, 5 of the participants slowed their speech during the set of interactions, 4 ending with a lower WPM in the final session than in the first session. The other 3 participants varied their speech rates during the sessions but ended with a higher WPM by the final session (see figure 6).

Although the small participant sample size does not allow any firm conclusions to be drawn, indications of gender and child caring experience may be relevant for future larger scale studies. The 5 participants who slowed their speech rate were all female, 4 with direct child caring experience whereas the other participants were all male with no direct child caring experience. This may indicate that teaching a robot as if it were a young child may be more amenable for those already predisposed to such undertakings with a human child.

5.2 Repetition

An objective measure of repetition is to assess how many new words (of all types including nouns) are added between interaction sessions. If the number of new words decreases then it is likely the same words are being used between sessions. Similarly an analysis of words dropped (used in one session but not the next) would indicate, if increasing, that words are not being repeated, if decreasing, then the same words are being reused. Figure 7 shows the average number of new and dropped words between the sessions and indicates that repetition is being used (and increases) during the sessions. Note however that there is a sharp increase in the number of

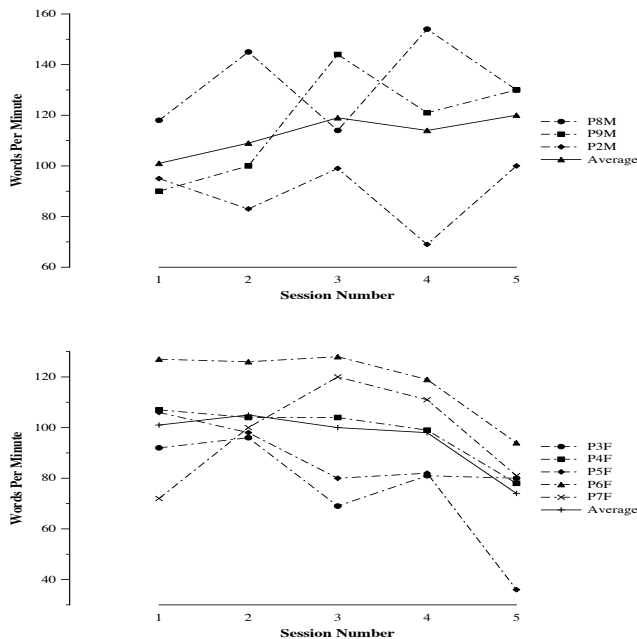


Figure 6. Speech rate. The graphs shows the speech rate for participants over the five session. Top graph (all male) shows words per minute (wpm) varying with a general small increasing trend. Bottom graph (all female) shows wpm general trend decreasing. Trends line based on average by session shown as solid line.

words dropped between the first and second sessions. This may indicate that once the robot starts to speak the human partner realizes the limited extent of its understanding and makes adjustments, in this case reducing the size of his/her expressed vocabulary.

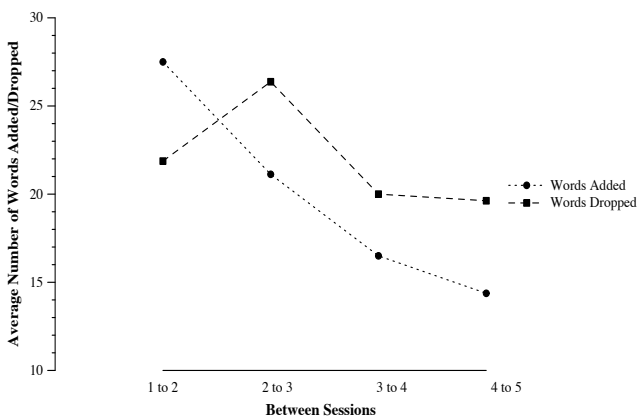


Figure 7. Words Added/Dropped Between Sessions. The average numbers of words added and dropped between sessions for all participants. The number of new words decreases between all of the sessions. The number of words dropped increases sharply between the first and second sessions and then starts to decrease (see text).

5.3 Word Duration and Placement

To assess whether participants actually placed the salient words referring to shape description at the end of their utterances and with a longer than average duration we considered the performance of the heuristic described in section 4.8 above. We compared the words extracted by the heuristic against all possible nouns used by each participant in each session. If the participants regularly placed the shape name at the end of an utterance (and with higher than average duration) the extraction rate should be higher than average. Figure 8 shows the average for this measure for each participant. The results indicate that for all participants this effect occurred with the majority placing nouns at end of utterance and with longer than average duration over 80% of the time.

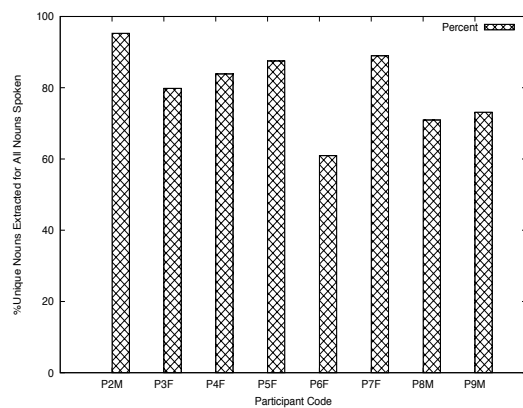


Figure 8. Noun Extraction. The graph shows for each participant the performance of the heuristic of extracting words based on duration and end of utterance placement.

5.4 Classification Performance

To assess the robot's learning ability each of the memory models was analysed using leave-out cross validation [25, pages 151-152] to predict classification performance (see figure 9). This method was used as it gave an unbiased and objective measure of classification accuracy over all of the participants sessions.

For all participants, performance according to the classification ability acquired by the robot had improved by the fourth session but started to worsen by the fifth session. This decrease in classification performance is due to the changes in the interaction style of the human. By session 4 the participants had noticed that the robot had started to correctly and consistently recognize the shapes, thus by session 5 their emphasis changed such that rather than place nouns at the end of the sentence various forms of reward utterances were used. For example, rather than continue saying 'this is a square' when the robot correctly and consistently recognized the square (and said 'square') the human would typically say 'well done' or 'good boy' or 'clever Kaspar'. This has the effect of diluting the previous association between the shape and the 'correct' word and thus reduced the classification performance. In this experiment we were not attending to such reinforcement signals, however the effect emphasized the fact that the participants were altering their behaviour in line with their perception of shared understanding between themselves and the robot.

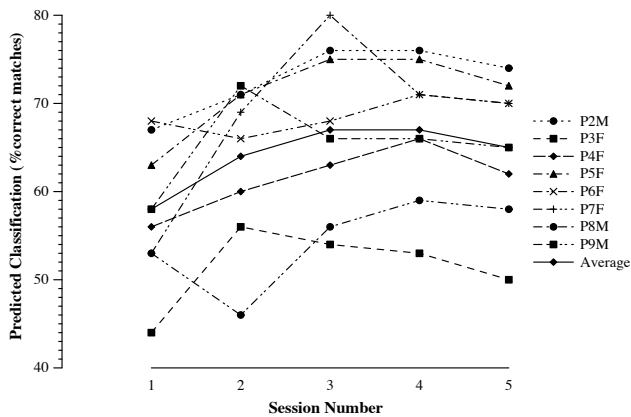


Figure 9. Predicted Classification Performance. Classification performance by participant generally increases between the first and the fourth sessions. A decrease in performance between sessions 4 and 5 indicates that the human focus of the interaction has changed (see text).

For the majority of the participants, there was also registered an increase in classification performance between the first and second sessions. This may be because the first feedback signals are being received from the robot in the second session (i.e. in the first session the robot is silent, in subsequent sessions it starts to speak). The robot speech feedback effectively signalling its level of understanding to the human tutor who subsequently adjusts his/her interaction appropriately. An alternative explanation is that participants had simply become familiar with the experiment, however participants P8M and P6F were the exceptions with a large performance drop of 13% for P8M and a marginal 2% drop for P6F, respectively. We believe that the large decrease exhibited by participant P8M was due to his interaction style i.e. not being engaged in the interaction. This behaviour becomes more apparent in the analysis below.

5.5 Identifying Meaningful Sensorimotor Attributes

Considering the fourth research question above (section 3). Here we are assessing whether sufficient information is available through the interaction to correctly enhance the key sensorimotor attributes (here the 'object id' returned from the ARTToolKit system) and thus weight for likely selection via the kNN algorithm. This is effectively indicating that a particular set of attributes are more closely associated with the given word and, in our interpretation, more 'meaningful' to the robot. In figure 10 we show an example of this effect in detail for one participant (P5F), however the separation occurred for the majority of the participants in the study with varying degrees of success.

Figure 11 shows a summary for all participants. In this graph the values for 'object id' and the head and image proprioception sensorimotor attributes have been averaged over the 5 sessions. Note that for all but 2 participants (P4F and P8M) the meaningful variable (object id) separates from the other less meaningful variables. Participant P8M's interaction with the robot was characterized by both the highest speech rate (a WPM value of 132 vs an average of 102), the highest number of unique words used per session (90 vs an average of 60 for the other participants) and the highest number of nouns used per session (22 vs. an average of 13). Our interpretation of these figures suggests that the participant had not changed

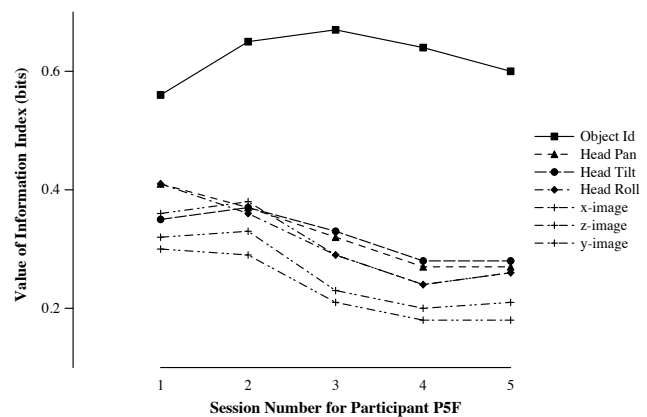


Figure 10. Separation of Meaningful Variables. The graph shows how during the sessions the sensorimotor attributes which contributes most to the meaning of the word separate from other less meaningful attributes. This is an example from participant P5F.

his interaction style to accommodate the robot's understanding. This is also reflected by the drop in classification performance over the initial sessions noted above (see 5.4).

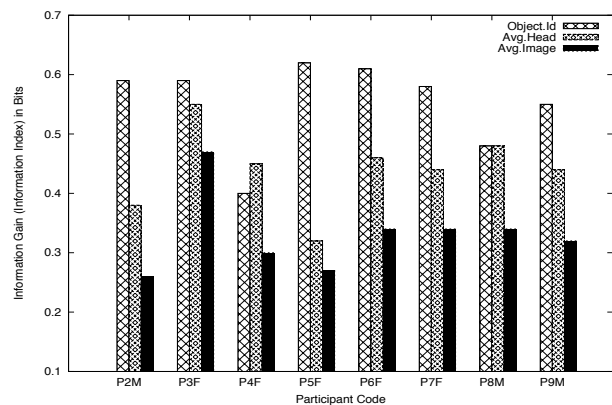


Figure 11. Separation of Meaningful Variables By Participant. For each participant the separation of the average value of the object sensorimotor attribute vs. average head and image proprioception attributes. The key sensorimotor attribute weighting the classification of words via the kNN algorithm is the object id. The object id weighting separates from the other sensorimotor attributes as the sessions progress.

Participant P4F achieved a classification performance similar to the other participants (average of 0.61 vs 0.64 for other participants) however the separation of the key object id attribute differed considerably. This would indicate that although the participant had engaged with the robot effectively she had not drawn sufficient attention to the object (thus she spoke about the object when the robot was attending elsewhere). The classification therefore indicated that the robot had made an association between head proprioception and the phoneme/word string rather than object and phoneme/word string. This was confirmed by subsequent review of the video for this par-

ticipant and showed Kaspar2 ‘saying’ words regardless of where it was attending.

5.6 Summary

Returning to the original research questions, we considered whether firstly the human interaction partner would engage in a verbal style similar to that of Child Directed Speech. Indications are that the participants generally talked more slowly and were more repetitive emphasizing the objects via word placement and word duration. Secondly, we asked whether their robot-directed speech would change as the linguistic capability of the robot improved. There did indeed appear to be a definite change once it was accepted by the human partner that shared understanding between themselves and the robot had been achieved. Thirdly, in considering the robot’s capacity to learn it was clear that the robot’s ability to recognize and correctly classify shapes improved during the initial 4 sessions but decreased in the final session. We believe that this final decrease was due to the human perceiving shared understanding and thus subsequently deemphasizing previously salient words. The salience detection heuristic was not designed to cope with this and thus the subsequent classification performance decreased. Finally, the robot was able to attach sensorimotor meaning to the presented object and this was further highlighted in the robot’s ability to distinguish which feature was important in identifying the presented shape from the set of attributes in the sensorimotor stream.

6 CONCLUSION

This study has attempted to demonstrate how interaction between a human and a robot can allow the robot to attach meaning to lexical items used by the tutor referring to a series of shapes. We have shown how the robot can derive meanings using some simple heuristics, developed with reference to human child language acquisition. Importantly no restrictions are placed on the human interactor other than to describe the shapes *as if* the robot were a young child. What is important is that the human adopts a language style similar to that employed in child directed speech and that some form of rudimentary shared reference is employed. This being the case the robot will learn the semantics of the shape words, based on its own sensorimotor feedback, with relatively few presentations.

Clearly, the case study discussed above is limited in that the number of participants is statistically relatively small (8 in total). We are therefore cautious in suggesting that the results are in any way conclusive. However we do believe that the results and the computational mechanisms outlined indicate interesting directions and novel approaches for future studies.

ACKNOWLEDGEMENTS

The work described in this paper was conducted within the EU Integrated Project ITalk (“Integration and Transfer of Action and Language in Robots”) funded by the European Commission under contract number FP7-214668.

REFERENCES

- [1] ARToolkit. <http://www.hitl.washington.edu/artoolkit>, 2003. [last visited on 30 June 2008].
- [2] R N Aslin, J Z Woodward, N P LaMendola, and T G Bever, ‘Models of word segmentation in fluent maternal speech to infants’, in *Signal to Syntax*, eds., J Morgan and K Demuth, Lawrence Erlbaum, (1996).
- [3] Paul Bloom, *How Children Learn the Meaning of Words*, MIT Press, 2002.
- [4] Noam Chomsky, *Aspects of the theory of syntax*, MIT Press, Cambridge, MA, 1965.
- [5] Eve V. Clark, *First Language Acquisition*, Cambridge University Press, Cambridge, UK, 2nd edn., 2009.
- [6] Walter Daelemans and Antal van den Bosch, *Memory-Based Language Processing*, Cambridge University Press, 2005.
- [7] Kerstin Dautenhahn, Chrystopher L. Nehaniv, Michael L. Walters, Ben Robins, Hatice Kose-Bagci, N. Assif Mirza, and Michael Blow, ‘Kaspar - a minimally expressive humanoid robot for human-robot interaction research’, *Applied Bionics and Biomechanics, Special Issue on ‘Humanoid Robots’*, **6**(3), 369–397, (2009).
- [8] eSpeak. <http://espeak.sourceforge.net/>, 2007. [last visited 31 July 2009].
- [9] S. Harnad, ‘The symbol grounding problem’, *Physica D.*, **42**, 335–346, (1990).
- [10] Carolyn B. Mervis and Laurel M. Long, ‘Words refer to whole objects: Young children’s interpretation of the referent of a novel word’, in *Paper Presented at Biennial meeting of the Society of Research in Child Development*, Baltimore, MD, (1987).
- [11] Microsoft. Microsoft speech technologies. <http://www.microsoft.com/speech/speech2007/default.aspx>, 2007. [last visited 30 June 2009].
- [12] Chrystopher L. Nehaniv, ‘Meaning for observers and agents’, in *IEEE International Symposium on Intelligent Control/Intelligent Systems and Semiotics (ISIC/ISAS’99)*, pp. 435–440, (1999).
- [13] Chrystopher L. Nehaniv, Kerstin Dautenhahn, Jens Kubacki, Martin Haegele, Christopher Parlit, and Rachid Alami, ‘A methodological approach relating the classification of gesture to identification of human intent in the context of human-robot interaction’, in *14th IEEE International Workshop on Robot and Human Interactive Communication (Ro-Man 2005)*, pp. 371–377, (2005).
- [14] openCV. <http://opencvlibrary.sourceforge.net>, 2006. [last visited 30 June 2008].
- [15] N. Otero, C. L. Nehaniv, D. S. Syrdal, and K. Dautenhahn, ‘Naturally occurring gestures in a human-robot interaction teaching scenario’, *Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems*, **9**(3), 519–550, (2008).
- [16] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA, 1993.
- [17] Deb Roy, ‘Grounding words in perception and action: Computational insights’, *Trends in Cognitive Sciences*, **9**(8), 389–396, (Aug 2005).
- [18] Deb Roy and Alex Pentland, ‘Learning words from sights and sounds: A computational model’, *Cognitive Science*, **26**, 113–146, (2002).
- [19] Jacqueline Sachs, Barbara Bard, and Marie L. Johnson, ‘Language learning with restricted input: Case studies of two hearing children of deaf parents’, *Applied Psycholinguistics*, **1**, 34–54, (1981).
- [20] J. Saunders, C. L. Nehaniv, K. Dautenhahn, and A. Alissandrakis, ‘Self-imitation and environmental scaffolding for robot teaching’, *International Journal of Advanced Robotic Systems*, **4**(1), 109–124, (March 2007). special issue supplement on Human-Robot Interaction.
- [21] Luc Steels, ‘The origins of syntax in visually grounded robotic agents’, *Artificial Intelligence*, **103**(1-2), 133–156, (1998).
- [22] SysMedia. Sysmedia word and phoneme alignment software. [Last visited 31 July 2009]. <http://www.sysmedia.com>.
- [23] TimBL. <http://ilk.uvt.nl/mbpl/>, 2005. [last visited 30 June 2009].
- [24] Michael Tomasello, *Constructing a Language: A Usage-Based Theory of Language Acquisition*, Harvard University Press, 2003.
- [25] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Elsevier, Morgan Kaufmann, San Francisco, 2005.
- [26] Ludwig Wittgenstein, *Philosophical Investigations (Philosophische Untersuchungen)* – German with English translation by G.E.M. Anscombe, Basil Blackwell, 3rd edn., 1968. (first published 1953).
- [27] C. Yu and D. Ballard, ‘A multimodal learning interface for grounding spoken language in sensorimotor experience’, *ACM Transactions Applied Perception*, **1**, 57–80, (2004).

Ethical issues in robot care for the elderly: Dystopia or optimism?

Amanda Sharkey¹ and Noel Sharkey¹

Abstract. The increasing proportion of elderly people in society, together with recent advances in robotics, makes the use of robots in elder care increasingly likely. Robots in elder care can be used (i) to assist the elderly and their carers; (ii) to monitor the health and safety of old people, and (iii) to provide them with companionship. Although robots could be usefully employed in each of these areas, there is a need to ensure that they actually benefit the old people themselves, and not just their carers. We discuss a number of ethical concerns and argue that, if introduced with foresight and careful guidelines, robots and robotic technology could improve the lives of the elderly. They could enable more independence, and create more opportunities for social interaction

1 INTRODUCTION

Could robots improve the lives of old people? It is often suggested that one way of coping with the increasing proportion of elderly people in our society would be to use robots and related technology to help in their care. As a result, there is a growing industry in developing robots for elder care. However, some ethical concerns are beginning to be expressed about these developments (e.g. [1,2,3,4]). There is concern for instance that using robots for elder care could result in increased social isolation, and could involve deception and loss of dignity. The three main ways in which robots might be used in elderly care are: (i) to assist the elderly, and/or their carers in daily tasks; (ii) to help monitor their behaviour and health; and (iii) to provide companionship. In an effort to determine the likely effects of robots on the lives of old people, we outline recent developments under each of these headings. We identify and discuss both the benefits that might be obtained from deploying robots in eldercare, and the ethical concerns that have been raised about their use. Finally we consider the relative likelihood that robot use in elder care will lead to a dystopia of neglect, or a more optimistic, empowering scenario; concluding with suggestions about ways to ensure that the latter is the case.

Underlying our consideration of the ethical issues involved in using robots in elder care, is the assumption that the human

rights of the elderly deserve as much protection as those of their carers and family. Amongst the fundamental human rights established in documents such as the Charter of the United Nations, and the Universal Declaration of Human Rights, the rights that are usually thought of as particularly relevant to the elderly include rights to a standard of living adequate for health and well-being; to private and family life; to freedom from torture and inhuman or degrading treatment; and to freedom from discrimination. If the human rights of the elderly are to be respected as much as the rights of other members of society, it is important to ensure robots that introduced into elder care do actually benefit the elderly themselves, and are not just designed to reduce the care burden on the rest of society. Hence our opening question, “could robots improve the lives of old people?” focuses on the effect of robots on the elderly themselves, rather than on the utility and convenience of using robots to help care for an aging population.

2 ASSISTIVE ROBOTS AND ELDER CARE

A number of robots have been developed to assist with some of the daily tasks involved in eldercare. Such robots include the Japanese Secom “My Spoon” automatic feeding robot, and the Sanyo electric bathtub robot that automatically washes and rinses. Riken have developed the Riba robot (Robot for Interactive Body Assistance) that has the face of a teddybear: it can pick up and carry humans from a bed to a wheelchair. It can recognise faces and voices and respond to spoken commands. In the US, Charlie Kemp at Georgia Tech, has developed an assistive robot, EI-E that can respond to many of the commands, and perform the same tasks in a house that a guide dog is capable of. When coloured towels are tied to doors, and drawers, the robot can be instructed to open them, using a combination of verbal commands, and a laser pointer to indicate the target location. It can pick up quite a wider variety of objects from flat surfaces [5].

There are some interesting new developments arising from robotic technology that could improve the lives of the elderly. Cyberdene has developed the Hybrid Assistive Limb suit (HAL) that is currently available to rent in Japan. It detects nerve signals sent by a person attempting to move, and then is said to automatically move the muscle as the person expects. It can apparently multiply original strength by a factor of 2 to 10. There are also walking machines from Honda: ‘Stride Management Assist’, and their ‘Walking Assist’ device with body weight support system (New Scientist online, November 2008), although these are some way away from mass

¹ Dept. of Computer Science, University of Sheffield, S1 4DP

. Email: {a.sharkey, n.sharkey}@dcs.shef.ac.uk

production. Wheelchairs for the elderly and the disabled are becoming increasingly sophisticated. For example, at MIT, Nicolas Roy and colleagues are working on an interface to a wheelchair that will enable it to take the user to a designated known location in an indoor environment, on vocal command [6].

3 ROBOTS THAT MONITOR AND SUPERVISE THE ELDERLY

A number of robots have been developed for monitoring the health and safety of old people. Pearl, developed at CMU [7], is a “nursebot” that reminds seniors about routine activities (e.g. eating, drinking, taking medicine and using the bathroom/toilet), and that can guide them through their environment. RP-7 is a tele-operated robot that is used to facilitate doctor-patient interactions without the doctor leaving her office. It has been used at the Silverado Senior Living Aspen Park [8]. The uBot5 robot, developed by University of Massachusetts, Amherst [9] has manipulators that enable it to perform a number of tasks such as picking up and moving objects around. It has been proposed as a means of doing house calls to check for signs of a stroke and could be used to monitor an old person for signs of a fall. It can also be used for social telepresence, since it can be remotely controlled by authorised users over the internet, allowing a virtual visit and two way conversation, with the visitor’s face appearing on a video screen. The robot can also be remote controlled to perform tasks in the home. Mitsubishi Heavy Industries developed the Wakamura robot mainly as a companion (see next section), but it can also be used for monitoring the elderly, delivering messages and reminding about medicine.

Gecko Systems International Corp has predicted that its sales of eldercare personal robots will reach 8.3 Billion dollars by 2014. They are developing the CareBot™, a personal robot equipped with multiple vital sign sensors, that can follow an elderly person in their home: home evaluation trials with the elderly began in November 2009. The CareBot is capable of verbal interaction, the delivery of medicine, video monitoring and two way interaction, and can provide verbal reminders, at predetermined dates and times, of the need for medical checks, or of other appointments. In Osaka, Japan, Matsushita Electric runs “Sincere Kourien”, an elderly retirement home with 106 beds, and used it as a test bed for robot teddy bears that monitor patients’ response times to spoken questions (Guardian online, 21st February 2002 <http://news.bbc.co.uk/1/hi/sci/tech/1829021.stm>). The bears record how long the residents spend performing various tasks, and can relay conclusions to staff, or alert them to unexpected changes. They form part of the monitoring network that exists throughout the building, and for instance alerts staff when residents leave their beds. Their use in the Sincere Kourien home illustrates a likely direction for monitoring robots, as a robotic component is integrated into a network of fixed sensors more typical of “smart” homes.

4 ROBOTS AS COMPANIONS FOR THE ELDERLY

It is sometimes suggested that robots could provide companionship for the elderly. Indeed, an old person might be more likely to tolerate, or even enjoy, the presence of a monitoring robot if it also had some ability to be a companion. For instance, the Gecko CareBot referred to in the previous section is described as “a new kind of companion that always stays close to them enabling friends and family to care from afar”. There is also a growing number of robot toys, some of which have been mooted as possible companions for the elderly. Such toys have the advantage of being relatively more affordable than some of the more elaborate robots we have considered so far.

Potential robot companions include Paro, a fur covered robotic seal, which was specifically designed for therapeutic uses with the elderly. Developed by AIST, it responds to petting by moving its tail and opening and closing its eyes. It reacts to sounds and can learn to respond to its name. It makes seal-like sounds, and is active in the day, preferring to “sleep” at night. It can detect light and dark by means of a light sensor, and recognise when it is being held, stroked, or hit, by means of posture and tactile sensors. Sony’s AIBO robotic dog, developed as an “entertainment robot”, has also been used in robot companions research. It has a metallic dog-like form, and can walk, or chase a ball. It has sensors that can detect distance, acceleration, sound, vibration and pressure. It can express six “emotions” (happiness, anger, fear, sadness, surprise and dislike), by means of its tail, body movements, and the colour and shape of its eyes. More recent versions can recognise voice commands, and the robot exhibits slightly different behaviour depending on the interactions it has experienced.

NeCoRo (OMRON) is a robotic cat covered in synthetic fur. Its behaviour depends on the history of its interactions and it can ‘learn’ to recognise its name. Stroking and hugging it elicits positive responses while treating it roughly elicits angry behaviour. Like Paro, it has internal rhythms that lead it to sleep, or seek to be cuddled. And My Real Baby (iRobot) is described as an “interactive emotionally responsive doll”. It can make different expressions, blink, and suck its thumb or bottle. It changes its expressed emotion depending on how it is treated – so that when fed and rocked for instance, it behaves as though it were sleepy. When unfed, it gets hungry and cries. It also makes a number of random sounds and words, which become more like sentences the longer it is played with.

There are other robots that have also been touted as possible companions for the elderly (see [10] for interesting discussions of digital companions in general). Toy robots that could entertain the elderly include: Pleo, Ifbot, and Primo Puel. Pleo is a robotic dinosaur with many sensors that responds with different behaviours depending on its treatment. The IfBot was developed by Business Design Laboratory Co. for elderly people and can converse with them by means of a large number of stored interaction patterns. Primo Puel is an interactive doll that talks, giggles, and asks

for cuddles. It was originally designed to stand in for a boyfriend for young single women, but proved unexpectedly popular with elderly women in Japan.

5 POTENTIAL BENEFITS

Assistive robots such as the Riba that can lift patients from bed to wheelchair, and back, could relieve those who look after the elderly from a physically demanding task. Robots that provide ways of remotely monitoring old people could potentially increase the likelihood that they could remain in their own home for longer. Studies that have looked at the effect on the elderly of interacting with robot companions have found evidence that old people can benefit from them. The positive effects are often said to be similar to those obtained from animal assisted therapy in which visits from a pet animal to residents of long term care facilities are found to reduce loneliness [11][12]. For instance, Kanamori *et al* [13] showed various improvements in elderly persons who interacted regularly with a Sony AIBO robotic dog – their loneliness scores were reduced, and their Quality of Life assessment scores improved. Banks *et al* [14], even found that old people in long term care facilities benefited as much from interacting with an AIBO robotic dog as from interacting with a real dog; both resulting in reductions in their reported levels of loneliness. Elderly dementia patients have also shown positive outcomes, including increased communication, as a result of sessions with an AIBO [15]. An interesting possibility that could be explored in future research is that some of the benefits that result from playing with, or “nurturing” robot pets are a consequence of increased levels of oxytocin.

Quite often the suggested benefits of robot eldercare are framed more from the perspective of the carer, than from that of the old person in receipt of such care. For example, GeckoSystems suggest that their robot will help the caregiver for an elderly person by giving them some time off, and that it will “decrease[s] the difficulty and stress for the caregiver that needs to watch over Grandma, Mom, or other family members most, if not much, of the time day in and day out due to concerns about their well being, safety, and security.” (from an article on GeckoSystems website, December 2009). To be fair, GeckoSystems do also consider the benefits of robot care for the care receiver, suggesting that such robots will make it possible for the elderly to stay in their own, or their family’s home for longer and avoid long term nursing care; but more emphasis is placed on the advantages for the caregiver.

6 ETHICAL CONCERNS

The concerns that have been raised with respect to the use of robots in eldercare have mostly focused on two issues: loss of human contact, and deceit. A decrease in social contact for the elderly could result if robots were used for tasks such as feeding them. The use of robots to monitor the health of the elderly could remove an opportunity for social interaction with a nurse, or doctor. Even the use of robots to clean floors

could remove the informal opportunity to chat to a human cleaner, as Sparrow and Sparrow [1] point out. Robot visitors, and robot companions could also be used as an excuse by society and the family of old people to avoid the need of actually calling in and checking on their well being – “don’t worry about Granny, she’s with the robot” could become a familiar response.

The argument has also been made that the introduction of robot companions involves deceit of the old person, who may be encouraged to think that a robot is something with which they could have a relationship. Sparrow [16] and Sparrow and Sparrow [1] argue that any beneficial effects of robot pets or companions are a consequence of deceiving the old person into thinking that the robot pet is something with which they could have a relationship, and are adamant that this should not be encouraged. Turkle *et al* [17] expressed similar disquiet: “the fact that our parents, grandparents and our children might say ‘I love you’ to a robot who will say ‘I love you’ in return, does not feel completely comfortable; it raises questions about the kind of authenticity we require of our technology” [17]. Wallach and Allen [4], in a discussion of the ability of robots to detect basic human social gestures, and respond with human-like social cues, suggest that, “From a puritanical perspective, all such techniques are arguably forms of deception” [4].

Sparrow argued that the relationships of seniors with robot pets, “are predicated on mistaking, at a conscious or unconscious level, the robot for a real animal. For an individual to benefit significantly from ownership of a robot pet they must systematically delude themselves regarding the real nature of their relation with the animal. It requires sentimentality of a morally deplorable sort. Indulging in such sentimentality violates a (weak) duty that we have to ourselves to apprehend the world accurately. The design and manufacture of these robots is unethical in so far as it presupposes or encourages this”, [16].

The issue of deceit however is complex. Although Sparrow and Sparrow [1] argued that robot companionship necessarily involves deceit, we suggest that this is not as simple an issue as they suggest. It is not straightforward because people in general are all too ready to anthropomorphise about machines, as discussed by Sharkey and Sharkey [18]. Zizek [19] the psychoanalyst, describes the way in which people can chose to act as though something were not real, “I know very well that this is just an inanimate object, but none the less I act as if I believe that this is a living being”. There may well be elements of a “willing suspension of disbelief” [20] when old people interact with robot pets. Indeed, there are likely to be considerable individual differences here in terms of old people’s beliefs about, and understanding of robotic technology. There is as yet little knowledge about the interactions between old people’s knowledge of, and attitudes to robots, and the possible benefits they might obtain from interacting with them. It is quite possible that old people might enjoy, and benefit, from interacting with a robot pet without thinking that it is actually sentient.

As well as concerns about the (possible) deception of old people, we can identify other potential downsides to robots in elder care. It is quite possible that, if used insensitively, they

could increase the feeling of powerlessness and objectification that can result from the “malignant care” that old people sometimes receive [21]. Being lifted and carried by a robot that is under the control of another person seems likely to make an old person feel more of a burden, and less in control of their life. Further ethical concerns are discussed by Sharkey and Sharkey [2]; the use of robots to monitor old people could infringe their privacy, especially if it was done without the consent, or even without the knowledge of a confused old person. Robots left to supervise an old person might well be required to prevent or restrict some of their activities; discouraging them from leaving the building, or from potentially dangerous activities. The issues involved in using robots in an authoritarian capacity to restrain human activities are complicated and multifarious, and would depend amongst other things on a belief that the robot was able to consistently recognise the intended action of the old person, and its likely consequences.

7 DYSTOPIA OR OPTIMISM?

It seems there is a real risk that robots in elder care could lead to a dystopian future in which old people were removed from society and left in the care of robots. There is also a need for caution in accepting the results of studies that seem to show that old people benefit from interacting with robot companions, since such benefits depend on what the alternative is. The results reported by Kanamori et al [13] showed improvements in well being over time between initial, and later sessions. Banks et al [14] showed that beneficial effects were obtained for those interacting with either the real or the robotic dog, as compared to their control group who received no such opportunities for interaction. However, such improvements could have been found because the alternative was so dire and lacking in stimulation. Someone in solitary confinement might benefit from being given a robot companion, but they would benefit far more if they were released into a friendly social environment. It is possible that the similar improvements might be found if the comparison were to a control group that received some other form of intervention. If the comparison group received some other individual attention, such as a visit by someone who massaged their neck and shoulders, or who sat holding their hand and chatting, the benefits of the robot dog might not be so apparent. It is also important to see that any benefits are maintained over time, because a robot that was interesting to interact with initially may rapidly lose its appeal. A Reuters article (21st September 2007) reported that the old people in a care home had liked the IfBot robot for about a month before losing interest. The director of the facility in Japan commented, “stuffed animals are more popular”.

Nonetheless, we suggest that it is possible that robots could be introduced into elder care in such a way as to improve the lives of the elderly themselves, and not just the lives of their carers. There is a need for further, carefully controlled research on their impact on the elderly themselves and their quality of life. There is a need to provide protection against the overuse of robots in the care of the elderly, and for some

guidelines, or preferably legislation, to ensure that robots do not result in a reduction in opportunities for human contact for the elderly. The problem is that, apart from fundamental human rights legislation, there is little protection for old people against the potential downsides of robot care. In particular, there are no obvious explicit restrictions on the amount of time that old people could be left in the care of robots, nor on the amount of human contact that they should experience. Like children, the very old and infirm can be seen as being in need of special protection. As discussed elsewhere [20], children could possibly be afforded some protection from extensive robot care by the legislation pertaining to neglect, but there is no such legislation for the elderly.

If such protection were in place, it is quite possible that robots and robotic technology could be used in such a way as to increase the degree of social interaction that old people experience. For example, the wearable suits discussed earlier could make it possible for the elderly to get out of their houses, or chairs, and to reach social meeting places. An intelligent wheelchair that responded to the commands of the user could have a similar effect. Assistive aids based on robotic technology could function more as tools, or even as extensions of the old person's body, as discussed by Clark and Chalmers [22].

As well as helping old people to get to social meeting places, such aids could also increase the feeling of autonomy of old people, and reduce their dependence on human carers. Unfortunately, human carers do not always have the time and inclination to respond to the needs of old people, as the following example from the Joint Committee report on ‘Human rights and elder care’[23] shows: the account is that of an elderly wife visiting her husband in hospital: “*I went to visit my husband on the first day and he is a very private person, he doesn't like anything to embarrass him and when I went in he was almost in tears which is not my husband. He said 'please, please go and get a bottle I am nearly wetting myself'. I rushed out I got a bottle and I said to him 'Well why didn't you just ring the nurse', in my innocence. 'I have, for an hour and a half I've been asking for a bottle'. Well when I went out [and] told the nurse she said 'Oh don't worry we would have changed the sheets'. Now his dignity at that stage would have gone out of the window. There was no dignity.*” (Pg 15). A robotic device such as a wheelchair that acted under the control of the old person themselves could substantially reduce such dependence.

As indicated above, a possible downside of robot companions is that they could be used as an excuse for leaving an old person on their own for longer. Yet it is also possible that a robot companion could create or promote social interaction with fellow humans. Robot companions can act as social facilitators, giving the people around them something to talk to an old person about, and giving the old person themselves something to show, and to talk about. A robot that facilitates conversation can be an attractor for visitors. Children may want to play with the robot and have fun with granddad's big toy. For instance, Kanamori et al [13] report the case of an 84-year-old man with cerebral apoplexy sequelae. He talked much more to his children after

the introduction of an AIBO robot dog. It gave both him and them a focussed object to talk about. Similarly, when Wada and Shibata [24] videoed interactions between a Paro robot seal and a group of elderly care home residents, they found evidence that the social interactions between the residents themselves increased, at the same time that physiological indicators showed reduced stress levels. It seems that Paro even encouraged positive communication, and resulted in a reduction of the “backbiting” that had previously characterised some of their interactions. One 75 year old female resident greatly increased her interaction with fellow residents. A friend of hers commented that she had been taciturn before Paro’s arrival, but that now she was more cheerful and talkative. Thus benefits could be obtained from robot pets that are due to the robots stimulating increased interaction with other people; benefits that have nothing to do with deceiving the old person.

8 CONCLUSIONS

We have reviewed the main uses of robots in eldercare, and considered both the claimed benefits, and the ethical issues that are involved. There are concerns that robots could be used to create a dystopian future, in which neglected old people are insensitively manipulated and controlled by machines. At the same time, it is possible that robots could be introduced in such a way that they resulted in increased social interaction between the elderly and their society, and that offered the elderly more control over their lives and the environment.

How can we ensure that the optimistic, rather than the dystopian future, is achieved? Some of the solutions proposed above depend on sufficient monies being provided - However, the argument is beginning to be made that the cost of providing residential nursing care is so high, that the provision of robotic technology that enables old people to continue to live in their own homes begins to seem more cost effective. Research that focuses on the views and experiences of the elderly who interact with robots will help, although as discussed there is a need to guard against the possibility that any benefits shown are not just because interacting with a robot is better than being left in complete, or near, isolation. In particular, we are more likely to achieve the optimistic rather than the dystopian future if guidelines and legislation are introduced to give the elderly rights to certain levels of human social contact, and to protect them from being left in the control of robots. It is not the robots themselves that are intrinsically good or bad for the elderly, but the way in which they are used.

9 REFERENCES

- [1] R. Sparrow and L. Sparrow (2006) In the hands of machines? The future of aged care. *Mind and Machine*, 16: 141-161.
- [2] N. Sharkey and A. Sharkey (in press) Living with robots: ethical tradeoffs in eldercare, In Y. Wilks, *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues*. Amsterdam: John Benjamins
- [3] N.E. Sharkey (2008) The Ethical Frontiers of Robotics, *Science*, 322: 1800-1801
- [4] W. Wallach, and C. Allen (2009) *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, New York.
- [5] H. Nguyen, C. Anderson, A. Trevor, A. Jain, Z. Xu, and C. C. Kemp, (2008) *El-E: An Assistive Robot that Fetches objects from Flat Surfaces, HRI Workshop on Robotic Helpers: User Interaction Interfaces and Companions in Assistive and Therapy Robots*.
- [6] F. Doshi, and N. Roy (2008) Spoken language interaction with model uncertainty: an adaptive human-robot interaction system. *Connection Science*, 20, 4, 299-318.
- [7] M.E. Pollack, S. Engberg, J.T. Matthews, S. Thrun, L. Brown, D. Colbry, C. Orosz, B. Peintner, S. Ramakrishnan, J. Dunbar-Jacob, C. McCarthy, M. Montemerlo, J. Pineau, and N. (2002) Pearl: A Mobile Robotic Assistant for the Elderly, *AAAI Workshop on Automation as Eldercare*, Aug 2002.
- [8] S. Winner, (2007) Beam me inside, Scotty! *Assisted Living Consult*
- [9] P. Deegan, R. Grupen, A. Hanson, E. Horrell, S. Ou, E. Riseman, S. Sen, B. Thibodeau, A. Williams, D. Xie. (2007) Mobile Manipulators for Assisted Living in Residential Settings, *Autonomous Robots, Special Issue on Socially Assistive Robotics*. Volume 24, Number 2.
- [10] Y. Wilks, (in press) Editor, *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues*. Amsterdam, John Benjamins
- [11] M.R. Banks and W.A. Banks, (2002) The effects of animal-assisted therapy on loneliness in an elderly population in long-term care facilities. *J Gerontol A Biol Sci Med Sci* 57A:M428–M432.
- [12] M.R. Banks and W.A. Banks (2005) The effects of group and individual animal assisted therapy on loneliness in residents of long-term care facilities. *Anthrozoos*;18:396–408.
- [13] M. Kanamori, M. Suzuki, & M. Tanaka, (2002) Maintenance and improvement of quality of life among elderly patients using a pet-type robot. *Japanese Journal of Geriatrics*, v. 39, p. 214-8.
- [14] M.R. Banks, L.M. Willoughby, and W.A. Banks (2008) Animal-assisted therapy and loneliness in nursing homes: use of robotic versus living dogs, *J Am Med Dir Assoc* 9, pp. 173–177.
- [15] T. Tamura, S. Yonemitsu, A. Itoh, D. Oikawa, A. Kawakami, Y. Higashi, T. Fujimoto, L. Nakajima. (2004) Is an entertainment robot useful in the care of elderly people with severe dementia? *J Gerontol A Biol Sci Med Sci*. Jan;59(1):83-5.
- [16] R. Sparrow, (2002). The march of the robot dogs. *Ethics and Information Technology*, 4, 305–318.
- [17] S. Turkle, W. Taggart, C. D. Kidd, O. Dasté, (2006) Relational Artifacts with Children and Elders: The Complexities of Cybercompanionship. *Connection Science*, 18, 4, pp 347-362.
- [18] N. Sharkey and A. Sharkey (2006) Artificial Intelligence and Natural Magic, *Artificial Intelligence Review*, 25, 9-19
- [19] S. Zizek, (2002) *The Zizek Reader*. London:Blackwell
- [20] N. Sharkey and A. Sharkey (to appear) The crying shame of robot nannies: an ethical appraisal. *Interaction Studies*.
- [21] T. Kitwood, (1997) *Dementia Reconsidered: The Person Comes First*. Open University Press.
- [22] A. Clark and D.J. Chalmers (1998) The extended mind. *Analysis* 58:10-23.
- [23] *Older people and human rights: Research and Mapping Report*. 2009-11-03 Age Concern England and British Institute of Human Rights.
- [24] K. Wada and T. Shibata (2006) Robot Therapy in a Care House: Its Sociopsychological and Physiological effects on the Residents. *Proceedings of the 2006 International Conference on Robotics and Automation*, Orlando, Florida, May 2006. 3966-3971.

Mirror my emotions!

Combining facial expression analysis and synthesis on a robot

Stefan Sosnowski¹ and Christoph Mayer² and Kolja Kühnlenz³ and Bernd Radig⁴

Abstract. Everyday human communication relies on a large number of different communication mechanisms like spoken language, facial expressions, body pose or gestures. Facial expressions are one of the main communication mechanisms and pass large amounts of information between human dialogue partners [22]. Therefore, the analysis and the synthesis of facial expressions are important steps towards an intuitive human-machine interaction and form valuable research targets. We present a system that tackles both challenges. It relies on a fully automated, model-based, real-time capable approach to distinguish universal facial expressions and their intensities from camera images. Facial expression synthesis is conducted via the robot head EDDIE, a flexible low-cost emotion-display with 23 degrees of freedom. Static facial expressions at continuous intensities are included, as well as smooth transitions based on the circumplex model of affect. Miniature off-the-shelf mechatronic components are used to provide high functionality at low cost. Evaluations conducted in a user-study show that emotions displayed by EDDIE are recognizable by humans very well. By combining facial expression recognition and display on the robot, a first demonstration is presented in which the robot mirrors the human's emotions, as a basis for further research in the field of emotional closed loop systems.

1 Introduction

Robots will be an integrated part of future living or working environments making fast and efficient human-robot interaction important. Traditional human-human interaction passes important information on many different communication channels like the auditory channel for spoken language or the visual channel for facial expressions. In contrast, traditional human-machine interaction via mouse and keyboard is often considered slow and non-intuitive. Today, dedicated hardware often derives the emotional state from blood pressure, perspiration, brain waves, heart rate, skin temperature, etc [5, 14, 13]. Humans, however, interpret emotion via video and audio information

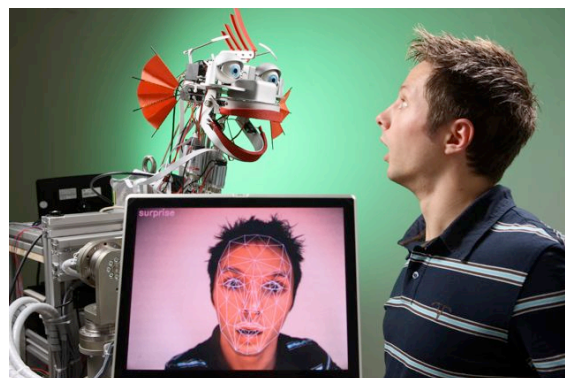


Figure 1. Demonstration setup with EDDIE (foto: Kurt Fuchs).

that is acquired without interfering with other persons. Fortunately, this information can be acquired by computers as well using general purpose hardware. For a comprehensive overview on applications in emotion recognition, we refer to Lisetti et al. [10]. Yet, recent research aims at enabling machines to utilize communication channels more natural to human beings, such as gestures or facial expressions. Knowledge about human behavior, intention, and emotion is necessary as well to construct convenient interaction mechanisms.

We present a robot system with combined facial expression recognition and synthesis, that aims at providing communication via facial actions between the robot and the human and therefore partially filling the gap between human-human and human-machine interaction. Facial expressions form an important aspect of human-human communication and pass important information about the interaction partners' emotional states, which robots should be able to utilize to understand and eventually influence the human interaction partner. The proposed system recognizes facial expressions from camera images, processes the information encoded in the facial expression and reacts with facial expressions of the robotic head EDDIE. In a preliminary version, the facial expression recognized from the human is merely mirrored on the robotic head, see Figure 1. However, since facial expressions are strong communication signals in human-human interaction, more complex reactions according to psychological insights will be integrated in the future.

¹ Institute of Automatic Control Engineering, Department of Electrical Engineering and Information Technologies, Technische Universität München, Munich, Germany email: sosnowski@tum.de

² Chair for Image Understanding and Knowledge-Based Systems, Computer Science Department, Technische Universität München, Munich, Germany email: mayerc@in.tum.de

³ Institute for Advanced Study (IAS) and Institute of Automatic Control Engineering, Department of Electrical Engineering and Information Technologies, Technische Universität München, Munich, Germany email: koku@tum.de

⁴ Chair for Image Understanding and Knowledge-Based Systems, Computer Science Department, Technische Universität München, Munich, Germany email: radig@in.tum.de

1.1 Related Work

Integrating human-inspired forms of communication in robotic systems will allow for a human-machine interaction with increased efficiency, security and robustness and ease of use [16]. The work presented in this paper presents a fusion of two research areas: facial expression recognition and facial expression synthesis. In this paper, we present a combination of two systems that have already been inspected in the past independent of each other. Please note, that the main contribution of this work is not to present a novel way of facial expression analysis or synthesis but rather to introduce a system that combines both approaches in a closed loop. Both are based on the Facial Action Coding System (FACS), which precisely describes the muscle activity within a human face that appear during the display of facial expressions [3]. Ekman and Friesen find six universal facial expressions that are expressed and interpreted independent of the cultural background, age or country of origin all over the world, see Figure 2 [2].

1.1.1 Facial Expression Recognition

Pantic et al. [12] propose a three-step approach for facial expression recognition which we also follow in our work. Firstly, the position and shape of the face visible in the image is determined. Secondly, features descriptive for facial expressions are extracted. Finally, in the third step a classifier is applied to identify the expression. Some approaches infer the facial expression by exploiting a face model. Kotsia et al. take this approach by fitting a face model to the example images showing either a neutral face or a strongly displayed facial expression and extracting the model parameters [8] to apply rules stated by Ekman and Friesen [3]. Similar to our approach, Gokturk et al. rely on person-specific information of a neutral face and track a three-dimensional model in image sequences [4]. In contrast, they utilize non-public image data for building the face model, training the classifier and evaluating their approach. De la Torre et al. [9] extract temporal segmentation of facial behavior from image data to assist professional FACS coders. They demonstrate their approach on image data that has not been taken with a computer vision application in mind and therefore face challenging context conditions. However, they rely on person-specific active appearance models (AAM) which renders the approach not applicable to previously unseen data. In contrast, we utilize publicly available data throughout



Figure 2. The six universal facial expressions as they occur in the Cohn-Kanade Facial Expression Database.



Figure 3. The face model is fitted to each image in order to estimate the facial expression currently visible.

the whole approach. The accuracy of model fitting is of great importance to model-based approaches since the accuracy of facial expression recognition depends directly on the model's capability to reduce the image data and reflect its content. Pure image data without prior data reduction is also applied for facial expression classification. Littlewort et al take this approach [27]. They utilize AdaBoost and Support Vector Machines to determine the facial expression from convolutions of image data with Gabor energy filters and their approach obtains very high recognition results. Local binary patterns are applied in recent work by Shan et al. [28]. These approaches suffer from large changes in face appearance which is usually not provided by standard databases. Similar, Michel et al extract the motion of 22 feature points within the face between a neutral image and an image that displays a strong facial expression to train a Support Vector Machine (SVM). [11].

1.1.2 Facial Expression Synthesis

Facial expression synthesis is utilized in virtual agents and robots for social interaction. In virtual systems there are two major methods: One is an abstraction based approach, using blend shapes, form interpolation or deformation based morphing. The other is muscle based, being separable in physical mass-spring systems and muscle group based approaches like FACS [26]. In robotic systems, mostly FACS is used (e.g. Kismet [16], Saya [25]).

2 Facial Expression Recognition

We rely on model-based techniques to determine the exact location of facial components such as eyes or eye brows in the image. Geometric model form an abstraction of real-world objects and contain knowledge about their properties, such as position, shape or texture in a model parameter $\mathbf{p}$, see Figure 3. The Candide-3 face model consists of 116 anatomical landmarks and has been specifically designed for facial expression recognition [8]. Its parameter vector $\mathbf{p} = (r_x, r_y, r_z, s, t_x, t_y, \sigma, \alpha)^T$ describes the affine transformation $(r_x, r_y, r_z, s, t_x, t_y)$ and the deformation (σ, α) [1]. The 79 deformation parameters indicate the shape of facial components such as the mouth width or the roundness of the eyes and partially refer to the Facial Action Coding System, see Figure 4. Model-fitting is the computational challenge of determining correct model parameterizations for single images without prior knowledge of the image content. We rely on the work of Wimmer et al. to tackle this challenge [15].

2.1 Feature Extraction

Two aspects generally characterize facial expressions: they turn the face into a distinctive state and the involved muscles show a distinctive motion. Our approach considers both aspects by extracting static and person-adapted features and assembles them in a feature vector $\mathbf{x}$. Since we consider the transformation parameters not related to

the facial expression, they are not considered in the feature assembly. The static features are assembled by calculating the model parameters for a single image and extracting the deformation parameters σ, α . To calculate the person-adapted features, we acquire a neutral reference image of the person visible in the image sequence and determine the model parameter change $\Delta\sigma, \Delta\alpha$ from the reference image to the inspected image.

2.2 Classification

The aim of this step is to infer the facial expression currently visible from the extracted features. For training purpose, the Cohn-Kanade Facial Expression Database [6] and MMI Face Database [7] serve as training and testing data. They provide the facial expression label for every image sequence. To determine the expression intensity, we manually annotate the facial expression starting, peak and ending image index in every sequence and model the facial expression intensity to increase or decrease linearly. The intensity is defined to range between 0 (neutral image) and 1 (peak expression image). We apply decision trees and Support Vector Machines (SMVs) to map the feature vector x on the facial expression visible in a single image. Since the databases do not provide the point of transition between a neutral face and a displayed facial expression within a single sequence, we decide to use all images with an intensity less than 0.33 as neutral face representatives. Once the facial expression is determined, a model tree is trained to infer its intensity from x . Model trees are flow-chart-like tree structures, in which internal nodes denote a test on a feature to branch the flow of the computation into one of the two subordinate nodes [20]. Furthermore, the leaf nodes contain line segments. Advantages of Model Trees over other means of classification are the low run-time and fast generation of the tree from a previously acquired set of training data.

2.3 Experimental Evaluation

Every sequence in the image databases starts with a neutral face and develops into the apex expression. In contrast to the image sequences of the Cohn-Kanade Facial Expression Database, the MMI Face Database image sequences also show decreasing facial expression intensity until a neutral face is reached again. We fit the face model to each image and compute the static and person-adapted features for each image individually.

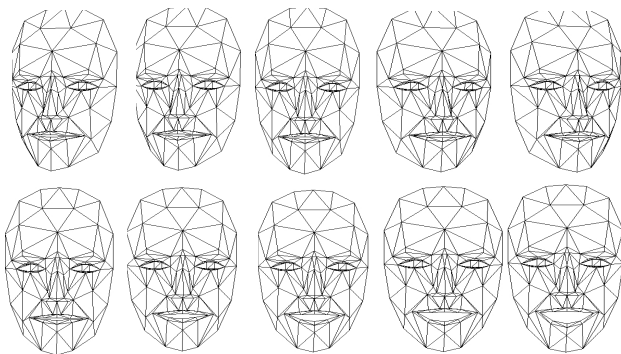


Figure 4. The face model reflects face properties such as gaze direction or facial component shapes.

We provide the classification algorithm with different sets of features. The classifier C1 is trained on the static features only and therefore it is applicable to single images, without requiring a comparative image depicting a neutral facial configuration. The classifier C2 is presented the person-specific features only and adapts to the person visible in the image. The classifier C3 includes both static and person-specific features, with the intention of combining the "best of both worlds". The classifiers are evaluated by a cross-validation of ten independent subsets of the training data, where nine sets are used to train a classifier and the remaining set is used for testing purposes. The intensity estimation is evaluated by correlation between the value calculated by the model tree and the image groundtruth, see Table 1.

	C1	C2	C3
decision tree	83.6% / 91.8%	84.7% / 91.8%	86.1% / 92.0%
SVM	87.9% / 93.1%	88.4% / 92.7%	91.4% / 87.8%

database	Cohn-Kanade	MMI
intensity correlation	92.8%	93.7%

Table 1. Recognition accuracies for different feature sets and classifiers. The numbers are given for Cohn-Kanade Facial Expression Database and MMI Face Database, respectively.

3 Facial Expression Synthesis

According to Mori's theory of the uncanny valley, a robotic interaction partner is considered to be visually most unpleasant, if it is designed to look very similar to a human, but not entirely human. Therefore, in order to design a comfortable robotic interaction partner, a reasonable degree of familiarity should be achieved, balancing well between a machine-like and a human-like appearance which guided the design process of EDDIE [18].

A distinctive feature of this robot head is that it includes not only human-like facial features, but also animal-like features. The ears can be tilted as well as folded/unfolded and a crown with four feathers is included. For further details on the actual design of the head, please refer to [21]. EDDIE has a total of 23 degrees of freedom. To acquire video data from a human interaction partner, IEEE1394 Point-Grey Dragonfly cameras are integrated in the eyes, covered with natural looking eyeballs. The mounting in the eyes ensures full frontal images of the interaction partner, with the partner normally fixating the gaze on the eyes of the robot. To compensate for movements, the eyes are movable in two degrees of freedom (pan/tilt). Video data is streamed with a resolution of 640x480 at 30 frames per second. The head can be controlled via an embedded control architecture over either a TCP/IP connection or an interface to the RealTime DataBase (RTDB) developed by Matthias Göbl [24]. The RealTime DataBase is a shared memory architecture, implementing a publisher and subscriber concept. In this demonstration setup the robot is controlled over RTDB, which is also used to pass the video-stream captured from EDDIE's eye cameras to the expression recognition software component.

3.1 Expression Generation

The emotional state of the display can be controlled in two ways, either referring to the discrete basic emotions found by Ekman et al.

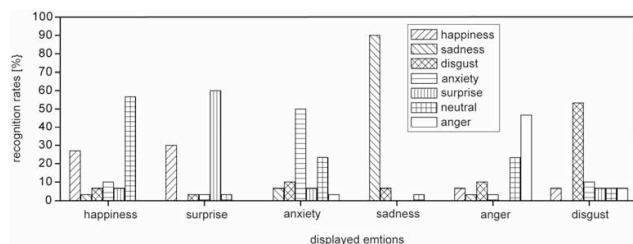


Figure 5. Results of the evaluation study

or referring to the circumplex model of affect proposed by Russel et al.[2, 17]. Each state of the discrete emotions was modeled according to the FACS description, transitions between states are animated by linear interpolation of the respective motor commands of the start - / end-state. The expression of the six emotional states can be seen in Figure 6. A second way to generate the emotional state of the robot and the according facial expressions is the circumplex model of affect. In this model, the state-space is spanned by a valence (pleasure) and an arousal (excitement) axis. Every mood-describing expression can be defined according to its pleasure- and excitement-component. With a state-space to joint-space mapping that we developed, arbitrary facial expression in this emotional space can be generated believable and with smooth transitions [21].

3.2 Experimental evaluation

A study with 30 participants (researchers of Technische Universität and the University of the Armed Forces, München, 11 female, 19 male, age-mean 30 years) has been conducted. Six basic emotions have been presented to each individual. The subjects' ratings of the displayed emotions were rated with a multiple-choice test with a seven-item-scale. The result of this study can be seen in Figure 5. On the abscissa the six displayed emotions are shown. For each displayed emotion the amount of assumed emotions by the subjects is presented. For example, 90% of the participants agree in seeing a sad face if a sad face is displayed. For the displayed emotions anger and anxiety about 50% of the answers were correct, the remaining 50% are shared between the other emotions. This evaluation study proves that the generated facial expressions can be correctly identified by untrained users and matched to the corresponding emotional state of the robot to a reasonable degree [23].

4 Demonstration setup

In this setup we are using the RealTime DataBase as an interface between the recognition and the synthesis components, providing a fast and reliable data exchange. Images captured from the robot head's eye cameras are passed to the recognition module via the RealTime DataBase. The face model is fitted to faces found in the images. The facial expression and it's intensity are extracted from the face model and passed to back to the RTDB. This information about the emotional class and the intensity is then pushed to the synthesis component on the robotic head. Although all six basic facial expressions are recognized and mirrored on both databases, a preliminary demonstrator version includes only the strongest facial expressions, laughing and surprise from live images. The display of the facial expression is computed as presented in Section 3.1. In order to account for the

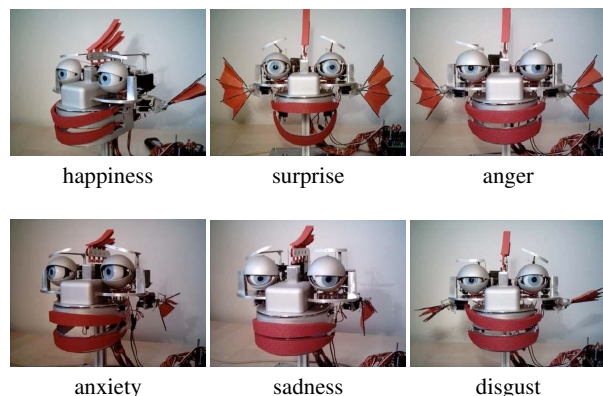


Figure 6. EDDIE displaying the basic facial expressions proposed by Ekman et al. [2].

intensity, the shown facial expression results from a linear interpolation between the emotional class and the neutral expression. The human interaction partner perceives the emotional state displayed by the facial expression, closing the loop, as can be seen in Figure 1 and 7.

5 Conclusion and Future Work

In this paper an approach is presented that recognizes six universal facial expressions together with their intensity and mirrors them on a robotic head, as a first step towards an emotional closed loop system. We integrate the Candide-3 face model and exploit its parameter vector to extract feature values. Accuracy on independent training data peaks at 91%. Our approach runs in real-time on standard hardware. The drawback of our approach is the inability to recognize very subtle changes due to the lack of texture information. Furthermore, the recognition accuracy is limited by the training data provided causing the facial expression recognition to fail if presented face movement not part of the training data.

Future work includes evaluation on and improvement of our approach on data obtained from more realistic real-world scenarios and extension to the live mirroring of all six basic expressions. So far the system components have been evaluated independently of each other. Although the recognition rates of facial expressions obtained by the system and the inspection of facial expression recognizability by humans indicate that the synthesized expression can be recognized with the one displayed by the human, there is no scientific proof, yet. Therefore, future work will include evaluating the complete analysis-synthesis loop on human test subjects. A test will be set up that asks human probands to match a synthesized facial expression with the original human face.

Real-time capabilities are important for future use in emotional closed loop systems, where delays in information processing should be kept to a minimum. While the mirroring in future experiments could be extended to more sophisticated versions, the final goal is to influence the human interaction partner in it's perceived emotional state with the induction of feelings through the emotional display. This might improve the acceptance or performance rating of the robot in human-machine-interaction.

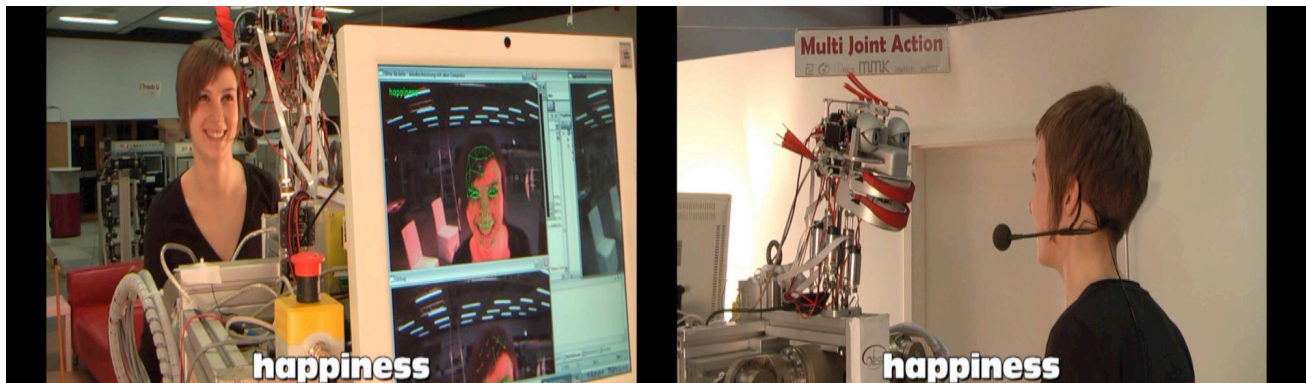


Figure 7. Facial expression recognition and synthesis demonstration

ACKNOWLEDGEMENTS

This work is supported in part within the DFG excellence initiative research cluster Cognition for Technical Systems CoTeSys, see www.cotesys.org, and by the Institute for Advanced Study (IAS), Technische Universität München, see also www.tum-ias.de.

REFERENCES

- [1] J. Ahlberg. Candide-3 – an updated parameterized face. Technical Report LiTH-ISY-R-2326, Linköping University, Sweden, 2001.
- [2] P. Ekman. Universals and cultural differences in facial expressions of emotion. *Nebraska Symposium on Motivation 1971*, University of Nebraska Press.
- [3] P. Ekman. Facial expressions. *Handbook of Cognition and Emotion*, New York, 1999. John Wiley & Sons Ltd.
- [4] S. B. Gokturk, J.-Y. Bouguet, C. Tomasi, and B. Girod. Model-based face tracking for view-independent facial expression recognition. *IEEE International Conference on Automatic Face and Gesture Recognition*, 2002.
- [5] C. S. Ikehara, D. N. Chin, and M. E. Crosby. A model for integrating an adaptive information filter utilizing biosensor data to assess cognitive load. In *User Modeling*, Springer Berlin / Heidelberg, 2003.
- [6] T. Kanade, J. F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *International Conference on Automatic Face and Gesture Recognition*, 2000.
- [7] M. Pantic and M. F. Valstar and R. Rademaker and L. Maat. Web-based Database for Facial Expression Analysis. Proc. IEEE Int'l Conf. Multimedia and Expo (ICME'05), 2005.
- [8] I. Kotsia and I. Pita. Facial expression recognition in image sequences using geometric deformation features and support vector machines. *Trans. On Image Processing*, 16(1), 2007.
- [9] F. D. la Torre Frade, J. Campoy, Z. Ambadar, and J. Cohn. Temporal Segmentation of Facial Behavior. In *International Conference on Computer Vision*, 2007.
- [10] C. L. Lisetti and D. J. Schiano. Automatic facial expression interpretation: Where human interaction, artificial intelligence and cognitive science intersect. *Pragmatics and Cognition, Special Issue on Facial Information Processing and Multidisciplinary Perspective*, 1999.
- [11] P. Michel and R. E. Kaliouby. Real time facial expression recognition in video using support vector machines. In *Fifth International Conference on Multimodal Interfaces*, 2003.
- [12] M. Pantic and L. J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE PAMI*, 2000.
- [13] E. M. Sheldon. *Virtual agent interactions*. PhD thesis, 2001. Major Professor-Linda Malone.
- [14] R. M. Vick and C. S. Ikehara. Methodological issues of real time data acquisition from multiple sources of physiological data. *Annual Hawaii International Conference on System Sciences*, 2003.
- [15] M. Wimmer, F. Stulp, S. Pietzsch, and B. Radig. Learning local objective functions for robust face model fitting. *IEEE PAMI*, 2008.
- [16] C.L.Breazeal. Designing sociable robots. *IEEE The MIT Press* 2002
- [17] J.A.Russell A circumplex model of effect *Journal of Personality and Social Psychology* 1980
- [18] J.Reichard Robots: Fact, Fiction and Prediction *Penguin Books* 1978
- [19] A.Mehrabian and J.A.Russell An Approach to Environmental Psychology *MIT Press* 1974
- [20] J. R. Quinlan Learning with continuous classes. In *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence*, pages 343-348, 1992.
- [21] S.Sosnowski and A.Bittermann and K.Kühnlenz and Martin Buss Design and Evaluation of Emotion-Display EDDIE In *Int. Conference on Intelligent Robots and Systems*, 2006
- [22] A. Mehrabian Communication without Words, *Psychology Today*, volume 2, pages 53-56, 1968
- [23] K. Kühnlenz and S. Sosnowski and M. Buss Evaluating Emotion Expressing Robots in Affective Space In *Advanced Robotic Systems and I-Tech*, Vienna, Austria, 2007
- [24] M. Goebl and G. Färber, "A real-time-capable hard- and software architecture for joint image and knowledge processing in cognitive automobiles," *Intelligent Vehicles Symposium*, pp. 737 – 740, June 2007.
- [25] T. Hashimoto and S. Hitramatsu and T. Tsujia and H. Kobayashi Development of the Face Robot SAYA for Rich Facial Expressions SICE-ICASE 2006. International Joint Conference
- [26] Z. Deng and J. Noh Data-Driven 3D Facial Animation Computer Facial Animation: A Survey. Springer London, 2007.
- [27] G. Littlewort and M. Stewart Bartlett and I. Fasel and J. Susskind and J. Movellan Dynamics of facial expression extracted automatically from video *Image and Vision Computing*, vol. 24, pp. 615–625, 2006
- [28] C. Shan and S. Gong and P. W. McOwan Facial Expression Recognition based on Local Binary Patterns: A comprehensive Study *Image and Vision Computing*, vol. 27, pp. 803-816, 2009

On Making Robots Invisible-in-Use

Leila Takayama¹

Abstract. A major challenge facing human-robot interaction is understanding how to people will interact and cope with increasingly agentic objects in their everyday lives. As more robotic technologies enter human environments, it is critical to consider other models of human-robot interaction that do not always require focused attention from people. Ubiquitous computing put forth the perspective that computers should not always be the focus of our attention, but that computing should weave itself into the fabric of our everyday lives. Similarly, robots might be the center of attention in some interactions, but might be even more effective when they fade into one's attentional background. In this line of thought, the current study presents results from interviews ($N=19$) and surveys ($N=46$) regarding personal experiences with tools that became invisible-in-use, shedding light upon ways that robots might do the same. We present the lessons learned from these open-ended interviews and surveys in the context of larger theories of making tools invisible-in-use [10], functional [18], ready-at-hand [9], proximal [16], and/or in the periphery of one's experience [27].

1 INTRODUCTION

As robots become increasingly pervasive in human environments and our everyday lives [5], it is critical to consider how people will come to deal with increasing amounts of autonomy in the tools and agents with which they will interact. One fundamental challenge to human-robot interaction is how to design interactions with these increasingly autonomous agents without overloading people's limits of attention, cognitive load, and patience. This is similar to the challenge posed in human-computer interaction (HCI) with the notion of ubiquitous computing.

1.1 The Ubiquitous Computing Perspective

Ubiquitous computing (ubiquomp) is often misunderstood as simply being about computers being everywhere. While it may be true that computers are becoming increasingly pervasive in our everyday lives, that is merely the premise, not the thesis, of ubiquitous computing. The most frequently cited description of ubiquitous computing is Mark Weisers *Scientific American* article: "The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it" [24]. His ideal for computers was that they recede into the background of life, just as written language, paper, and electricity have disappeared from our conscious attention despite the fact that we use them all the time. This vision of computing was once deemed the Third Paradigm of computing [26, 27], following after its predecessors of mainframe computing and personal computing (PC).

The goal of making something "disappear" in the ubicomp sense is to quietly empower people. As Weiser put it, "only when things disappear in this way are we freed to use them without thinking and so to focus beyond them to new goals" [24]. This notion extends the definition of computers to include "things that think" [6, 8], i.e., appliances and other objects that have computational abilities, e.g., robots.

This perspective of ubiquitous computing extends into human-robot interaction issues of making robots so natural, easy, and invisible-in-use, that they fade into the background of everyday experiences. This contrasts against a model of interacting with robots that demand large amounts of conscious attention directed at the robot in order for the person to get anything done.

1.2 Invisible-in-Use

One formulation of this ubiquitous computing perspective is to make tools (e.g., computers and robots) invisible-in-use. Being invisible-in-use means that something is phenomenologically invisible: "the experience of direct interaction with artifacts and tools largely free of conscious monitoring" [10], i.e., it is experienced from a first-person perspective as though it is not present when it is used.

The ideas that tools can become invisible-in-use is not new. It has been put forth in different forms by prominent philosophers, psychologists, and graphic novelists alike. Heidegger [9] wrote about the carpenter's hammer became ready-at-hand (i.e., invisible-in-use) when he became accustomed to using it, but that the carpenter could easily shift to reflecting upon the hammer as being present-at-hand (e.g., noticing its shape, materials, weight) if desired. See Figure 1.

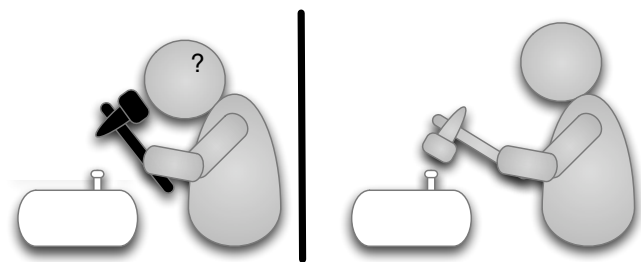


Figure 1. Example of a hammer being present-at-hand (left) and ready-at-hand (right)

Heidegger's concepts of present-at-hand vs. ready-at-hand have influenced much of HCI research, primarily through the focus upon the phenomenological experience of using computers as articulated by Winograd and Flores [28].

Another philosopher who heavily influenced the ideas behind ubicomp was Michael Polanyi, who wrote about the tacit dimension of

¹ Willow Garage, USA, email: takayama@willowgarage.com

human experience [17]; in his conceptualization, there are proximal and distal elements of human attention: “In an act of tacit knowing we attend from something for attending to something else; namely from the first term to the second term of the tacit relation. In many ways the first term of this relation will prove to be nearer to us, the second further away from us” (p. 10). He uses the example of someone learning to use a probe, shifting from experiencing the impact of the probe against his hand (as a distal object, i.e., present-at-hand) to automatically sensing the meaning of feeling impacts from the tip of the probe (as a proximal object, i.e., ready-at-hand).

Phenomenological philosopher Merleau-Ponty [14] expressed this same phenomenon in terms of a person’s sense of “I can,” which is shaped by the tools that the person has mastered; when a person has mastered a tool, she or he has a “maximum grasp” [4] of it (that is goal and context specific) and will perceive and interact with the world from a different perspective. As an example, if you have mastered a hammer, then you see nails-like things in terms of what they afford, according to your tools and skills. Merleau-Ponty provided the example of a blind man using a walking stick and the opportunities for action afforded to a blind man with a maximum grasp of this tool: “the stick is no longer an object perceived by the blind man, but an instrument with which he perceives. It is a bodily auxiliary, an extension of the bodily synthesis.”

Ecological psychologist, James Gibson [18] conceptualized an entitive vs. functional point of view that also corresponds to present-at-hand vs. ready-at-hand perspectives. He presents the notion of a field of possible actions such as those afforded by the expert use of a car: “Within the boundaries of the road lies, according to our hypothesis, an indefinitely bounded field which we will name the field of safe travel. It consists, at any given moment, of the field of possible paths which the car may take unimpeded. Phenomenally it is a sort of tongue protruding forward along the road” (p. 120). Faster sports cars might have longer tongues than slower ones and more readily maneuverable cars might have larger, wider tongues than cars that are difficult to handle. Dant further delved into this particular example as a driver-car experience [3]. For both Merleau-Ponty and Gibson, the main idea was that “animal behavior is best understood in terms of alertness to opportunities for action” [19].

Psychologist Mihaly Csikszentmihaly [2] similarly wrote about the state of “flow,” which is characterized as a state of complete absorption or concentration upon an activity (e.g., being in “the zone”). He provides examples of athletes, musicians, programmers, etc. reaching the state of flow when being totally immersed in their given activity in a way that feels positively energized, rather than overwhelming or negative.

Graphic novelist Scott McCloud [13] provides many other vivid examples of people fusing with their cars, telephones, forks and knives, and hats. He provides a starting set of previously lacking empirical examples from this literature and idea.

These related concepts each its own nuances, but all broadly speak to the same first-person experience of a tool becoming invisible-in-use. It is this phenomenon that we explored in the current study, eliciting personal experiences and stories to gain insights into our understanding of what it would mean for tools to become invisible-in-use and how robots and other tools might become invisible-in-use, too.

1.3 Robots as Invisible-in-Use

As explicated in prior work [21], robots could become invisible-in-use when they fade into the background of one’s experience in-the-moment even though one objectively knows that the robot is not a

part of oneself upon further reflection; this is the ultimate immersive telepresence experience, as depicted in Figure 2.

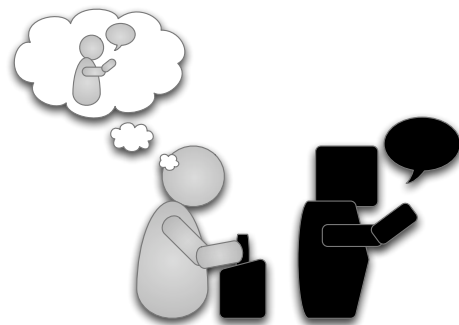


Figure 2. Robots can become invisible-in-use as a medium through which a teleoperator is acting

Similarly, more autonomous robots can provide services to help people in ways that are unobtrusive and do not rely upon explicit commands from the user. As an example, a person working on task (e.g., trying to hammer a nail or stay awake long enough to complete a homework assignment) might be helped by robots in ways depicted in Figure 3.

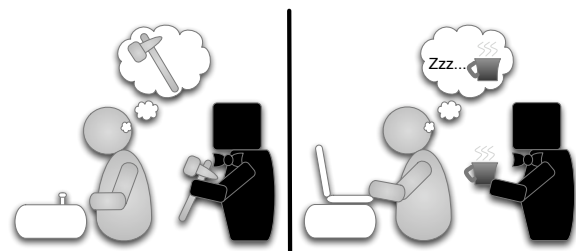


Figure 3. Robots can become invisible-in-use by unobtrusively providing services to people to help them to achieve the person’s primary tasks

Although the spirit of ubiquitous computing ran against the idea of using interface agents [25], Weiser got stuck on the point that one type of invisible agent that might be butlers. This may not necessarily correspond to the real world service that butlers provide, but it indicates a lean toward providing services to people in unobtrusive way (e.g., ubiquitous robots [11] and butler-like agents [20]).

2 STUDY DESIGN

The current study aimed to gain insights into how tools become invisible-in-use, allowing people to use their own definitions of tools. We did not prompt them to think about robots, computers, or any particular technology. We merely provided the example of the carpenter’s hammer to give respondents a sense of what we meant by “invisible-in-use.”

2.1 Respondents and Interviewees

Respondents for the survey ($N=46$) and participants for the interviews ($N=19$) were recruited through campus announcements. All participants were over 18 years of age, granted informed consent to

participate in the study, and were paid in course study participation credit. This study was approved by the research institution's internal review board (IRB) to ensure that it respected the rights of the participants involved.

2.2 Hypotheses

Based upon the idea that breakdowns are key to understanding how tools become present-at-hand [28], we hypothesized that (H1) breakdowns would make invisible-in-use tools visible again. Because we associated the ideals of ubiquitous computing with the ease of use, we also hypothesized that (H2) tools that are easy to use and comfortable would also make them invisible-in-use.

Although we had hypotheses about what would make tools invisible-in-use and visible again, we only used these hypotheses to pose open questions as a starting point for anticipating responses, but we did not drive questions in these particular directions.

3 METHODS

3.1 Survey

The survey was administered online with this set of open-ended questions: Please take a moment to think about a tool that you have used that became invisible-in-use.

- Please describe the tool that became invisible-in-use to you.
- For what activities do (or did) you use the tool?
- How did you learn to use the tool? How did it become invisible-in-use?
- Why did the tool become so invisible? Was there something about the properties of the tool itself, the context in which you learned to use it, or were some other factors involved?
- Approximately how long did it take for the tool to become invisible-in-use for you?
- Please write about a time (or times) when a tool broke down or otherwise required you attention.
- In general, what makes a tool visible (i.e., brought to the center of your attention)?
- Are there any other thoughts you would like to share about the tool or about this concept of invisible tools?

Upon completion of the survey, respondents read a debrief of the study and were given contact information in case they had any questions or concerns about the study. Thus, the survey only covered one tool per respondent.

3.2 Interviews

The interviews were conducted in person and consisted of the following set of open-ended questions:

- Some people believe that tools can become a part of one's self when using them. Please tell me about a tool that became a part of you at some point in time.
- Please tell me about a tool that you have completely mastered.
- Please tell me about a tool that you feel completely comfortable using.
- Please tell me about a tool that you have sometimes felt is actually a part of yourself.
- Please tell me about a tool that became so familiar to you that it became invisible-in-use.

For each tool mentioned, we asked:

- How did you learn to use the tool?
- How did it become so familiar?
- Why do you think the tool became so invisible?
- How long did it take to become invisible-in-use?

Finally, we asked:

- Please tell me about a time when a tool became visible again.
- Please tell me about a time when an invisible tool broke down.
- What do you see as the difference between invisible and visible tools?
- What is your definition of a tool?

Interview responses were audio recorded (when permitted by interviewees) and transcribed. Upon completion of the interview, participants were debriefed about the purpose of the study and were given the opportunity to discuss questions or concerns. Thus, the interviews elicited stories about four to five tools per respondent.

In collecting data for this study, we were cautious to refrain from judging people's responses and made every effort possible to make respondents feel comfortable and confident in their answers. We were careful to ask questions in terms of their own first-person perspectives [22] and emphasized that they could use their own definitions of tools and invisible-in-use tools, tell us about tools that they felt reflected the concept of the invisible-in-use tools, and discuss personal experiences with particular tools rather than tools in general.

4 DATA ANALYSIS

In line with the phenomenological perspective presented earlier, we chose to base our data collection methods informed primarily by empirical phenomenological research [7]. Our goal in these open-ended interviews was to gain an understanding of what it is like for each person when they relate stories of their personal experiences with tools that have become invisible-in-use to them. The main phases of empirical phenomenological research are: (1) data constitution, (2) transforming raw data into phenomenological descriptions, (3) psychological reflection on each example to yield structure of each example, and (4) identifying general psychological structure from examples (p. 161).

In transforming the raw data into phenomenological descriptions, we analyzed each invisible tool example down to the essence of what made the tool invisible-in-use to the respondent, how it became invisible-in-use, and what (if anything) it was about the tool itself that the respondent believed made it invisible-in-use. In searching for structures for each example, we reviewed the responses and their distilled forms to identify themes and structural patterns that cut across invisible tool examples. Finally, we present general psychological structures from the examples in the following section.

5 RESULTS AND DISCUSSION

5.1 Type of Invisible-in-Use Tools

Even though the category of tool has its prototypical members (e.g., hammer, wrench, anything found in a tool box), people largely did not discuss these types of tools. Out of the 132 invisible tools reported by respondents and interviewees, 33 of them were unique. Items mentioned only once included: ballet shoes, braces, chopsticks, credit card, ice skates, light switches, pool cue, saxophone, staplers, swimming goggles, tennis racket, tweezers, watch, and yoga mat.

Those items that were reported by more than one respondent are presented in Figure 4.

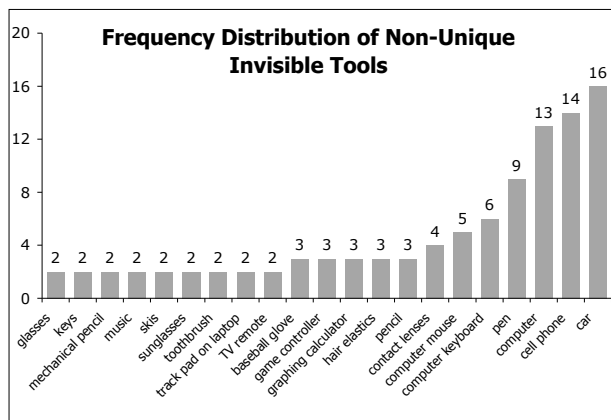


Figure 4. Frequency distribution of invisible tools reported (items mentioned more than once)

As can be seen in Figure 4, cars, cell phones, computers, and pens were the most frequently mentioned items. Out of the 132 invisible tools reported, we found that 51 of them involved a computer in some way (e.g., computer, computer keyboard, computer mouse, graphing calculator, cell phone), 55 of them were purely mechanical (e.g., baseball glove, keys, mechanical pencil, pen, skis, toothbrush), and 38 of them involved items worn on one's body (e.g., cell phone, contact lenses, hair elastic ties, hat, shoes, sunglasses, wallet, watch).

5.2 What Makes Tools Invisible-in-Use

Several themes emerged from the descriptions of their psychological experiences of having tools become invisible-in-use. The strongest themes included:

1. **Reliability:** Being able to rely upon the performance of a tool is absolutely key to tools becoming invisible-in-use. A major theme that cut across many stories was that the tool can be taken for granted. When invisible tools disappointed users, it was often because of unreliable behavior, which typically results in disrupting their everyday routines (e.g., computers failing, cars breaking down, contact lens drying out). Exemplar statements include: "If you're wearing high heels and you have to run, you don't feel confident. Even if you're not wobbly... the heel isn't really supporting you so you have to run on your toes. With figure skates, you have to trust that it's fully under you because you're going to jump in the air and land on it. You can't look down or worry that it's going to buckle under your foot," and, "as soon as it breaks, my whole routine totally falls apart and... everything becomes prioritized behind replacing the cell phone, making the cell phone work again, charging the cell phone, going to the... store and yelling."
2. **Predictability and consistency:** Being able to predict what the tool would do was also very important for almost all invisible-in-use tools. When expectations were not met or people were surprised by the behavior of the tool, that made the tool visible (i.e.,

present-at-hand). Exemplar statements include: "The [computer] keys always followed the same layout, even across platforms and different keyboard versions," and, "a tool will become invisible if it is consistent and my mind is able to grasp in patterns necessary for its input, output, and desired use." Once the user knows what to expect from using the tool and there are no surprises, the tool can become invisible-in-use.

3. **Familiarity:** Most examples of these invisible tools were objects that the person had owned for many years and that the person used very often (e.g., hair elastics, cell phones, hats). Some tools were only invisible-in-use if it was the person's own tool (e.g., 4-year old cell phone), whereas others could quickly adapt to using many instances of the tool (e.g., pen). Exemplar statements include: "over time, the positions of the keys became more familiar... With more practice, my fingers eventually remembered the exact locations of the keys," and, "[your own car is] a familiar environment and everything is just the same. You just feel like you're at home. I don't have that feeling when I get into other cars."
4. **Sense of control:** Having a sense of control over the tool as opposed to feeling out of control with the tool (e.g., driving a car) was also critical to a tool becoming invisible-in-use. Exemplar statements include: "I honestly miss the ease and complete control with which I maneuvered my old [cell] phone which I had for about three years."

Contrary to our original hypothesis (H2), we learned that several aspects we expected to be important for tools to become invisible-in-use were not actually required.

1. **Comfort:** Although comfort was often mentioned as one aspect of people's invisible tools, it was not always necessary. Ballet shoes and figure skates are not comfortable and take a very long time to break in, but comfort helps to speed up the time until one reaches a point of the tool becoming invisible-in-use (e.g., estimated a few days for new running shoes to become invisible-in-use).
2. **Ease of use and simplicity:** The ease of using a tool was sometimes cited as the reason why a tool became invisible-in-use, but this aspect of tools was not necessary either. Graphing calculators, TV remotes, and computers are not necessarily easy to use (and often are not), but these persistent users were able to reach a point of feeling that such complex tools are invisible-in-use. As with comfort, easy to use tools seem to become invisible-in-use more quickly (e.g., stapling, swiping a credit card), but ease of use is not absolutely necessary for becoming invisible-in-use.
3. **Design:** Thoughtful product design was often cited as why some tools became invisible-in-use (e.g., Mac operating systems, video game controllers), but this was not absolutely necessary (e.g., poorly designed toothbrushes, unusual mechanical pencils). Even though some toothbrushes and mechanical pencils can be quite awkward to adapt to, even these types of tools were reported as becoming invisible-in-use to our respondents.

While these are certainly not the only aspects of these tools that made them invisible-in-use to our respondents, these were the primary themes that cut across their personal experiences with their invisible tools.

5.3 Qualia: What it is Like

Because a tool becoming invisible-in-use is a very personal, first-person experience [22], we aimed to gain a sense of the qualia (i.e., what it is like) for a person to experience a tool that is invisible-in-use. While it might be fundamentally impossible to fully understand

the qualia experienced by others [15], we did notice that respondents consistently used these phrases to describe their experiences with these tools:

- Don't need to think about it
- Effortless
- Extension of me
- Forget you're using it
- Instinctual
- Natural extension
- No longer think of them as new or technology
- Part of me
- Second nature
- Take for granted and don't miss until it's gone
- Use it without even noticing
- Very fluid motion

Using more specific examples, touch typing on computer keyboards and driving cars were experiences that several respondents described to us. They talked about both experiences in terms of the tool being an extension of one's sense of self: "The keyboard had become an extension of myself. There was no more thought required for the input of text." Similarly, in the driving example, one respondent said, "Usually, you are very aware of driving, but there are definitely times when I'm completely unaware and all I'm focused on is the road so it's like the car is an extension of me. I'm not conscious of how much I have to turn [the steering wheel] to turn [the car] a certain way. I'm not conscious of braking. It just happens naturally."

On the other hand, other respondents talked about the same tools as projections of their desired outcomes. Regarding the keyboard, one respondent reported, "When I'm typing notes, those are the thoughts that are going through my head that are just on there. It's kind of like a projection of what's in my head." Regarding driving, another respondent said, "I don't think about the steps of driving or what I'm doing when I'm driving. They become part of this natural process where I can do other things. Obviously, I'm in a car and it's not invisible in the sense physically but it's invisible in the sense mentally. I don't think about it anymore. It's a natural projection."

Some responses focused upon the fusion of the tool with one's self: "When you're skiing or you become decent at it, then there's a kind of synergy that happens between your legs, your boots, and your skis. It just becomes one system for getting around the mountain and then the tool becomes an extension of you, as you have a different organ that allows you (like how fish have gills that you can get around underwater) skis are sort of an organ that you can use... to get around mountains."

Altogether, these descriptions of what it is like to experience a tool as being invisible-in-use provide a grounding for evaluating how much a computational system such as a robot is experienced as being invisible-in-use, too.

5.4 How Tools Become Invisible-in-Use

While each tool had its own story about how it became invisible-in-use, a few themes emerged from reviewing the responses.

Active practice was the most often described way that people said their tools became invisible-in-use. Sometimes this practice was a part of formal training (e.g., driver's education or writing with pencils/pens) and sometimes not (e.g., using a cell phone or chopsticks). This process requires repetition over long periods of time, which is demanding of the user, but many respondents engaged in prolonged

practice because they saw using the tool as a necessity (e.g., commuting to school, communication needs while traveling, learning writing/literacy). Practiced routines (e.g., brushing teeth or hair) and social norms (e.g., wearing makeup, text messaging on cell phones) were other unanticipated reasons we learned about that motivated people to persevere and practice with their tools.

Passive exposure, observation, and use was another way that people's tools became invisible-in-use. Contact lenses were described by several respondents as being something they just got used to; even though they had to actively practice putting the contacts in and taking them out, wearing the lenses throughout the day was something that just took time to get used to. Similar stories were told about braces, including becoming accustomed to using headgear. Although many respondents were unable to recall how they learned to use more everyday tools (e.g., light switch, toothbrush, hair dryer), most speculated that they learned from watching their parents routinely use those tools.

Trial and error was a common response, as well. Tools such as the TV remote control, tweezers, computers, and cell phones were described as becoming invisible-in-use in this way. What respondents meant by this was that no one formally taught them how to use the tool and they did not read any manual or watch a tutorial about how to make use of the tool; they just tried using the tools themselves and worked out effective ways to use them.

What is most notable about these findings is that the most prevalent way that these tools became invisible-in-use to our respondents was simple practice through actual use, not book knowledge such as reading manuals. Several respondents described learning through watching other people using the tool and imitating them; this is consistent with learning theories of communities of practice [12]. By engaging in a process of legitimate peripheral participation, people learn how to master the tools such that the tools become invisible-in-use.

5.5 How Tools Become Visible

In analyzing the ways that people reported their tools becoming visible again, we were surprised to find that tools becoming visible again had to do with much more than just breakdowns (H1). Among the many ways that tools became visible to respondents were:

- Requiring conscious attention: A tool is new (e.g., new car) or breaks down (e.g., flat tire)
- Being annoying: A tool's presence and/or use is irritating (e.g., having to carry cell phone everywhere, feeling dried contact lenses in one's eyes)
- Being absent: This is consistent with the idea that invisible tools are taken for granted and yet indispensable (e.g., forgetting one's cell phone at home while traveling, thereby making it difficult to coordinate a pick-up from the airport)
- Forgetting how to use it: Losing one's skills in using the tools (e.g., forgetting how to play the saxophone after a long period of no practice)
- Considering alternatives: When alternative tools become available, one's existing tools can become more visible (e.g., waiting for a laptop to compile large amounts of code because it has less RAM than most computers on the market)
- Transferring knowledge to forms of the tool: When an instantiation of a tool is too different from one's own, the new one becomes quite visible (e.g., shifting from computer mouse to trackpad or from Windows to Mac OS)

- Anticipating urgency: When a tool absolutely must work reliably (e.g., waiting for mobile phone call from sick mother 113) or when a tool is known to fail catastrophically, even in predictable situations (e.g., driving across the country in a car that has known problems)

Contrary to our original hypothesis (H1), it is not simply the breakdown of a tool that makes it visible. Simply the threat of a tool breaking down can make a once invisible-in-use tool come to the forefront of one's conscious attention. One respondent described this situation while on a road trip across the country: "If I kept the car above 65 mph and kept the [gas] tank above half full, then it drove fine, but if I went above 70 [mph] or the [gas] tank went below a half, it wouldn't be able to get enough gas to the engine to keep going. It became very visible because I was very nervous. I didn't want the car to break down in the middle of nowhere so I'd keep looking at the dials, the meters, keep everything low. Once I did that, it would work, but of course it was no longer invisible." This is consistent with the earlier finding that a tool must be reliable to become invisible-in-use.

5.6 Implications for Theory

While there are many philosophical theories about the concept of tools become ready-at-hand, proximal, functional, and phenomenologically invisible-in-use, there is limited empirical data to ground this concept and to extrapolate and predict what tools more readily afford becoming invisible-in-use and how they do so. That is the gap that this study aims to begin to fill.

From this gathering of personal stories, we have gleaned at least a few lessons that implicate theories surrounding tools becoming invisible-in-use.

1. Types of invisible-in-use tools: While most examples provided by philosophers consist of purely mechanical tools (e.g., blind man's walking stick, carpenter's hammer), the notion of invisible-in-use tools extends to include computational tools (e.g., computers, cell phones) and increasing levels of autonomous behavior (e.g., automatic braking systems in cars).
2. Perspective upon the tool: A given tool can go in and out of one's awareness over very short periods of time, mostly depending upon when where one is attending. It is not an inherent property of the tool; in fact, the very same tool can be totally invisible-in-use to one person and not to another person.
3. Becoming invisible-in-use: Tools become invisible-in-use in ways that extend beyond only physical and psychological means; they also become invisible-in-use through social pressures and norms (e.g., cell phone adoption).
4. Becoming visible again: Tools that were once invisible-in-use can quickly become visible again through many means, not limited to breaking down.

This study presents a step toward explicating the concept of being invisible-in-use. By drawing from the first-person experiences of people who are not already steeped in concerns of making computers or robots invisible-in-use, we have drawn from the essences of their experiences to more thoroughly understand the experience of a tool becoming invisible-in-use.

5.7 Implications for HRI Design

Many of the lessons learned from this set of inquiries can be applied to the design of computational tools, including robots, to become invisible-in-use.

1. Learning how to use a tool so that it can become invisible-in-use can take a long time (on the scale of years, not just hours or days). That is not unreasonable if the tool is worth using and/or if there is enough motivation for the person to continue to learn to use the tool over long periods of time (e.g., driving cars, touch typing, skiing).
2. Because becoming invisible-in-use requires a certain psychological orientation toward a tool, it is important to encourage users to take this sort of stance toward the given tool. Highlighting too many details of the nuances of using a tool may hinder the user from being able to master the tool, whereas encouraging continuous practice with the tool may prove to be more effective.
3. While comfort, ease of use, and design can improve the user experience of a tool, none are necessarily required for a tool to become invisible-in-use.
4. It is not always good to be invisible-in-use. Speed dial features of cell phones and the mindless swiping of credit cards are not necessarily desirable because these tools can make the user rusty in his or her ability to use other phones to make phone calls and to not fully realize just how much money one is spending.
5. Any given tool can shift from moment to moment between being invisible-in-use and visible again. Therefore, it is important to design for both orientations toward robots, ideally enabling the user to effectively use and interact with the robot to achieve one's goals while remaining invisible-in-use (through practice, routine use, minimal attentional requirements, etc.) and to effectively deal with the robot when visible again (e.g., support the user in coping with breakdowns, annoyances, absences, memory lapses, etc.).

If it is desirable for a computer or robot to become invisible-in-use, then some guidelines to take away from the current study include:

- Reliability is critical to becoming invisible-in-use. Even the possibility of breaking down can ruin the experience of using the tool.
- Design robots to behave predictably and/or consistently across instantiations.
- Enable the user to feel completely in control of the tool. Taking steps toward providing the user with a sense of self-efficacy [1] can be used to achieve this goal.
- Do not design an interaction to require constant or frequent conscious attention.
- Encourage practice with the tool so that the user can become familiar with the tool through use.

While this is not an exhaustive list of lessons to be learned from tools that become invisible-in-use, it is a starting point from which to build and refine our understandings of what it is like for a tool to fade into the background of experience and how tools shift in and out of conscious experience.

6 FUTURE WORK

There is clearly room for further work in this research and design space, particularly in more specific cases of particular types of robots (e.g., humanoid robots), particular task domains (e.g., cleaning), and particular user groups (e.g., people from different demographic categories). In particular, a more direct comparison between these invisible-in-use tools and more agentic objects such as robots would flesh out the lessons that can be used to inform human-robot interaction design. The current study was purposefully designed to

be open-ended, not limited to the domain of robots, but more focused explorations into this domain could narrow the scope of inquiry to particular robots or types of robots. Drawing the connections between such invisible-in-use tools and acceptance/adoption of the technologies could be further explored through models such as the Unified Theory of Acceptance and Use of Technology [23], particularly in terms of effort and performance expectancies.

7 CONCLUSIONS

This review of the theoretical background of tools becoming invisible-in-use and empirical study of examples of such experiences provides the groundwork from which we can begin to understand how it is that people come to incorporate tools into their first-person skills and experiences. It also provides the HRI research and design community with guidance for how to enable robots to fade into the fabric of everyday life instead of being attention-demanding agents, taking a step toward exploring how to reach the goal of empowering people through robots that are more invisible-in-use.

8 ACKNOWLEDGEMENTS

Thanks go to Clifford Nass for his support, Stephanie Saenz for her assistance with interviews, the survey respondents and interview participants who made this project possible, and to the reviewers who provided insightful feedback to the earlier draft of this paper.

REFERENCES

- [1] Albert Bandura, *Self-efficacy: The exercise of control*, W. H. Freeman and Co, San Francisco, 1997.
- [2] Mihaly Csikszentmihalyi, *Flow: The psychology of optimal experience*, Harper Collins, New York, 1991.
- [3] Tim Dant, 'The driver-car', *Theory, Culture, and Society*, **21**(4/5), 61–79, (2004).
- [4] Hubert Dreyfus, *What computers can't do: A critique of artificial intelligence*, Harper, New York, 1972.
- [5] Bill Gates, 'A robot in every home', *Scientific American*, **296**(1), 58–65, (January 2007).
- [6] Neil Gershenfeld, *When things start to think*, Henry Holt and Company, New York, 1999.
- [7] Amedeo Giorgi, *Phenomenology and psychological research*, Duquesne University Press, Pittsburgh, Pennsylvania, 1985.
- [8] Rich Gold. The plenitude: Design and engineering in the era of ubiquitous computing.
- [9] Martin Heidegger. Basic writings: From being and time (1927) to the task of thinking (1964), 1992.
- [10] Shabeh Esearch Ivison, Jeffrey Heer, and Peter Khooshabeh, 'Seeing the invisible', in *In: Workshop on Invisible and Transparent Interfaces at AVI 2004, ITI Workshop, part of AVI 2004 (Advanced Visual Interfaces)*, (2004).
- [11] Jong-Hwan Kim, Yong-Duk Kim, and Kang-Hee Lee, 'The third generation of robotics: Ubiquitous robot', in *International Conference on Autonomous Robots and Agents*, (2004).
- [12] Jean Lave and Etienne Wenger, *Situated learning: Legitimate peripheral participation*, Learning in doing: Social, cognitive, and computational perspectives, Cambridge University Press, New York, 1991.
- [13] Scott McCloud, *Understanding Comics: The Invisible Art*, Harper Perennial, New York, 1993.
- [14] Maurice Merleau-Ponty, *Phenomenology of perception*, Routledge, New Jersey, 1962.
- [15] Thomas Nagel, 'What is it like to be a bat?', *The Philosophical Review*, **LXXXII**(4), 435–450, (1974).
- [16] Michael Polanyi, *Personal knowledge: Towards a post-critical philosophy*, University of Chicago Press, Chicago, 2nd edn., 1962.
- [17] Michael Polanyi, *The Tacit Dimension*, Doubleday, Garden City, NY, 1964.
- [18] *Reasons for realism: Selected essays of James J. Gibson*, eds., Edward Reed and Rebecca Jones, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1982.
- [19] John T. Sanders, 'Affordances: An ecological approach to first philosophy', in *Perspectives on embodiment: The intersections of nature and culture*, eds., Gail Weiss and Honi Fern Haber, 121–142, Routledge, New York, (1999).
- [20] Tim Sohn, Raphael Ballagas, and Leila Takayama, 'At your service: Using butlers as a model to overcome the mobile attention deficit', in *CHI '09: Proceedings of Human Factors in Computing Systems*, pp. 4219–4224. ACM Press, (2009).
- [21] Leila Takayama, 'Making sense of agentic objects and teleoperation: in-the-moment and reflective perspectives', in *HRI '09: Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pp. 239–240, New York, NY, USA, (2009). ACM.
- [22] Francisco J. Varela and Jonathan Shear, 'First person methodologies: What, why, how?', *Journal of Consciousness Studies*, **6**(2-3), 1–14, (1999).
- [23] Viswanath Venkatesh, Michael G. Morris, Gordon B. Davis, and Fred D. Davis, 'User acceptance of information technology: Toward a unified view', *MIS Quarterly*, **27**(3), 425–478, (2003).
- [24] Mark Weiser, 'The computer for the twenty-first century', *Scientific American*, 94–100, (1991).
- [25] Mark Weiser. Does ubiquitous computing need interface agents? no., October 1992 1992.
- [26] Mark Weiser and John Seely Brown, 'The coming age of calm technology', in *Beyond calculation: The next fifty years of computing*, eds., Peter J. Denning and Robert M. Metcalfe, Springer-Verlag, New York, (1997).
- [27] Mark Weiser and John Seely Brown, 'Center and periphery: Balancing the bias of digital technology', in *Blueprint for the Digital Economy*, ed., Don Tapscott, 317–335, McGraw-Hill, (1998).
- [28] Terry Winograd and Fernando Flores, *Understanding Computers and Cognition*, Ablex Publishing Corporation, Norwood, NJ, 1986.

Exploratory Analysis of Operator:Robot ratio in Search and Rescue Missions

Alberto Valero¹ and Chiara Saracini² and Paloma de la Puente³
and Diego Rodríguez-Losada⁴ and Fernando Matía⁵

Abstract. Mobile robots are increasingly becoming an aid to humans in accomplishing dangerous tasks. Examples of such tasks include search and rescue missions, military missions, surveillance, scheduled operations (such as checking the reactor of a nuclear plant), and so forth. The advantage of using robots in such situations is that they accomplish high-risk task without exposing humans to danger: robots go where humans fear to tread. When dealing with robot teams for exploration missions, the human operator needs to control simultaneously multiple robots. This paradigm highly increases the difficulty of the operator task. In this paper, based on extensive experimental data, we are studying the problematic associated to the Single-Operator-Multiple-Robot Interaction. This analysis can help to ascertain the critical points in multi-robot control and supervision. We propose two autonomy management modes and analyse the optimal operator robot:ratio for each of them. The experiments have been done in a Search and Rescue scenario using USARSim.

1 Introduction

An HRI system should permit humans and remote robots to cooperate in order to accomplish cognitively-demanding tasks, working in the spectrum between a fully autonomous system and a fully tele-operated one. The HRI system should then improve the accomplishment of the task by benefiting from the capabilities of the artificial agent and the human. The increment of the degree of autonomy, and the robot-coordination techniques are conducting researchers to deploy teams of robots for exploration missions. This brings a new challenge to the Human-Robot-Interaction system: *how the information coming from robots distributed spatially needs to be integrated into a Graphical User Interface?* and, *how the robots can be controlled as singular entities and/or as a team?*. The answers to these questions should permit one operator to be capable of controlling and/or supervising a team of robots deployed on an unknown environment. Consequently researchers are being encouraged to analyse, design, and test, multi-robot teams controlled by a single operator. Military missions, surveillance, scheduled operations, and so forth, are examples of application of this research.

In this paper we present the resulting analysis of two controlled experiments, in which a Search and Rescue Operation was simulated using USARSim. We modified the number of robots, from 1 to 4, for the first experiment, and from 1 to 3, for the second, and analysed the performance of the operators. The analysed data allows us to identify the critical points in multi-robot tele-operation. **The goal of this study is to set the basis for an adequate single-operator-multi-robot Interaction System design.**

2 Single-Robot vs. Multi-Robot for Exploration Missions

Developments in Graphical User Interfaces (GUIs) for tele-operating a mobile robot have mainly focused the single-robot paradigm. They are mostly concerned on how to provide the operator with the required Situational Awareness for controlling a remote mobile robot [18]. A multi-year study made in a rescue simulated competition [17], revealed poor interface designs. Yanco et al. designed and evaluated a video-centric interface [16], raising a great interest on the GUI design. The Human Robot Interaction INL research group⁶, designed a map centred interface [7], giving to the operator a first person point of view of the environment. Both interfaces were experimentally compared in [4]. The limitation of these two interfaces is that they are difficult to extend to the multi-robot paradigm without considerable re-design. In first mentioned GUI, the video information is tightly related with the robot localization, and thus unfeasible for managing a team of spread robots; the INL GUI, map centred, could be easier to extend, as the mapped environment is common to all robots, but it has been designed in order to keep the operator point of view mostly fixed over the robot, presenting consequently the same problem.

These works have set the basis for a good single-robot interface design. Now, the next step drives through multi-robot systems. There are many advantages of using multiple robots in complex exploration tasks like Search and Rescue: properly coordinated robots can cover faster an area than a single robot, the key problem to be solved is to coordinate the robots so that they simultaneously explore different regions of their environment [9][2]. SLAM can be benefited from the use of multiple-robot sensing by reducing the uncertainty associated to localization and map building [10][12]. To act effectively under uncertainty it is required to accurately estimate the state of the environment, providing information of the dynamically changing situation; individual robots, with uncertain sensors, may not be able to accurately determine the current situation, the team as a whole

¹ Centro de Automática y Robótica, Universidad Politécnica de Madrid, email: avalero@etsii.upm.es

² Lab. of Cognitive Science and Psychology, SAPIENZA - Università di Roma, Italy, email: chiara.saracini@uniroma1.it

³ Centro de Automática y Robótica, Universidad Politécnica de Madrid, email: paloma.delapuerta@upm.es

⁴ Centro de Automática y Robótica, Universidad Politécnica de Madrid, email: diego.rlosada@upm.es

⁵ Centro de Automática y Robótica, Universidad Politécnica de Madrid, email: fernando.matia@upm.es

⁶ <http://www.inl.gov/adaptiverobotics/>

can exchange information to perform a better situation assessment by verifying hypothesis.

From the operator point of view, the control and supervision of a team of robots brings a challenging novelty with many associated difficulties. It is not only a question of operator's available time. He does not only need "free" time to switch control from one robot to another, and supervise them sequentially, he also needs to process all the information coming from the deployed robots and integrate it in order to have a global Situation Awareness of the scenario and the robots [18]. He must be able to send the proper commands, not only to the single robots, but also to the team, in order to carry out a coordinated action, and so forth. The more the Interaction System enhances the operator in this task, the better he will perform.

Robots, considered as entities are individual sources of information, which cannot be "processed" by a human if not properly integrated. One operator cannot tele-operate simultaneously several robots. If he just switch control from one to another, the performance of the operation is highly reduced [3]. Adams et al. address this problem proposing not to consider the single robots, but the team as a whole entity, integrating the information from the team and sending team commands [15]. The limitations of this approach is that it presupposes that the robots are able to autonomously carry out the tasks assigned to them by the team coordination layer, what is not feasible in hazardous scenarios like Urban Search and Rescue. Goodrich et al., propose to dynamically adjust the autonomy level to keep the operator workload within tolerable ranges responding to environment and workload changes [6]. This would minimize the time required by each robot, permitting the operator to switch among them. The limitation is that the robots are not treated as a team. Burke and Murphy, based on field studies consider that an operator is not suitable for driving more than one robot simultaneously in *real* Search and Rescue Missions [1]. Scholtz et al. experimented in autonomous off-road driving, analysing the number of times that an UGV requires operator assistance to complete the task. Based on this analysis they conclude that one operator can only supervise two UGVs simultaneously in outdoor environments [11]. Fong et al. proposed a system in which the operator is just a supervisor, who is asked for help by the robot when his assistance is required [5]. Unfortunately no study of the scalability of Fong's system was made.

Based on experimental data, we analyse the case of multi-robot control in Search and Rescue Missions. This application scenario is more complex than the off-road autonomous navigation studied by Scholtz et al. as in their application the UGV should not deal with exploration. We are more flexible than Burke and Murphy, not applying all the hazardous conditions that a *real* mission would present. We have used as test-bed the NIST simulator: USARSim; the "emergency scenarios" are those used in the 2006 RoboCup Competition⁷.

3 Interface description

Our first interface prototype was inspired by the works of [16] and [7], which are probably the most evaluated designs in the literature for single robot GUIs. In addition, we analysed the multi-robot problem, and re-designed their GUIs to support the control and supervision of a team of robots. This interface was used in the 2008 RoboCup Competition, for controlling a team composed by three ground robots and an aerial vehicle. An experimental evaluation of the first prototype can be seen in [14].

The second prototype includes two displays:

The Complex Display (Fig. 1(a)): It includes allocentric and egocentric views of the scenario. It consists of three views: a *Local View of the Map*, that can be set to have the robot facing up, or to have the map north oriented; a *Global View of the Map* giving a bird's eye view of the explored area, and a *pseudo-3D View* giving a first person view, including the video feed-back on its real position in the environment.

The Team Display (Fig. 1(b)), which gives a bird's eye point of view of the explored area with the robots within, for an easier team control and supervision. The operator can zoom a particular area or see the whole explored area. He can switch from one robot to another by clicking on it, and set the target point clicking on the map. Additionally he can tag the map by adding information about the scenario. Victims found, snapshots taken, tagged areas, robots path, a-priori maps, and other retrieved information, can be visualized on the map.

The operator can control the robot in four modalities: *tele-operation*, *safe tele-operation*, *shared control* and *autonomy*. In the *safe tele-operation* mode the system prevents the robot from colliding with obstacles, limiting the speed according to the obstacles distance. In the *shared control* mode the operator sets a target point for the robot by directly clicking on the map, which the robot must try to reach. The *autonomy* policy depends on the mission. For Search and Rescue, our robot explores autonomously the area visiting the unknown places and taking pictures of detected victims [2].

4 Interface Evaluation

As we said at the beginning, an effective operation of multiple robots requires that the operator's time be distributed among the robots; this allows him to supervise the whole team and intervene whenever this is required. The experiment we are presenting here aimed to evaluate scalability of the mechanism of autonomy adjustment that we implemented. Furthermore, subjects were asked to fill out questionnaires as a way of measuring the usability of the interface for multiple-robot control and supervision.

4.1 Experiment Design and Procedure

Forty subjects (thirty-five males, five females) participated in the experiments. All of them were either master students, PhD students, or senior researchers in the field of engineering. None of them had had previous experience with the interface. The subjects were randomly divided into four groups. Each subject in a given group had to accomplish a Search and Rescue Mission while controlling a team of robots (*number of robots* $\in [1 - 4]$). Every subject went through a twenty-minute training program to acquire a basic knowledge of the functionalities provided by the interface. The training scenario was taken from the RoboCup Competition and was similar in difficulty to the scenarios used for the experiments.

The subjects were asked to run two experiments. In the first experiment, they had to explore an unknown office building (indoor scenario) in order to search for victims. They were given 15 minutes. After completing this experiment, they filled out a usability questionnaire. In the second experiment, they were asked to explore an unknown urban outdoor area (with the same number of robots), also with the aim of finding victims. In this paper we will only analyse the results concerning the indoor scenario. Figure 2 shows a screen-shot of the indoor scenario.

⁷ Available at <http://usarsim.sourceforge.net>



(a) Complex Display

(b) Team Display

Figure 1: Desktop-interface.



Figure 2: Indoor Scenario used in the Experiments

4.2 Data analysis

The focus of our analysis is the influence of the factor *number of robots* $\in \{1, 2, 3, 4\}$. The variables studied are:

- The area explored divided by the total time of the mission.
- The percentage of the total time in which the robots were operating (moving) simultaneously.
- The relation between the time the robot remained stationary in *Shared Mode* and the total time in that mode;
- The relation between the time the robot remained stationary in *Tele-operation Mode* and the total time in that mode;

4.2.1 Area Explored

A measure of operator performance is the area covered by the team of robots. When an operator is controlling several robots simultaneously, he may neglect some of them, thus decreasing the time of simultaneous operation, as we have just seen. That said, this "low utilization" might be justified if the area explored by the robots is significantly large.

We made an Analysis of Variance (ANOVA) for the explored area⁸. The best level of the factor corresponds to one robot, but it is not significantly different from 2, 3 or 4 robots, because the intervals are overlapped. Figure 3(a) shows this result.

4.2.2 Simultaneous Operation

When controlling a team of robots, the operator must try to use all of the robots simultaneously. As he can tele-operate only one robot at a time, the rest of the robots have to be working at some level of autonomy: *Shared Control*, when a target point is set, or, when it is not, then *Full Autonomy*. In the best case scenario, all the robots composing the team should be operating (moving) at any given time. An effect of operator neglect on the system performance can be seen in [3]. The present study measures the amount of time that all the robots were moving simultaneously.

An ANOVA was run. The main effect of the *number of robots* is significant. Fig. 3(b) shows the main effects of *number of robots*⁹. The simultaneous operation is significantly greater when the operator is controlling one robot.

4.2.3 Stop Condition Analysis

We also found it useful to analyse the conditions in which a robot was neglected by the operator. A robot is neglected when the operator is not aware of the state of the robot. This may occur in two different situations: A robot has finished its task and so is waiting for a new command; a robot is stalled. As the number of robots increases, the operator controlling a team can lose "*Robot State Situation Awareness*" [4]. This lack of SA indicates an overloaded operator who can be expected to yield a worse performance.

The lack of SA can increase in severity depending on the level of autonomy on which the robot is working. Robot inoperability can occur in a variety of different situations:

⁸ Due to the extension of this paper the detailed ANOVA data is not presented, as it has been considered more explanatory the figures of the confidence intervals. For detailed data please refer to [13]

⁹ Bonferroni adjustment has been applied to solve the multiple comparisons problem.

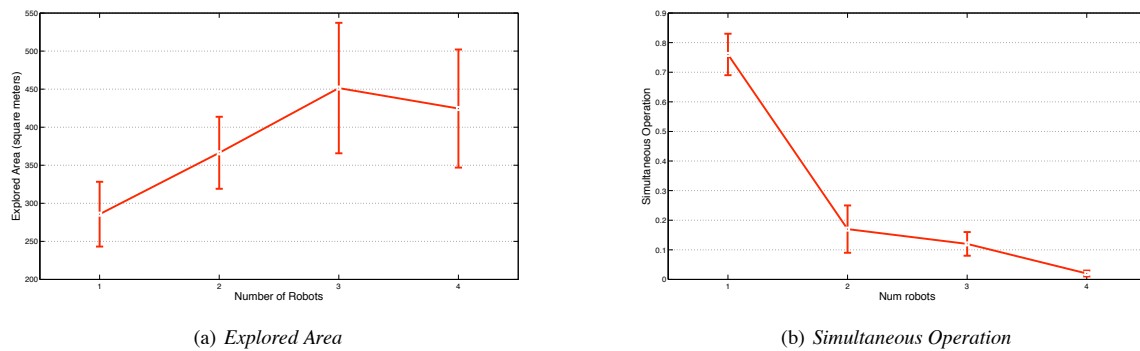


Figure 3: Means and 95% Bonferroni intervals.

Stop in Shared Mode We analysed the time the robots were stopped in *Shared Mode*. When the robot has been sent to a target point (*Shared Mode*), the robot may be stalled. Most usually, robot standstill occurs when the robot has already reached the destination point. The robot remains stationary as it awaits new commands. This situation indicates an even greater lack of SA than the one previously described. After all, the operator knows that the robot will navigate autonomously only until it arrives at the target point (supposing that it does in fact reach this point), and he should therefore be able to estimate the time required to carry out the given task. Figure 4(b) shows the main effect of *number of robots*. The analysed data is significantly smaller when operating one robot. In addition, there are no significant differences between 2, 3 and 4 robots, since the intervals are not overlapped.

Stop in Tele-Operation The greatest lack of SA is manifested when a robot is left standing still in *Tele-operation Mode*. We analysed the $\log(Y)$. The *number of robots* is a significant factor. In Figure 4.2.3, the main effect of the *number of robots* is shown. The best performance in Tele-Operation is with one robot.

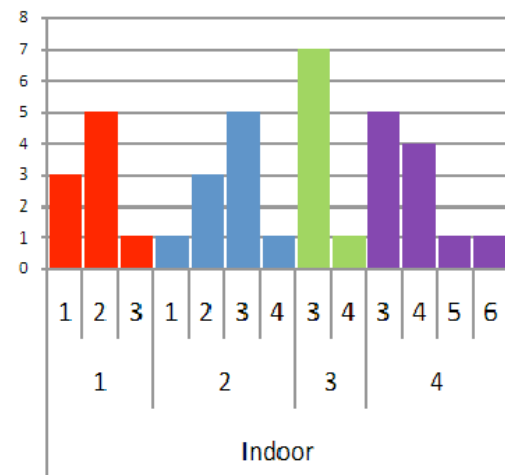
4.3 Questionnaires

After completing the task the subjects were asked to fill a questionnaire. We will not present all the collected data but only the subjective perception of the dimension of the team. That is, subjects were asked how many robots they think they can control in order to maximize the performance. This data is represented in Figure 5(a). Subjects scored each operation mode (1-worst, 5-best). This data is represented in Figure 5(b)

4.4 Discussion

4.4.1 Interface Design and Operation Modes

A first observation is that for one robot, the subjects almost always used the *Tele-operation Mode*. This suggests that the interface provides enough information to an operator to enable him to control a robot remotely and to manoeuvre it in a precise and secure way. If the operators' situational awareness had been insufficient for this task, the resulting feeling of lack of control, or of the inability to send precise commands, would have led them to prefer *Shared Mode* instead. In fact, a few subjects that showed small capabilities to drive



(a) Preferred Number of Robots. X Axis: Number of Robots Controlled. Y Axis: Preferred Number of Robots

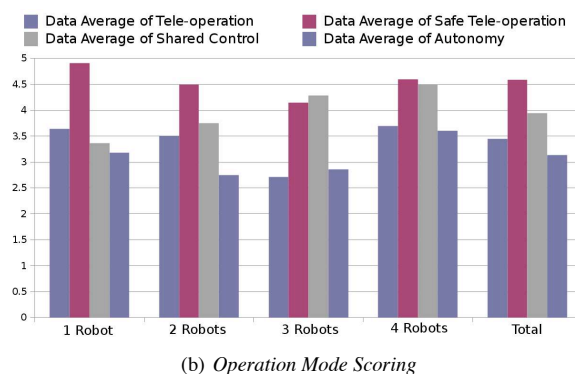


Figure 5: Questionnaires Analysis

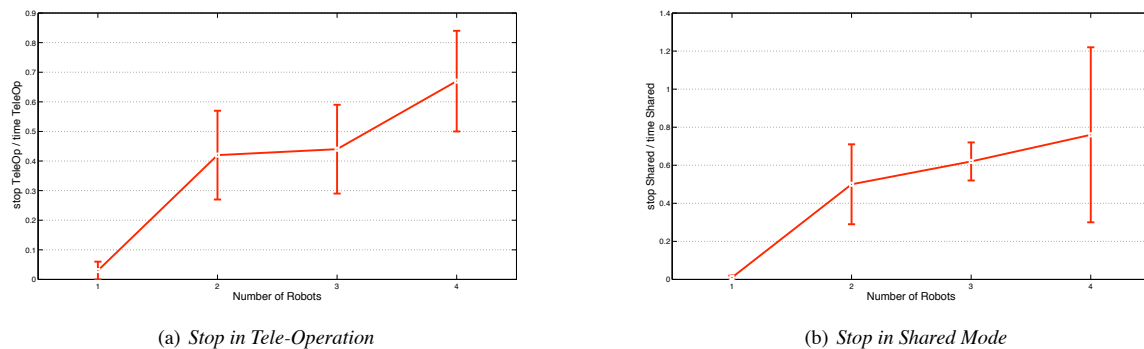


Figure 4: Means and 95% Bonferroni intervals.

the robot, used mostly the Shared Mode, commanding through target points.

Analysing the scoring the operator gave to the Operation Modes (Figure 5(b)) we conclude that the Tele-operation mode is the less appreciated. Operators considered this mode very dangerous, as collisions were not prevented at all. Instead the Safe Tele-operation Mode received a good scoring, which showed that the changes applied to the second version of the interface were successful. The Autonomy Mode received also a low scoring.

4.4.2 Operator Performance

The data analysis concludes that the optimal robot:human ratio, when account is taken of the utilization of robots and of the area explored, is 1:1. When we consider the utilization of the robots, this result is not surprising. When an operator controls more than one robot, he inevitably neglects some to a certain degree, which results in correspondingly less optimal utilization.

The use of several robots would be justified, even when they are not fully exploited, if the area explored increased with the number of robots. We had hypothesized that the area would expand to its with 2 or 3 robots, and would subsequently decrease due to an excessive operator workload. In actuality, the result of the experiments shows that the area covered by the robot team increases only when we use 3 or 4 robots, but its not better with 2 robots in relation with one. We believe that this result is caused by the great decrease in robot utilisation due to the shift from one robot to two. As we can see in Figure 3(b), when the operator is controlling two robots, they are moving simultaneously less than the half of the time. Thus, even if both robots explore different areas, the total area is not bigger than when an operator controls just one robot.

The *Stop Condition Analysis* helps us to understand why this performance fall-off occurs when the operator is using more than one robot. Let us suppose that the robot is working in *Shared Mode* (in other words, when the operator has set a target point that the robot must try to reach autonomously). If the operator is controlling multiple robots, he usually does not realize immediately that the particular robot has arrived at the target point. When the operator is controlling one robot, however, this does not happen, as he maintains supervision of the robot's actions during the entire task. Again, the problem arises only when he is controlling two or more robots. This is consistent with the usability questionnaires. Most users said that they felt

unable to keep track of the state and task of all the robots composing the team.

Something slightly different happened when the robot was left stationary in *Tele-operation Mode*. Direct observation of the experimental runs revealed to us the reason why this situation happened. When a robot was moving in *Autonomy Mode* or *Shared Control* and got stalled (that is, became unable to follow its path), as soon as the operator realized this, he switched control to the stalled robot, leaving the robot he was controlling previously in whatever state it happened to be in at the time of the switch. For example, if this old robot was in Tele-operation, it was left stationary, and the operator usually forgot about it and simply continued to manoeuvre the the new robot until he realized that he had forgotten the old one. When we asked the operators if they were aware of their mistake in this regard, they typically answered that it was very difficult to keep track of everything they were doing on account of the complicated nature of the task they were performing.

Analysing the subjective perception that the subjects had of the optimal team dimension (Figure 5(a)) we observe that operators considered they were able to command a team of up to 3 robots in the indoor scenario and up to 2 for the outdoor scenario. This perception changes with the number of robots they effectively commanded. Operators commanding 4 robots considered that the number was excessive, operators controlling 3 robots felt the number was ok, while for 1 and 2 robots they thought they could control more robots. We conclude that at this point the interface was not yet well designed for supporting robot teams, as the operator "*optimism*" indicates that he is capable of more. We concluded that the improvement should be made the autonomy adjustment management.

5 Autonomy Management Evolution

The major weakness that appeared in the previous study concerns the excessive amount of time the robots were neglected when they were not performing a task, that is, when they had stopped and were waiting for incoming commands. Taking account of system performance and the test subjects' input, we hypothesized that the problem might be caused by one of the following two factors:

- The robotic system was unable to complete the task autonomously and no error message was sent to the operator.
- The operation modes were not enough adequate to the operator's task

All these issues led to the development of a new interface prototype 3, focused on enhancing the operation modes, particularly the Shared Mode, understood as the spectrum between full autonomy and the Safe Tele-operation. We will describe these improvements in this paper. After describing the interface, we will present a set of controlled experiments performed for the sake of analysing the improvements.

5.1 Operation Modes Evolution

In the previous version there were four independent operation modes, and the operator decided which one was more suitable to the task requirements. In version 3, the management of the autonomy level changed. The system no longer waits for a "precise command" for each operation mode: target point in Shared Control, speed and jog in Tele-Operation. In the third version, the autonomy is adjusted in a multi-layered form. The layered system may be seen in Figure 6. The following list describes the different autonomy levels, beginning with full autonomy and ending with pure tele-operation (based on [8]).

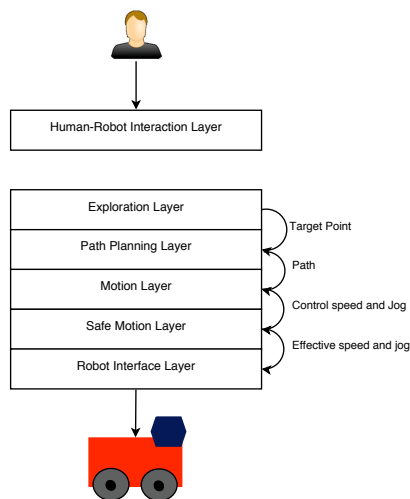


Figure 6: System Layers

1. The system proposes a target point;
2. if there is no input from the operator the system proposes a path to reach it, or
3. the operator sets a target point and the system proposes a path to reach it;
4. if the operator does not change the path the system follows the path, or
5. the operator sets a path and the system follows it;
6. the operator sets a control speed and jog and the system adjust the real speed and jog to avoid collisions;
7. the operator sets the real speed and jog.

From this list we can see that the operator has five command possibilities: set a target point (or sequence of target points), set a path, set a control speed and jog, set the real speed and jog. Depending on the operation mode of the robot: Tele-Operation, Safe Tele-Operation, Shared Control or Autonomy, the system will react differently to each input. The possibilities are:

Tele-Operation and Safe Tele-Operation. If the operator sets a target point or path, the robot changes to shared control.

Shared Control. In shared control mode, the system tries to reach the target point, or sequence of target points, set by the operator. This is done by means of two mechanisms: the calculation of a path and the control of robot motion to follow that path [2]. The operator can give the following commands.

- Speed Commands: The system remains in Shared Control, but while the operator continues sending speed commands, the robot's motion is not controlled by the system. When the operator stops sending commands, the system retakes motion control. This is very efficient when the robot is following a path and remains stalled, because the operator can get the robot out of its stalled condition without modifying its former task. As soon as the operator releases the control, the robot continues the previous task.
- The operator sets a path, the robot follows that path, and it then calculates a new path to the former target point, if there was one. In this way, if the operator disagrees with the path proposed by the system, he can modify it, and the robot adjusts its proposed path to the path set by the operator.

Autonomy. In full autonomy, the system chooses the most suitable target point in order to accomplish the mission [2]). The operator may send the following commands:

- Speed, Jog and Path, as in Shared Mode. The system gives the suitable control to the operator, but remains in autonomy, so that, when the operator stops commanding, the system keeps on working in autonomy. Like shared mode, this mode is suitable in situations in which the operator does not approve the actions taken by the system.
- The operator sets a target point or sequence of target points. The system tries to reach them, and, once they have all been reached, it keeps on exploring autonomously.
- The operator sets a desired path. The system tries to follow it, once it has completed the task, and if there are not target points, it keeps on exploring autonomously (if there is a desired path and a sequence of target points it will follow firstly the desired path, and afterwards it will try to reach the target points).

Each layer is in charge of performing an action. The more layers are working, the higher the autonomy level is. Thus, if all the layers are running, the system is working in full autonomy, while, if none of them is working, the operator has full control of the robot. The operator can command at every level of the layered system, substituting the actions of the higher layers. It is important to notice that this does not require him to change the operation mode. For example: The system may be operating in "Autonomy Mode," so that, by default, all the layers are running, but at any moment the operator can command at any level, while the system remains in "Autonomy Mode". At the same time, every layer provides feed-back indicating whether or not it has performed its respective task, so the operator has correct system state awareness and knows at which level he must act.

The new autonomy management is thought to provide the operator with two advantages:

- Fast recovery from navigation and exploration errors or bad performance. The operator can act at the error level, without re-configuring the robot task or changing the operation mode.

- Long-term commands and more granulated autonomy levels. The operator can set a sequence of target points, giving the robot a longer-term task. Moreover, since he can set the path, something that was not possible in previous version, he can send the robot along safe paths.

In this way, the system takes more advantage of the operator's expertise.

6 Evaluation

As we said at the beginning, an effective operation of multiple robots requires that an operator distribute his operating time among the robots, so that he can supervise the whole team and intervene whenever this is required. The experiment is equivalent to the previous one and aims to evaluate the scalability of the mechanism of autonomy adjustment that we implemented.

Forty-five subjects (forty-one males, four females) participated in the experiment. All of them were either master's students or Ph.D. candidates. The subjects were randomly divided into three groups. Each subject in a given group had to accomplish a Search and Rescue Mission while controlling a team of robots ($number\ of\ robots \in [1 - 3]$). Every subject went through a forty-minute training equivalent to the training performed for the previous experiment.

The participants were asked to explore an unknown office building (indoor scenario) in search of victims. This scenario was the same as that used in the previous experiment.

6.1 Data analysis

The focus of our analysis is the influence of the factor *number of robots* $\in \{1, 2, 3\}$. The variables studied were:

- The area explored divided by the total time of the mission.
- The percentage of the total time in which the robots are operating (moving) simultaneously in Shared Control Mode and in Tele-Operation.

6.1.1 Area Explored

The area explored was measured in square meters. The number of subjects in each group is 15. A one-way ANOVA was carried out to study if there were any significant differences. Confidence intervals can be seen in Figure 7

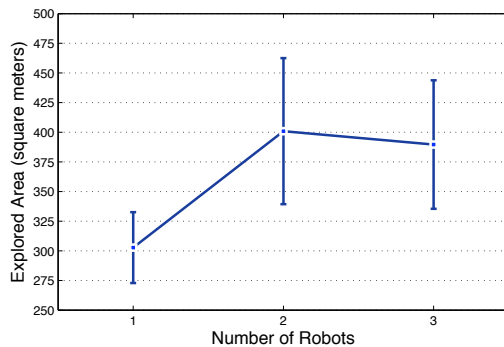


Figure 7: Explored Area - Confidence Intervals (95 %)

6.1.2 Stop Condition Analysis

The previous experiment showed that one of the reasons of a bad performance was that operator were neglecting some robots of the team. We hypothesized that this was caused by an insufficient autonomy management paradigm. We modified this paradigm in order to improve performance. This analysis, together with the previous one aims to analyse whether there is a significant improvement. We analysed stop condition in Tele-Operation and in Shared Mode. The results are displayed in the form of a graph in Figures 8(a) and 8(b).

6.2 Discussion

The major improvements made to the third version of the interface regarded the autonomy management and the new Shared Control commands. The results obtained for the previous version showed that the optimal number of robots for maximizing the area explored and minimizing the number of robots deployed was one. This experimental result led us to conclude that the interface did not provide the operator with suitable mechanisms for adjusting robot autonomy, which is crucial for the control of multiple robots. At the beginning of this paper, we presented a work of Scholtz et al., in which the authors conclude that the optimal operator ratio for off-road autonomous driving is 1:2. We set out to obtain a similar result for the current interface version. The analysis of the area explored reveals a considerable performance improvement over the previous version. While in the previous version increasing the number of robots did not lead to an increase in the area explored, in the current version the explored area is greater for 2 robots than for one robot. The differences are statistically significant. For 2 and 3 robots, there is no statistical difference (Figure 7).

The rest of the data studied enabled us to understand the key to this improvement. This phenomenon receives some explanation in light of the stop condition analysis (Section 6.1.2). In the previous version, there was a considerable jump in the stop condition between one robot and more than one robot, a jump that then entailed a corresponding significant loss of performance capability. In the data analysed above, this jump is mitigated, as can be seen in Figures 8(a) and 8(b). Furthermore, comparing the stop condition in Shared Mode in the previous interface with that of the current version, we see that the time the robots were stopped in Shared Mode is significantly less for 2 robots.

7 Conclusions

In this paper, we have evaluated the two different versions of our interface. This evaluation was presented on the basis of controlled user experiments.

The major change between the versions was the inclusion of new command capabilities for the Shared Mode as well as a different adjustable autonomy management. In version 2, the operator could set a target point only when working in Shared Control Mode. In version 3, the operator can set a sequence of target points or a path. Furthermore, each operation mode allows him to bypass any of its "proper" functionalities. For example, even if he is working in Autonomy Mode, the operator can set a target point, or temporarily control the speed and jog of the robot. These changes led to an improvement in the optimal operator robot:ratio. While in version 2 the performance of the operator decreased for more than one robot, in version 3 the operator achieves a better performance with 2 and 3 robots.

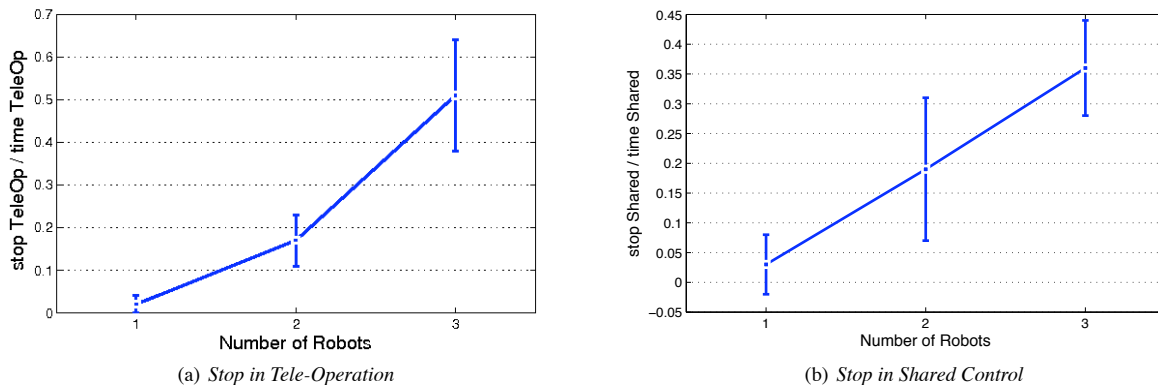


Figure 8: Means and 95% Bonferroni intervals.

We have learnt that commanding a team of robots requires proper shared control capabilities which enable the operator to act at the required autonomy level according to the task he must perform and the team state. If some of the robots are performing well in autonomy he can send higher level commands while commanding at a lower level those robots in most need of attention. Splitting the former Shared Mode into 3 commanding possibilities has been proven to be more suitable for this, as it has been seen in the experiments. Furthermore, long term commands, like setting the desired path or the sequence of target points decrease the negative effect of neglecting a robot.

The third version of the interface was used in the last RoboCup Edition (2009, Austria), in the Rescue Simulation League. It received a Technical Award for the most innovative interface. A fourth version has been developed introducing new functionalities in the *Full Autonomous* mode. Further work will include the conducting of user experiments to evaluate the improvements brought by Version 4. We hypothesize that, with the new functionalities offered in Autonomy Mode, operators will use it more often. The goal is to increase optimal team size to three or even four robots.

REFERENCES

- [1] J. L. Burke and R. R. Murphy, 'Human-robot interaction in usar technical search: Two heads are better than one', in *IEEE International Workshop on Robot and Human Interactive Communication*, pp. 307–312, (2004).
- [2] D. Calisi, A. Farinelli, L. Iocchi, and D. Nardi, 'Multi-objective exploration and search for autonomous rescue robots', *Journal of Field Robotics, Special Issue on Quantitative Performance Evaluation of Robotic and Intelligent Systems*, **24**, 763–777, (August - September 2007).
- [3] Jacob W. Crandall, Michael A. Goodrich, Dan R. Olsen, and Curtis W. Nielsen, 'Validating human-robot interaction schemes in multitasking environments', *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, **35**(4), 438–449, (2005).
- [4] Jill L. Drury, Brenden Keyes, and Holly A. Yanco, 'Lassoing hri: analyzing situation awareness in map-centric and video-centric interfaces', in *Proceedings of the Second ACM SIGCHI/SIGART Conference on Human-Robot Interaction*, pp. 279 – 286, (2007).
- [5] Terrence Fong, Charles E. Thorpe, and Charles Baur, 'Advanced interfaces for vehicle teleoperation: Collaborative control, sensor fusion displays, and remote driving tools', *Auton. Robots*, **11**(1), (2001).
- [6] Michael A. Goodrich, Timothy W. McLain, Jeffrey D. Anderson, Jisang Sun, and Jacob W. Crandall, 'Managing autonomy in robot teams: observations from four experiments', in *Proceedings of the Second ACM SIGCHI/SIGART Conference on Human-Robot Interaction, HRI 2007, Arlington, Virginia, USA, March 10-12, 2007*, pp. 25–32, (2007).
- [7] Curtis W. Nielsen and Michael A. Goodrich, 'Comparing the usefulness of video and map information in navigation tasks', in *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction, HRI 2006*, pp. 95–101, (2006).
- [8] Raja Parasuraman, Thomas B. Sheridan, and Christopher D. Wickens, 'A model for types and levels of human interaction with automation', *Systems, Man and Cybernetics, Part A, IEEE Transactions on*, **30**(3), 286–297, (2000).
- [9] Sarvapali D. Ramchurn, Alex Rogers, Kathryn Macarthur, Alessandro Farinelli, Perukrishnen Vytelingum, Ioannis A. Vetsikas, and Nicholas R. Jennings, 'Agent-based coordination technologies in disaster management', in *7th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2008), Estoril, Portugal, May 12-16, 2008, Demo Proceedings*, pp. 1651–1652, (2008).
- [10] Diego Rodriguez-Losada, Fernando Matia, Agustin Jimenez, and Ramon Galan, 'Local map fusion for real-time indoor simultaneous localization and mapping', *Journal of Robotic Systems*, **23**(5), 291 – 309, (April 2006).
- [11] Jean Scholtz, Brian Antonishek, and Jeff Young, 'Operator interventions in autonomous off-road driving: effects of terrain', in *Proceedings of the IEEE International Conference on Systems, Man & Cybernetics: The Hague, Netherlands, 10-13 October 2004*, pp. 2797–2802, (2004).
- [12] Sebastian Thrun and Yufeng Liu, 'Multi-robot slam with sparse extended information filters', in *Robotics Research, The Eleventh International Symposium, ISRR, October 19-22, 2003, Siena, Italy*, pp. 254–266, (2003).
- [13] Alberto Valero, *Evolutionary Design of Human-Robot Interfaces for Teaming Human and Mobile Robots in Exploration Missions*, Ph.D. dissertation, Sapienza - University of Rome, January 2010.
- [14] Alberto Valero, Gabriele Randelli, Chiara Saracini, Fabiano Botta, and Massimo Mecella, 'The advantage of mobility, mobile tele-operation for mobile robots', in *Proceedings of the AISB-HRI Symposium "New Frontiers in Human - Robot Interaction (To appear)*, (March 2009).
- [15] Lovekesh Vig and Julie A. Adams, 'A framework for multi-robot coalition formation', in *Proceedings of the 2nd Indian International Conference on Artificial Intelligence, Pune, India, December 20-22, 2005*, pp. 347–363, (2005).
- [16] Holly A. Yanco and Jill L. Drury, '"where am i?" acquiring situation awareness using a remote robot platform', in *Proceedings of the IEEE International Conference on Systems, Man & Cybernetics: The Hague, Netherlands, 10-13 October 2004*, pp. 2835–2840, (2004).
- [17] Holly A. Yanco and Jill L. Drury, 'Rescuing interfaces: A multi-year study of human-robot interaction at the aaai robot rescue competition', *Auton. Robots*, **22**(4), 333–352, (2007).
- [18] Holly A. Yanco, Jill L. Drury, and Jean Scholtz, 'Awareness in human-robot interactions', in *Proceedings of the IEEE Conference on Systems, Man and Cybernetics, Washington, DC, October 2003*, (2003).

Development of HRI Models by means of a Cognitive Walkthrough Approach

Astrid Weiss and Florian Foerster and Daniela Wurhofer and Manfred Tscheligi¹

Abstract. Well established interaction models are the source of successful human-robot interaction user studies. We propose a methodological adaption of the Human-Computer Interaction expert evaluation method cognitive walkthrough as approach to develop interaction models based on video data from previous user studies. To investigate this methodological adaption we conducted a cognitive walkthrough taking advantage of videos of two user studies with two different humanoid robots. Based on the implications gained from the cognitive walkthrough, a set of implications for a general interaction model was developed which served as basis for new interaction scenarios with reduced usability problems.

1 INTRODUCTION

Two key elements for successful human-robot interaction are interaction *models*, representing procedures and characteristics of the interaction in a general way and interaction *scenarios*, representing scripts of the story behind the interaction. Scenarios are also a proven method in the area of HCI/HRI to describe concrete situations of human-robot interaction [8], involving personas and specific tasks, whereas interaction models are more general, focusing on the input-output stream of an interaction. The development of interaction models helps to increase usability aspects like intuitiveness and learnability on a basic level and assures these aspects for all interaction scenarios derived from it. Thus, interaction models build the basis for a broad range of interaction scenarios with a minimal number of usability problems.

But how to generate such general human-robot interaction models? In this position paper, we propose an adaption of the classical cognitive walkthrough method from Human-Computer Interaction to identify the critical areas of human-robot interaction scenarios and to derive implications for more general human-robot interaction models rooted in the respective context of use. In other words this paper aims to present a method for structuring scenarios in action sequences. By describing a concrete application of our adapted cognitive walkthrough in two case studies, the reader is introduced to the stepwise translation process of HRI scenarios to HRI models. Our adaption is mainly characterized by the usage of concrete interaction scenarios represented in video-taped, task-based human-robot interaction user studies. These scenarios are then presented to a group of experts who walk through the interactions and identify the major usability problems regarding intuitiveness and flexibility. Based on the combination of a user-centered hierarchical task analysis of the examined scenario and the identified usability problems, a general interaction model can be proposed.

The paper is organized in the following way: First, we give an overview on the related work concerning the cognitive walkthrough method and scenario-based approaches in HRI, as well as high level interaction models for human-robot interaction. Next, we describe the methodological set-up of our cognitive walkthrough adaption for getting more general interaction models of a human-robot interaction out of concrete scenarios. By presenting two case studies we then illustrate the concrete usage of our method. Finally, we discuss the scope and limitations of the presented methodological adaption and show possible future use cases of it.

2 RELATED WORK

Our approach is characterized by an indirect assessment of human-robot interaction, i.e. the observation of the interaction scenarios presented on video. In general, user interaction scenarios help to make a system's use explicit [9]. According to Rosson and Carroll, user interaction scenarios represent an effective design tool for inspiring and systematizing the iterative development of interactive systems.

Videos can be used as instantiations of scenarios in order to represent prototypes, thus helping to bridge the gap between abstraction and detail in the design process [4], [5]. Up to now, video-based methods are not very common in HRI. However, video trials could represent an alternative to live trials, as running a live human-robot interaction trial under controlled conditions with statistically valid results requires a relatively high investment of time, resources, and personnel [15].

A fundamental question when using video based methods in HRI is if people's perceptions from live and video trials are comparable. This question is discussed by Woods et al. [15]. In a pilot study, they could show that videotaped trials are a valuable technique for investigating people's perception of presented human-robot interaction scenarios, as the participants' preferences and opinions were similar in both trials (video-based and live). The authors indicate that video trials should not be seen as a replacement for live trials but rather as a complementing research tool.

In a study of Kidd [3], it could be shown that there are no fundamental differences between present and remote interaction with a robot head (via video) when rating personality traits.

A study of Sydral et al. [11] focused in particular on how participants evaluated three robotic systems in terms of user experience, after having presented a brief video showing how a human interacts with a robot. The authors found out that video-prototyping represents a valuable instrument for evaluating how prototypical systems are experienced by users.

In a study of Buur et al. [1], videos were used to discuss different types of mobile interfaces. Therefore, the participants had to select

¹ HCI&Usability Unit, ICT&S Center, University of Salzburg, Austria, email: firstname.lastname@sbg.ac.at

three problematic situations, which in their view would be good test cases. After that, each group prepared a short video scenario which was later shown to the other participants and then discussed.

The studies presented above let us assume that videos are a promising instrument for investigating human-robot interactions under certain conditions (i.e. at an early stage of development).

In order to describe the way a system acts, interaction models can be used. Such models act as a base for successful design of systems including high usability. Interaction involves at least two participants, the user and the system; interaction models are helpful for understanding exactly what is going on in the interaction and for identifying the likely root of difficulties. In general, interaction models are used for planning, describing, and analyzing systems/interactions with systems. What is more, a model gives the possibility to compare different applications/systems/prototypes and provides a unified language. By applying interaction models, possible procedures and interactions can be simulated and possible problems can be identified [2].

In the area of HRI, interaction models do not have a long tradition and thus only exist sparsely and incompletely. One of the best known interaction models in HRI is that of Scholtz [10], which is based on Norman's seven stages model [6]. Scholtz assumes that HRIs are of five varieties, each needing different information and being used by different types of users. Therefore, five roles are defined: supervisor, operator, mechanic, bystander, and teammate role. Supervisory and teammate roles imply the same relationships between humans and robots as they do when applied to human-human interactions. An operator is needed to work "inside" the robot; this includes adjusting various parameters in the robot's control mechanism to modify abnormal behavior, changing a given behavior to a more appropriate one or taking over and tele-operate the robot. The mechanic role is adopted when a human needs to adjust physical components of the robot, such as adjusting the camera or various mechanisms. A bystander does not explicitly interact with a robot but needs some model of robot behavior to understand the consequences of the robot's actions. Based on these roles, possible models of interaction for the various HRI roles are presented, namely supervisor interaction, operator interaction, mechanic interaction, peer interaction and bystander interaction.

3 METHODOLOGICAL SET-UP

3.1 Video-based cognitive walkthrough

3.1.1 *The Cognitive Walkthrough as a Usability Inspection Method*

The cognitive walkthrough is a usability inspection method to evaluate the design of a user interface [7]. It puts special emphasis on the ease of learning, primarily by exploring the application without a prior instruction or training. The method is based on observations showing that users prefer to learn a software by exploring it. To evaluate a proposed interface, experts are asked to walk through action sequences of a specific task and assess the cognitive effort required by the user when interacting with the interface in the optimal course of action. The method aims at finding "mismatches between users' and designers' conceptualization of a task" [14].

The cognitive walkthrough used within usability inspection in the field of HCI consists of two phases. In a first preparatory phase the context for the use of the system is defined. This includes concrete information on the target group, the system and the tasks that are to be performed. The second phase consists of the actual walkthrough

and the locating of potential usability problems. The central question within this second phase is whether the user would behave the way it is required by the interface to successfully solve the given task.

3.1.2 *Adaption of the Cognitive Walkthrough for HRI*

In order to use the cognitive walkthrough for the purpose of examining human-robot interaction scenarios, we slightly adapted the cognitive walkthrough method. However, the main aim of the adapted cognitive walkthrough still was to increase usability aspects like intuitiveness and learnability - which we believe essential for human-robot interaction, especially when designed for unexperienced users - and integrate these into the scenarios. For the first phase - the analysis of the context of use - we resorted to information and scenarios that had already been deployed in the two user studies (see 3.2). According to the scenario, the participant would have the role of a teammate as defined by Scholtz [10]. The scenarios further described the target user and the tasks to be analyzed. These scenarios were already tested in user studies and videotaped at this occasion. In classic cognitive walkthrough the interface or a description of the interface is given in detail at each action sequence. If the interface is not fully developed paper descriptions of the interface can be used. As human-robot interaction is not limited to a graphical user interface, we opted to use video material to present each action sequence to the group of experts. We wanted to give the experts a more complete picture of the interaction instead of only handing out a list of all possible commands and feedback given from the robot. The video-based material of the user studies served as basic input for expert evaluators. We assumed that the video material could help the evaluators to better relate to the potential user. We selected video clips of participants who solved the task without any errors in order not to bias the evaluators. These clips were cut into single action sequences for an easy handling during the cognitive walkthrough sessions.

3.2 The user studies

Two user studies preceding the cognitive walkthrough served as preliminary studies to evaluate the usability of the initial interaction sequence represented in the scenarios. They further had the goal of investigating how novice users experience the collaboration with the humanoid robot. The first of the two user studies was conducted at LAAS Institute of CNRS in Toulouse, France.



Figure 1. Study Setting: HRP-2

The participants had to accomplish two tasks which were based on speech recognition and on interaction with the HRP-2 robot - see Figure 1. Their performance and behavior was observed, recorded on video tape and analyzed at a later date. In task 1, the participants' job was to tell the robot to pick up the ball; in task 2, the participant's job was to tell the robot to put the ball on the table. The participants were introduced into the scenario. They received a list of commands needed in order to fulfil the task as well as some additional commands to interact with the robot. An interaction model itself was not developed. In order to fully understand the task we conducted a hierarchical user-centered task analysis of the interaction scenario, see Figure 2. More details on this user study can be found in [12].

The second user study took place at the LASA Institute of EPFL in Lausanne, Switzerland. In order to find out more about human-robot interaction with the HOAP-3 humanoid robot, the participants had to accomplish two tasks based on learning by demonstration. The tasks were video-taped and accompanied by a retrospective think aloud. The two tasks - both based on demonstration of specific motions - had to be conducted: In task 1, the participants had to teach the robot to push a box away from its working space into the users' direction; in task 2, the participants' job was to teach the robot to close the lid of a box.



Figure 4. Study Setting: HOAP-3

Before starting with the task, the scenario was described by the instructor: "Imagine you are working at an assembly line in a big fabrication plant. A new robot is introduced, which should support you in solving tasks. You can teach the robot specific motions by demonstrating them (meaning move the robot's arm like you expect it to move it later on its own); the robot will repeat the learned motion. You can repeat this "demonstration-repetition-cycle" as long until you are pleased with the result". The interaction model for the scenario can be seen in Figure 5. Based on the scenario and the interaction model we conducted a hierarchical user-centered task analysis of the interaction scenario as already done with the HRP-2 interaction. Figure 3 represents the hierarchical user-centered task analysis of the interaction scenario. More details on this user study can be found in [13].

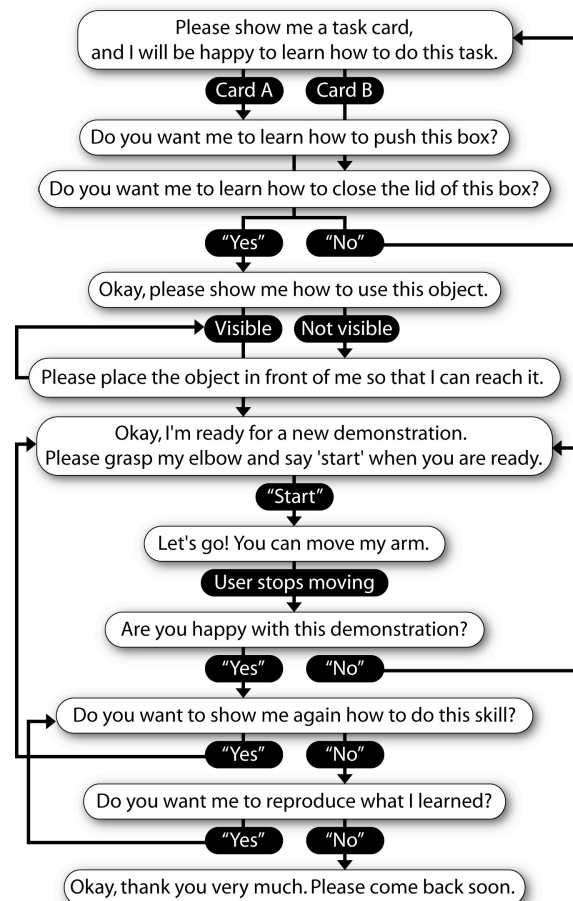


Figure 5. Study Setting: HOAP-3

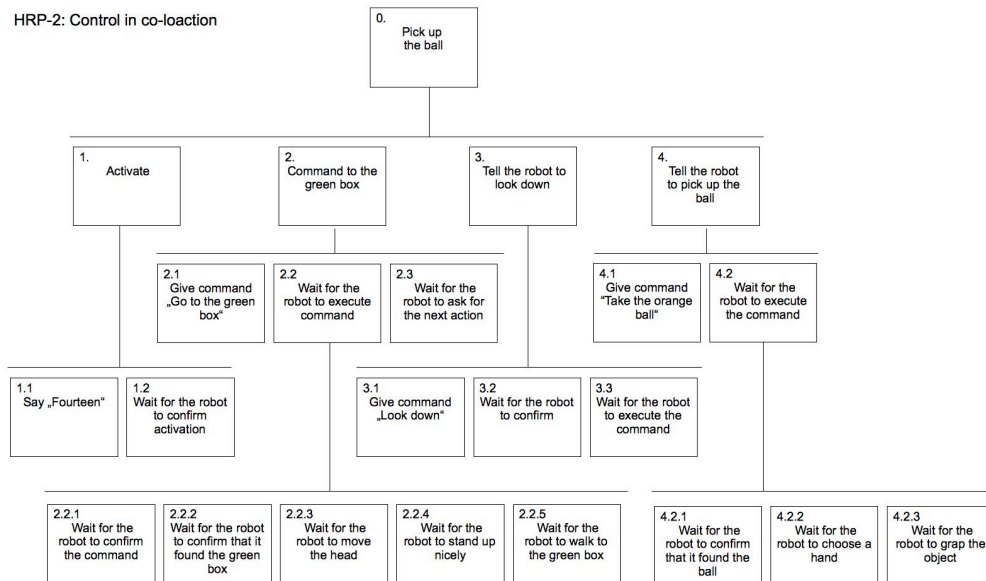


Figure 2. Task Analysis: HRP-2

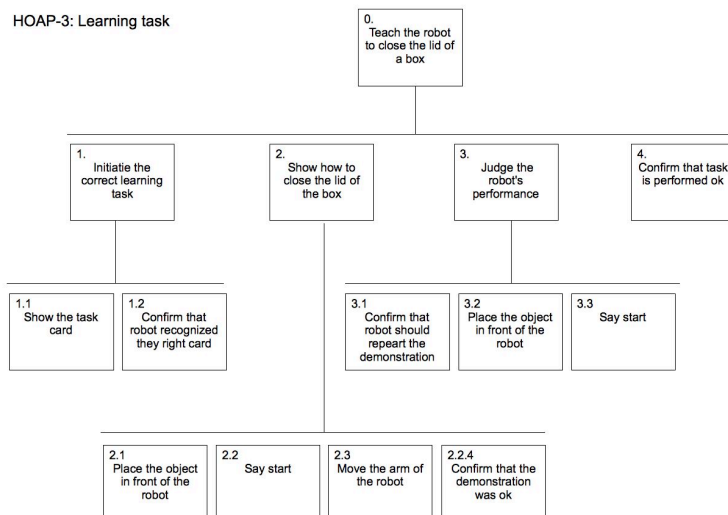


Figure 3. Task Analysis: HOAP-3

3.3 Evaluation setting and procedure

For each task of the user studies a separate cognitive walkthrough was conducted. A group of four experts from the field of Human-Robot Interaction participated in the walkthroughs that were organized and moderated by a researcher. Each session lasted about two hours. An assistant took notes during the whole procedure.

At the beginning of each session, the moderator introduced the experts to the scenario. Each participant received a scenario card (see Figure 6) with a detailed description of the scenario as well as of the two tasks which had to be analyzed by the expert group. For the actual walkthrough, each task was divided into action sequences, the video clips taken from the user studies were prepared accordingly. The evaluators went through each step of the action sequence by watching the video clip first and discussing if and why a user would choose the correct action at each step afterwards. Special attention was paid to whether the robot's feedback was sufficient. After the sessions were completed, the transcript of the sessions was analyzed by two researchers and the results, summarizing cognitive effort and possible problems of each step within the task, were collected.

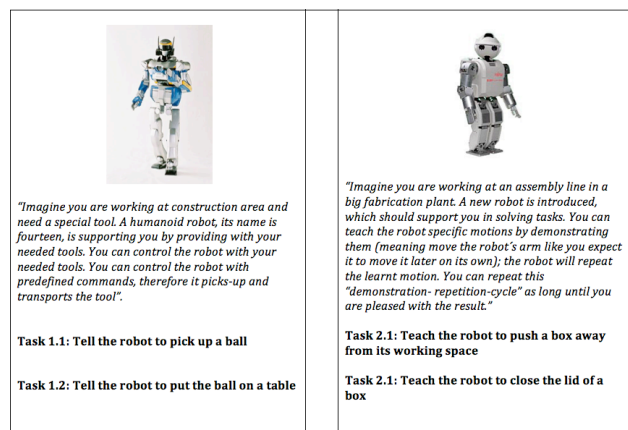


Figure 6. Scenario Card

3.4 Cognitive walkthrough results

Based on the results of the cognitive walkthrough (see problem list of Table 1), we developed a set of implications for general interaction models supporting a high level of learnability and intuitiveness. As theoretical base, we took the definition of learnability from Dix et al. [2]. According to the authors, learnability deals with the characteristics of an interactive system that allows novice users to understand how to use it initially and attain a maximal level of performance. For describing learnability in detail, Dix et al. define five principles which affect learnability. These principles are predictability, synthesizability, familiarity, generalizability and consistency. The results of the cognitive walkthrough and the implications for improving certain aspects/principles of learnability for general interaction models are listed below.

Results concerning synthesizability and consistency Giving feedback is related to the principle of synthesizability, enhancing

Table 1. Overview Cognitive Walkthrough Results

Problem	HRP-2	HOAP-3	Resulting Implications
No information on status of the robot	X		Imp 1
One-dimensional feedback (not multimodal)	X	X	Imp 2
Not feedback from the robot at all	X	X	Imp 1, Imp 3
No predictability of robot action	X		Imp 4
No information on the vision field of robot	X		Imp 4
Low flexibility	X	X	Imp 5
No options of stopping or canceling an action	X	X	Imp 6
No possibility to start interaction/an action	X		Imp 6
Unclear output from the robot		X	Imp 7

learnability in a positive way. Synthesizability should support the user to assess the effect of past operations on the current state. Consistency is a further principle affecting learnability, dealing with the likeness of input-output behavior arising from similar situations or similar task objectives.

Imp 1: Provide visual feedback about status of the robot: It is essential for the user to know the robot's status at every point of interaction. Therefore, cues for understanding whether the robot is activated / is in a mode of receiving commands / executing any commands have to be provided. Visual feedback via lights (on the head, shoulders, in the eyes), body language, gaze direction, gestures and posture could be used for indicating the status of the robot. For example, an inactive robot looks down, whereas an activated robot looks into the user's direction. The advantage of visual feedback is that it is understood easily and intuitively by the users.

Imp 2: Provide multimodal feedback: One feedback modality is often not sufficient for getting clear and unambiguous feedback. More information about the robot's current status could be provided for example by accompanying movements or gestures. For more details concerning this issue, see below Imp 2a.

Imp 2a: Support auditory feedback by other feedback modalities: The understanding of auditory feedback must not be crucial. Especially when a robot uses auditory feedback, this can pose problems. For example, if the noise level of the environment is high, the verbal output of the robot can be difficult to understand. Therefore, other feedback modalities should enhance the auditory feedback.

Imp 3: Provide consistent feedback: The robot's behavior should be consistent and unambiguous (e.g. eyes have to be illuminated every time the robot is active). Furthermore, in case of multimodal feedback, consistency of the different output channels / feedback modalities has to be ensured. Moreover, response times of the robotic system must not be too long, otherwise the user cannot correlate his actions with the system's reactions. Unnecessary feedback/information should be avoided as it distracts the user. Giving immediate and relevant feedback thus supports learnability. In order to provide coherence, feedback should not only be given in case of success (of an action), but also in case of failure.

Results concerning predictability Predictability is a principle which affects learnability. It means that the user should be supported determining the effect of future action based on past interaction history.

Imp 4: Inform about the robot's perspective: For a human interacting with a robot, it is not always clear how the vision of a robot works and which objects can be seen or recognized by the robot. For example, objects that are too close to the robot might not be detected. Information about the robot's perspective/vision range could be given by an ambient display, giving information about the robot's field of vision. This helps the user in learning to evaluate the robot's vision field correctly and in turn supports efficient interaction with it.

Results concerning familiarity Familiarity is a principle of learnability, saying that the extent to which a user's knowledge and experience in other real-world or computer-based domains can be applied when interacting with a new system.

Imp 5: Allow flexibility in interaction: A robot should be flexible concerning the sequence of actions, the input modalities and the length of the feedback. Flexibility in the sequence of actions means that some steps of one sequence can be skipped if they are not necessary in the interaction. The input modalities for controlling a robot should also be kept as flexible as possible, meaning that the user can select one of several different input modalities. For example, when collaborating with a robot in a factory, switching between controlling the robot with verbal input or a remote control can be beneficial depending on the situation. Generally, verbal input is faster and does not require any devices (both hands are free). However, if there is a lot of ambient noise, controlling the robot via remote control can be more efficient. The length of the feedback given by the robot should be reduced to a minimum. Only feedback which is informative for the user should be given. Needless information / feedback should be avoided, as it is distracting and time-consuming for the user. The possibility to skip certain steps should be provided in order to shorten interaction times for more advanced users. When interacting with a new system, novice users often rely on the principle of trial and error. In case of error, the system has to react accordingly and give feedback about the error. In order to enable trial and error, it is necessary to let the user control the situation (and not the robot). Flexibility accounts to the principle of familiarity in so far, as the user can adapt the system's behavior so that it becomes familiar for him/her.

Imp 6: Provide distinct possibilities for controlling the robot:

Stopping the robot, cancelling a current task / action of the robot as well as putting the robot in stand-by state should be always provided. These actions are familiar for the user and are therefore easy to understand and learn. It is important to provide the possibility to stop or cancel wrong actions. The difference between stand-by and activation must be clear and unambiguous. Automatic stand-by after certain time is especially important due to safety reasons.

Results concerning generalizability Generalizability is a principle of learnability, saying that there should be support for the user to extend knowledge of specific interaction within and across applications to other similar situations.

Imp 7: Support easy commands: In order to support the principle of generalizability and in turn enhance learnability, commands have to be short, easy to remember, unambiguous and consistent. Therefore it is essential to use verbal cues and commands which are familiar and common for humans. For example, activating a robot by saying its name is intuitive for the human user, as this is also a common practice in human-human interaction. When controlling the robot verbally, possible interferences have to be taken into account. For example, efficient control of the robot is strongly dependent on the robot's position in relation to the users' position.

4 General interaction model

Based on the results and implications regarding usability (with a special focus on learnability) the expert evaluators developed a general interaction model for "speech-based human-robot interaction" and "learning by demonstration interaction". These general interaction models focused on the human input, the robot output, and the procedure on a very general level. Based on the general interaction model the interaction scenarios of the two case studies were re-designed and are presented in the following.

4.1 Re-designed HRP-2 interaction scenario

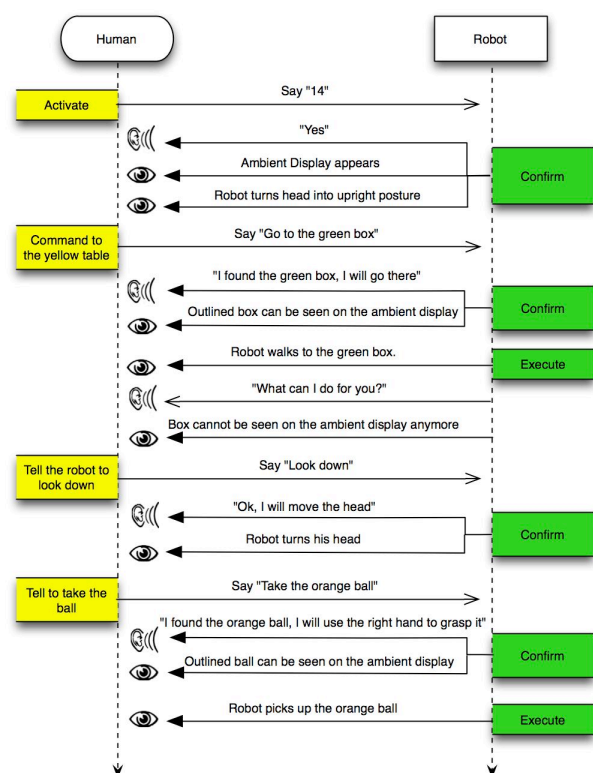


Figure 7. Interaction Model: HRP-2

Figure 7 shows the re-designed HRP-2 scenario. The tasks of the human are on the left side, the arrows towards the right represent the

commands of the human. On the right hand side the output (confirmation) and the actual reaction of the robot (execute) can be seen. The re-designed scenario indicates how major problems detected in the human-robot interaction can be solved: feedback and output modalities. In the interaction scenario of the preliminary field study the feedback was limited to audio output (voice) and the actual action (physical movement) of the robot.

Multimodal feedback is therefore advisable. The re-designed interaction scenario suggests a set of additional feedback modalities. As can be seen in the first confirmation of the robot in the diagram the body movement of the robot can be used as feedback on its status, easily allowing status detection. E.g. an inactive robot has its body crouched down. An additional possibility (not included in the sequence diagram) for giving feedback on the robot's status and its actions could be a light on the robot (eg. On the head or on the shoulders) that indicates the status of the robot. Four different statuses are suggested: (1) Stand-by, (2) ready for input/commands, (3) processing a command, and (4) executing the command.

Another way that is presented in the model diagram for enriching the feedback is by providing an ambient display. Through the display the vision field of the robot could be made traceable and thus comprehensible for the user. Its importance can for example be seen in the step where the participant needs to command the robot to look down. As he/she has just commanded the robot to the green box, it is illogical that the robot cannot detect the green box anymore, once being there. The ambient display giving information on the robot's field of vision helps to solve this issue. It shows the user that the robot cannot detect the green box at this point but the head needs to be turned downwards first. The ambient display can also show next actions or movements taken by the robot. This can increase predictability of the robot's action. In the re-designed model some of the robots auditory outputs were modified and reduced as the cognitive walkthrough revealed that not all information is necessary and reducing the amount of information output of the robot can ease learnability. Apart from feedback modalities, it is important to give the user the feeling of control. Hence, the options of stopping or canceling an ongoing action are crucial. As the sequence diagram portrays an ideal course of action this command is not included. However, they should be considered in the re-designed model.

4.2 Re-designed HOAP-3 interaction scenario

Figure 8 presents the re-designed HOAP-3 interaction scenario. As feedback modalities showed to be of highest importance during the cognitive walkthrough of the HOAP-3 scenario, the first step in re-designing the scenario was to rework the feedback model. The issues of consistent feedback upon all feedback channels (including body language, illuminated eyes, speech output, and actions) and multi-modal feedback were considered. First, it is important to stress that gestures, gaze direction, and speech output of the robot need to be consistent, e.g. the moment the person puts a box on the table the robot should turn its head towards it indicating that it detected the object. The speech output of HOAP-3 was reduced and clarified in the re-designed interaction model according to the suggestions of the experts that took part in the cognitive walk through. To help a novice user to successfully interact with the robot, feedback on the user's actions would be of help: Is the box put onto the correct position or in which ways can the robot's arm be moved (by demonstrating the different degrees of freedom). Such additional help should at the same time be easily skippable.

Apart from the feedback the re-designed interaction model offers

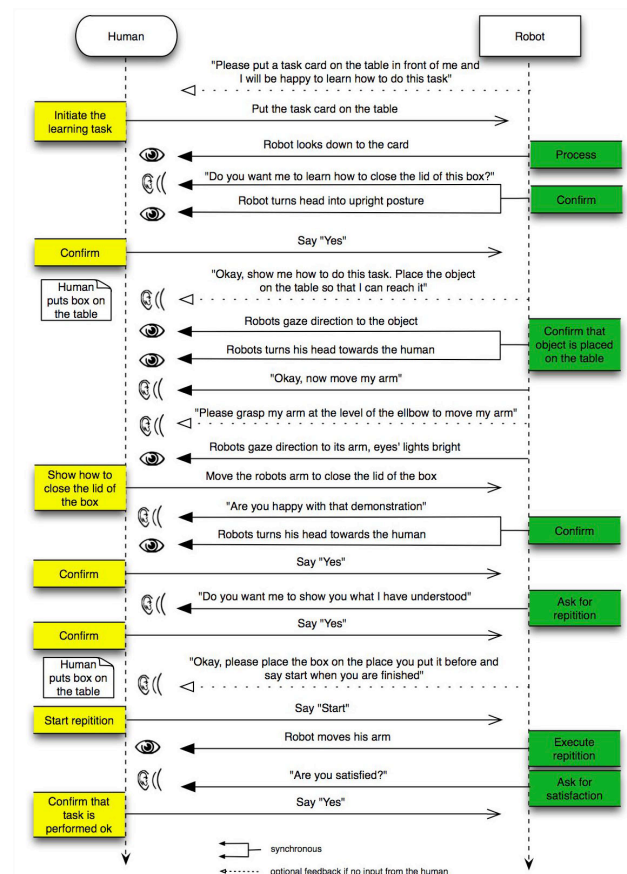


Figure 8. Interaction Model: HOAP-3

more flexibility to the user. First of all it is important that the user is given the possibility to start the interaction. That will avoid causing the feeling of the robot being in control of the situation. In the sequence diagram of the re-designed interaction model this is indicated by dotted lines. Only in case that no input/command is given by the user the robot will ask for a task card or give some information on what it requires the user to do. Giving the user the possibility to stop or cancel an ongoing action could even further increase flexibility. Therefore additional speech commands would be necessary which were not taken into account for the presented general interaction model.

5 Discussion and Outlook

This position paper reported about a methodological adaption of the cognitive walkthrough approach to allow the abstraction from specific HRI scenarios to more general HRI models in a specific application domain. Our main assumption was that the general models can serve as main narrative for other HRI interaction scenarios in order to reduce potential usability problems and increase intuitiveness and flexibility of the interaction. Two case studies were presented which described how a video-based cognitive walkthrough can be used to develop such general interaction models.

Generally speaking, the use videos of concrete human-robot interaction scenarios as input for expert evaluations to develop general interaction models seems to be a reasonable approach. As numerous severe usability issues could be identified by means of the adapted cognitive walkthrough and diluted in the subsequently developed interaction models, we consider the approach as effective. Thus, we suggest that applying this methodological adaption can be valuable for (1) improving human-robot interaction scenarios (e.g. for iterating the set-up of a user study before all participants are tested) and (2) developing general human-robot interaction scenarios as narratives to derive variations of interaction scenarios.

However, to verify our proposed method, comparative proof-of-concept user studies will be necessary to investigate if the scenarios based on the two presented interaction models actually offer less usability problems in the areas of intuitiveness and flexibility once in the field. For the future we plan to use the presented interaction models as source for the development of different, specific interaction scenarios to validate our proposed methodological approach. Moreover, it is desirable to apply the proposed cognitive walkthrough approach at an earlier stage of the development phase. Instead of video material of a first user study an animation video of the foreseen scenario could be used for the expert evaluation. Additionally, we want to gain insights on the question whether a user-based cognitive walkthrough even produces more robust interaction models compared to expert evaluations.

We conclude that our methodological adaption of the cognitive walkthrough represents a first step towards a beneficial approach to detect potential usability problems in human-robot interaction scenarios, especially at an early stage in the development phase, and thus is a valuable complement for human-robot interaction research. Even so, further research to proof evidence of the feasibility of this approach in terms of developing general interaction models will be indispensable.

ACKNOWLEDGEMENTS

We would like to thank all researchers supporting the expert evaluation. This work is supported in part within the European Commission as part of the Robot@CWE project, see also www.robot-at-cwe.eu.

REFERENCES

- [1] J. Buur, T. Binder, and E. Brandt, 'Taking video beyond 'hard data' in user centered design', in *Participatory Design Conference-PDC'00*, New York, NY, USA, (2000).
- [2] Alan J. Dix, Janet Finlay, and Gregory D. Abowd, *Human-computer interaction*, Pearson Prentice-Hall, Harlow [u.a.], 3. ed. edn., 2004.
- [3] C.D. Kidd. Sociable robots: the role of presence and task in human-robot interaction, 2003.
- [4] Wendy E. Mackay, Anne V. Ratzer, and Paul Janeczek, 'Video artifacts for design: Bridging the gap between abstraction and detail', in *In Proc. ACM Conference on Designing Interactive Systems, DIS 2000*, pp. 72–82. ACM Press, (2000).
- [5] Heather Martin and Bill Gaver, 'Beyond the snapshot from speculation to prototypes in audiophotography', in *DIS '00: Proceedings of the 3rd conference on Designing interactive systems*, pp. 55–65, New York, NY, USA, (2000). ACM.
- [6] Donald A. Norman, 'Cognitive engineering', in *User Centered System Design: New Perspectives on Human-Computer Interaction*, eds., Donald A. Norman and Stephen W. Draper, L. Erlbaum Associates Inc., Hillsdale, NJ, USA, (1986).
- [7] John Rieman, Marita Franzke, and David Redmiles, 'Usability evaluation with the cognitive walkthrough', in *CHI '95: Conference companion on Human factors in computing systems*, pp. 387–388, New York, NY, USA, (1995). ACM.
- [8] B. Robins, E. Ferrari, and K. Dautenhahn, 'Developing scenarios for robot assisted play', in *IEEE RO-MAN 2008: The 17th IEEE International Symposium on Robot and Human Interactive Communication*, 2008., pp. 180–186, (2008).
- [9] Mary Beth Rosson and John M. Carroll, 'Scenario-based usability engineering', in *DIS '02: Proceedings of the 4th conference on Designing interactive systems*, pp. 413–413, New York, NY, USA, (2002). ACM.
- [10] Jean Scholtz, 'Theory and evaluation of human robot interactions', in *HICSS '03: Proceedings of the 36th Annual Hawaii International Conference on System Sciences (HICSS'03) - Track 5*, p. 125.1, Washington, DC, USA, (2003). IEEE Computer Society.
- [11] Dag Sverre Syrdal, Nuno Otero, and Kerstin Dautenhahn, 'Video prototyping in human-robot interaction: results from a qualitative study', in *ECCE '08: Proceedings of the 15th European conference on Cognitive ergonomics*, pp. 1–8, New York, NY, USA, (2008). ACM.
- [12] Astrid Weiss, Regina Bernhaupt, Manfred Tscheligi, and Eiichi Yoshida, 'Addressing user experience and societal impact in a user study with a humanoid robot', in *AISB2009: Proceedings of the Symposium on New Frontiers in Human-Robot Interaction*, p. accepted for publication, (2009).
- [13] Astrid Weiss, Judith Igelsböck, Sylvain Calinon, Aude Billard, and Manfred Tscheligi, 'Teaching a humanoid: A user study on learning by demonstration with HOAP-3', in *Ro-Man2009: Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication*, pp. 147–152, (2009).
- [14] Cathleen Wharton, John Rieman, Clayton Lewis, and Peter Polson, 'The cognitive walkthrough method: a practitioner's guide', 105–140, (1994).
- [15] S.N. Woods, M.L. Walters, K.L. Koay, and K. Dautenhahn, 'Comparing human robot interaction scenarios using live and video based methods: towards a novel methodological approach', in *Proceedings of 9th IEEE International Workshop on Advanced Motion Control*, 2006., pp. 750–755, (2006).

A Robot as Persuasive Social Actor A Field Trial on Child-Robot Interaction

Astrid Weiss and Thomas Scherndl and Roland Buchner and Manfred Tscheligi¹

Abstract. This article presents an exploratory user study designed in order to explore the persuasive effect of the humanoid RS Media robot by means of an experiment on uncertain knowledge. In the study an overall of 171 children (99 male, 72 female) between the ages of 6 and 17 took part. The experiment was set up as a card game based on the theoretical “Monty Hall Problem” in which the children had to choose that card out of three that would disclose the treasure. During the game the robot advised the child to change his/her mind and choose another card. Three different pieces of advice could be given by the robot varying in their level of certainty. The basic assumption was that the more specific the hint is, the more credible and thus the more persuasive the robot will be experienced. The whole game was accompanied by a questionnaire in which the children had to state their level of certainty about their choice and the credibility of the robot on a rating scale between 0% and 100%. The results indicate a high persuasive effect on young children. As regards the hints it can be said that only easily understandable cues improve the persuasive effect.

1 INTRODUCTION

Child-Robot Interaction as a specialized research area of Human-Robot Interaction is paid more and more attention to in recent years. Robots can be seen as social interface for children in a variety of roles: companions, teachers, nannies etc. Several toy robots already exist which are commercially available to entertain children, like the robotics toy series of the WowWee robotics company [2]. However, entertainment will not be the exclusive future interaction scenario for Child-Robot Interaction. Research is already conducted with the aim to investigate the possibilities of robot therapy for children [9, 23, 25, 28]. Insights from these studies show the potential of robotic technology for children lies not only in entertainment, but also in supporting children’s development. It is likely that robots will be introduced in class rooms and children’s rooms in an educational role.

Based on the media equation, stating that media equal real life [30], we assumed that the social role of a teacher could be replaced by a robot. However, in that case it would matter to what extent children would follow the advice of a robot or trust in a robot. In this article we present an exploratory user study conducted with 171 children to explore the persuasive effect of the humanoid toy robot RS Media. We focused on how the detail level of the hint influences the persuasive power of the robot and the interplay between the perception of credibility and the child’s decision certainty.

Persuasive technology can be defined as technology especially designed to change beliefs, attitudes, or the behavior of another person or a group of people [12]. Key factors for successful persuasion are (1) the source of the message, (2) the strength of the argument, (3) and the influenced person being himself. According to Fogg computing systems which offer social cues and elicit social responses have a high persuasive power. Previous studies on the persuasiveness of robots have already shown promising results, indicating the high feasibility of robots as persuasive social actors [29, 15]. Based on this indication we decided to use the commercially available RS Media robot as interaction partner.

We assumed that in an uncertain situation in Child-Robot Interaction the level of detail of the argument and the child’s personal certainty concerning with the decision play a major role for the persuasive effect. In uncertain situations humans have to base their decisions on likelihood estimations, e.g. in the outcome of an election. These beliefs that lead to a decision are determined by influencing factors, like representativeness of the argument and availability of knowledge. To create a situation of uncertainty the experiment was set up as a game, based on the so called “Monty Hall Problem”.

Furthermore, according to the “Elaboration Likelihood Model” developed by Heesacker et al. [16], credibility is a major source factor for persuasive effects. This model states that people are more likely to be persuaded when the source of information is perceived as credible. This correlation of the information source and the persuasive effect could already been shown for screen agents [27]. In general, credibility can be seen as a key aspect for successful Child-Robot Interaction on the long term. Moreover, credibility plays a major role in supporting children’s decisions in situations of uncertainty.

The paper is structured as follows: a short survey on related work is presented in Sec. 2, the user study including the theoretical concept of the “Monty Hall Problem”, the research questions, study setting, and procedure are presented in Sec. 3. Sec. 4 discusses the results and insights gained, and research directions for future work will be given in Sec. 5.

2 RELATED WORK

Research on credibility has got a long history. Credibility has often been the object of research especially in the field of communication science, for instance the comparison of radio, TV, and newspaper [14] as well as recent research including online media [3],[20]. Hovland et. al [17] showed in their study that credibility consists of the two main principles *trustworthiness* and *expertise*. Based on this research Fogg and Tseng [13] investigated to what extent credibility matters in Human-Computer Interaction. Hereby they suggested a set of basic terms for assessing computer credibility. Bartneck [5] adapted Fogg’s and Tseng’s concept of credibility [13] to his model

¹ HCI&Usability Unit, ICT&S Center, University of Salzburg, Austria, email: firstname.lastname@sbg.ac.at

of convincingness. He could show that convincingness and trustworthiness highly correlate.

One of the most famous experiments regarding the credibility and persuasiveness of computing systems was conducted by Weizenbaum [37] with a virtual agent called ELIZA. This agent was able to simulate a psychotherapist and to keep the conversation going on by passively asking leading questions. At the end of the experiment, it was disclosed that people thought that they were talking to a human and that the conversational partner appreciated their problems. Throughout the study the subjects did not know that they were actually speaking to a computer program.

A similar study was conducted by Sundar and Nass [34] where people favored to interact with the computer over the human interaction partner. Studies which have been conducted lately, confronted their participants with questions regarding trustworthiness and believability of screen characters and fully embodied robots (see e.g. [5]; [31]).

In order to assess robots, scales were either adapted as done by Shinozawa, K., et al. [33] or newly developed [6]. Such scales were often used to assess the credibility of a source of information (TV, newspapers, web pages, etc., or an individual). Therefore, the source credibility is the amount of credibility attributed to a source of information by the receivers [8].

Shinozawa, K., et al. [33] could show that a 3D model of a robotic agent was rated higher in terms of source credibility by the means of McCroskey's credibility scale [24] than a 2D on-screen agent.

Kidd [19] also used a scale [7], actually developed for the media as well as McCroskey's one, to rate the robot's credibility. Hereby he found out that women tended to rate higher in terms of credibility than men. Additionally, people trusted robots that were physically present more, which is similar to the findings of [33]. The physical appearance of robots plays also a huge role in HRI. Powers and Kiesler [4] found out that certain physical attributes of the robot could change the human's mental model and that traits of the robot in that model are linked to people's intentions. For instance, certain physical characteristics, like the robot's voice and physiognomy, predicted impressions of a sociable robot, which in turn predicted the intention to take the advice of the robot and changed the people's perception of the robot's human likeness, knowledge and sociability.

Furthermore, it makes a difference if a robot recommends something or not. In their study Imai and Narumi [18] conducted a user study with a robot that gave recommendations. The participants in the experiment inclined to want what the robot recommended, e.g. this cup of tea is delicious and the other is not. Therefore, recommendation plays also an important role in HRI and in the credibility of robots. Additionally, credibility is tightly interwoven with trustworthiness and recommendation [13], whereas recommendation could be seen as a synonym for advice.

Another approach researched by Desai et al. [11] is to design an interface where the user can adjust the level of trust of a robot so it can decide more autonomously. For instance if the user controls a robot with a joystick and the robot notes many errors made by the user, so the robot suggests the user to switch to a higher level of trust. In such a level the robot could make more decisions without asking the human operator.

3 USER STUDY

The goal of the user study was to gain insights on the credibility and persuasive effect of the humanoid toy robot RS Media. The user study was set-up as a card game based on the "Monty Hall Problem", which will be described in more detail in the next section. A

preliminary study [36] with adults could already prove the suitability of the "Monty Hall Problem" to create a situation of uncertain knowledge and to investigate credibility and persuasive power. However, the study revealed that adults tried to solve the "Monty Hall Problem" by probability reasoning which influenced the study setting negatively. We assumed that conducting the same study with children should limit this effect. The study was conducted in a big shopping mall during an "open University day" in Salzburg, Austria, in November 2008.

3.1 Theoretical Basis

The user study was based on the famous paradox called "Monty Hall Problem", also known as "Three Door Problem" or "Car Goat Problem". The roots of the name "Monty Hall Problem" go back to the 80's game show, "Let's make a deal" moderated by Monty Hall. In that show the candidate had the chance to win a car but he/she had to choose one door from three given ones. After the candidate had chosen the door, the show moderator opened one of the two doors left, which always contained a goat. Then he asked the candidate if he/she wants to change his/her decision and take the other door. Years later, the problem reached a wider publicity when Marilyn vos Savant published a puzzle in her *Parade*[1] magazine column as an answer to the following question of a reader:

Suppose you are on a game show, and you are given a choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say number one, and the host, who knows what is behind the doors, opens another door, say number three, which has a goat. He says to you, "Do you want to pick door number two?" Is it to your advantage to switch your choice of doors?

The answer to this question is, yes. You should switch because the chance to win the car in the first choice is only $1/3$. After unraveling one goat, not the candidate's door, the chance to win, if you stay, is still $1/3$ (see Figure 1). However, if the candidate changes his/her opinion at this point of the game, the probability to win would be twice as high ($2/3$) as if the candidate would stay with his/her initial decision. This kind of logic is very difficult and only a few people realize it.

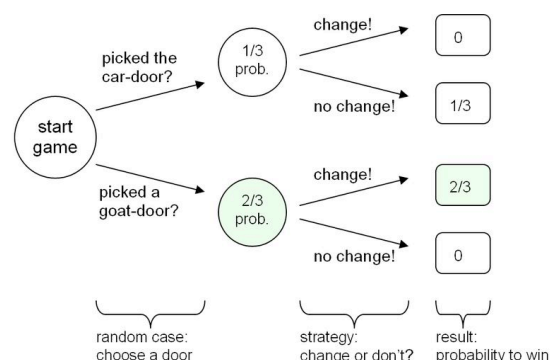


Figure 1. The "Monty Hall Problem"

The majority of the people would not take the other card due to the fact that they are thinking the probability of their first decision

increased to 1/2. So, if the goods and odds are equal, why should I change my primary choice? In the end these people are acting intuitively wrong and even people with extensive statistical skills can struggle [26]. Therefore, this kind of probability problem can be used to study how to deal with uncertainty [22].

The study described in the following was inspired by Komatsu et al. [21], where people estimated the robot's abilities by acquiring how accurately the robot can detect a gold coin under three mountains in an imaginary video game.

3.2 Study set-up

The study was set up as a card game where the participating children should play together with the humanoid toy robot RS Media [2], see Figure 2. To test our assumption that the detail level of the hint influences the credibility and persuasive power of the robot, three experimental conditions were set up:

1. **Recommendation without hint** → The robot says: "I would take the other card."
2. **Recommendation with hint** → The robot says: "I would take the other card, because the chance to win is higher."
3. **Recommendation with explicit hint** → The robot says: "I would take the other card, because the chance to win is 66%."



Figure 2. Child interacting with the RS Media robot

During the user study, the RS Media robot was placed on a table and in front of it sat the child, see Figure 2. The built-in loudspeakers of the robot were muted because the ambient noise in the shopping mall environment was very high. Therefore, external loudspeakers

were placed on the table with the robot and plugged into the line-out of a MP3-Player. The RS Media was still able to react to the environment, e.g. by rotating its head to the child in front of it, which is an important behavior when interacting socially [38]. A Wizard-of-Oz approach [10] was chosen for the audio output of the robot, in order to adapt the speed of the user study to the child's speed of decision making. When the experiment reached the point, when the robot's hint is requested, the experimenter pressed a button on the MP3-Player which triggered the robot's output depending on the experimental condition. The MP3-Player was put under the table where the subjects could not see it. During the study the experimenter changed the conditions after about ten to fifteen participants. In the end there was a distribution of 52 children with condition 1, 69 with condition 2 and 49 with condition 3. The experiment was intentionally set-up as between subject-design to eliminate a learning effect in terms of the "Monty Hall Problem".

3.3 Research questions

As our main research interest was to explore how the detail level of the hint influences the persuasive power of the robot and the interplay between the perception of credibility and the child's decision certainty as source for persuasion, the following main research questions were set up for the user study:

1. (RQ1): Do children follow the advice of the RS Media robot?
2. (RQ2): Which factors influence the credibility of the RS Media robot?

Based on these research questions the following hypotheses were supposed:

- H1: The experimental condition leads to a difference in the frequency of opinion change.
- H2: There is a difference in credibility depending on the experimental condition.
- H3: Credibility increases once the robot's prediction was correct.
- H4: There is an interaction between the child's opinion change and decision certainty.

3.3.1 Procedure

At the beginning of the user study the experimenter showed the child the three cards, one of which hiding the treasure (see Figure 3) and explained the game. Each card had the same back side and it was not possible to look trough or under it, in other words cheating was not possible. As the child was ready, the experimenter mixed the cards and told him/her to choose one card under which he/she supposed the treasure. After that the child was asked how certain he/she felt about the decision. Then the experimenter took a look below the other two cards, laid one blank card open and asked again how certain he/she felt about the choice. Now, the robot entered the game. The experimenter asked the robot, if the child should change his/her mind and go for the other card. Then the robot answered depending on the experimental condition. At this point the child had the possibility to switch or stay with his/her choice. Again the child was asked how certain he/she felt about the decision after the advice of the robot, if he/she believed in the robot's advice and if the advice was helpful. In the end the experimenter disclosed the secret and explained the "Monty Hall Problem".



Figure 3. The Card Game

3.4 Instruments

As beliefs in uncertain situations are often expressed in numbers [35], the whole user study was accompanied by a questionnaire containing questions on the level of certainty of the choice (which had to be rated on a scale from 0% until 100% certainty, see Figure 4) and the credibility of the robot. According to [32], children at the age of seven are able to use such scales. For children older than 5.6 years using scales was the best predictor 84% of accuracy.

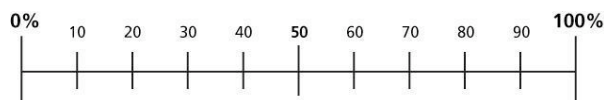


Figure 4. The Rating Scale for Certainty and Credibility

4 RESULTS

171 children took part in the study. Some of them knew the robot, as they have seen it before in toy stores or on television in an advertisement for the “open University day”. However, none of the children ever before played with the robot. 66 of the participating children were between 6 and 10 years old, 60 between 11 and 12 years, and 45 were older than 13 years. In 96 of the cases the robot gave the right advice in 73 the wrong one.

4.1 Did the children follow the advice of the robot (RQ1)

Regarding the question if the children followed the advice of the robot, 46.2% of the children did not take the advice and 53.8% of the children changed their decision based on the advice of the robot. The detail level of the advice influences the chance of children to switch to the other card (H1). However, an administered χ^2 test could not show a difference between the probability of change between the three different experimental conditions ($\chi^2(2) = 1.73, p > .05$). Nevertheless, there was a tendency that a reason for the given advice increased the probability of change. If the robot gave no reason

why to change to the other card (condition 1), there was no opinion change by trend, as can be seen in Table 1 (see count vs expected count). If the reason was easy to comprehend, the child will more likely change his/her opinion. Interestingly the effect did not occur for the third experimental condition with the highest detail level of advice. We assume that this hint was hard to comprehend for the children and thus the effect did not occur.

Table 1. Cross table: age, experimental condition, and opinion change

Age	Con		No Change	Change	Total
6-10	Con1	Count	5.0	10.0	15.0
		Expected Count	4.5	10.5	15.0
	Con2	Count	6.0	17.0	23.0
		Expected Count	7.0	16.0	23.0
	Con3	Count	9.0	19.0	28.0
		Expected Count	8.5	19.5	28.0
11-12	Con1	Count	8.0	6.0	14.0
		Expected Count	7.6	6.4	14.0
	Con2	Count	13.0	13.0	26.0
		Expected Count	14.1	11.9	26.0
	Con3	Count	11.0	8.0	19.0
		Expected Count	10.3	8.7	19.0
13+	Con1	Count	15.0	8.0	23.0
		Expected Count	13.8	9.2	23.0
	Con2	Count	10.0	10.0	20.0
		Expected Count	12.0	8.0	20.0
	Con3	Count	2.0	0.0	2.0
		Expected Count	1.2	0.8	2.0

A strong effect could be observed regarding age groups. Younger children were significantly more often persuaded to change their previous decision, whereas older children (13+) did change their decision not as often as expected ($\chi^2(2) = 11.39, p < .01$).

4.2 Factors influencing the credibility of the RS Media robot (RQ2)

The age group had an influence on the credibility of the RS Media ($F(2, 159) = 5.29, p < .01, \eta^2 = .06$). Age group1 children had a higher belief in the robot (age group1 mean: 59.77, SD: 32.00; age group2 mean: 44.49, SD: 30.10; age group3 mean: 40.68, SD: 31.52). There was a significant interaction between age group and the prediction correctness ($F(2, 159) = 4.29, p < .05, \eta^2 = .05$). As can be seen in Figure 5, the younger the children were, the higher they rated the initial credibility of the robot. When the prediction was correct, the credibility of the robot was rated higher in the final assessment by all age groups. When the prediction was incorrect, the credibility was rated lower in the final assessment by all age groups, but the drop of credibility was strongest in the youngest age group. Gender and education did not significantly influence the credibility assessment.

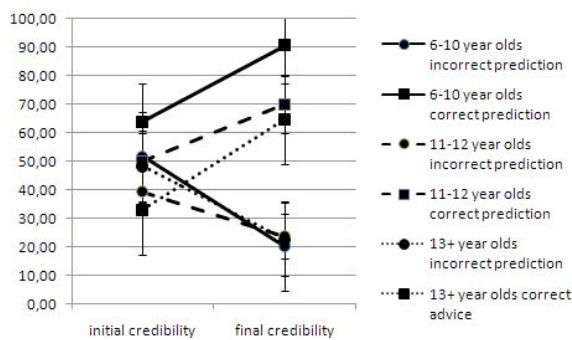


Figure 5. Influence of Age and Prediction Correctness on Initial and Final Credibility

4.2.1 Increase in credibility if the robot was right

In general, there was no difference between the conditions concerning final credibility of the robot ($F(2, 158) = 1.07, p > .05, \eta^2 = .01$). However, after considering the effect of an incorrect advice of the robot, the credibility dropped substantially indicating a main effect of the correctness of advice ($F(1, 158) = 179.08, p < .01, \eta^2 = .53$). In other words, the final credibility was mainly dependent on the correctness of the prediction and not so much on the level of detail of the advice. Moreover, an increase in credibility during the game could be observed, when the robot gave a correct prediction, whereas if it gave a wrong prediction the credibility substantially decreased ($F(1, 163) = 89.10, p < .01, \eta^2 = .38$) (see Figure 6).

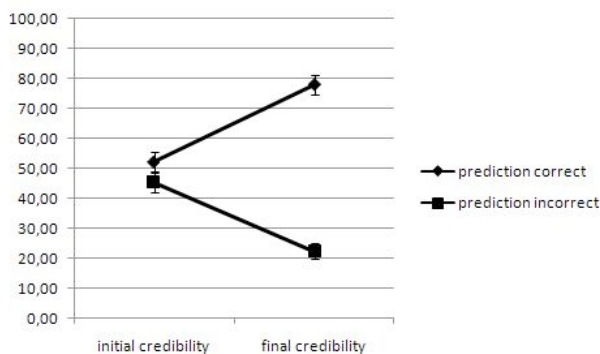


Figure 6. Interaction between Credibility and Prediction Correctness

4.2.2 Difference of credibility depending on the level of detail of the advice (H2)

When considering the correctness of the robot's prediction in the analysis, we found a significant interaction between the experimental condition and the correctness of the precondition ($F(2, 158) = 3.52, p < .05, \eta^2 = .04$). Children who received an advice with the

highest level of detail (condition 3) rated the final credibility especially low, if the advice turned out to be incorrect.

4.2.3 More detailed - more helpful (H3)

The helpfulness of all hints was perceived very heterogeneously as the broad standard deviations for the mean ratings in all three conditions show (mean: 35.58, SD: 34.86, mean: 45.65, SD: 32.70, mean: 52.31, SD: 35.99).

However, there was a significant difference of the perceived helpfulness between the three conditions ($F(2, 166) = 3.04, p = .05, \eta^2 = .04$). The condition with no reason given was rated less helpful than the condition with the most specific hint ($p < .05$). Interestingly this finding could be partial due to the fact that the factor of age was strongly influencing the perception of helpfulness, as the three age groups were not equally distributed in the experimental conditions (see Table 2). General, younger children assessed the helpfulness of all hints higher than older children, whereas gender and education had no effect.

Table 2. Mean helpfulness of advice

Con	Age	Mean	SD	N
1	6-10	49,00	38,79	15
	11-12	37,50	30,43	14
	13+	25,65	32,94	23
	Total	35,58	34,86	52
2	6-10	61,09	28,12	23
	11-12	39,04	29,09	26
	13+	36,50	36,89	20
	Total	45,65	32,70	69
3	6-10	65,78	36,73	27
	11-12	35,53	28,23	19
	13+	30,00	14,14	2
	Total	52,31	35,99	48
Total	6-10	60,25	34,52	65
	11-12	37,54	28,67	59
	13+	30,67	34,11	45
Total		44,44	34,74	169

4.2.4 Opinion change and level of certainty (H4)

The children were asked about the level of certainty with respect to their choice three times during the game. It was found that the children's perceived certainty was influenced by the time of the game play and the opinion change. Children who decided to change after the robot's advice were more insecure in the beginning than children, who did not change their preference, but the discrepancy between the children vanished after the robot gave its advice ($F(2, 336) = 19.32, p < .01, \eta^2 = .10$) (see Figure 7).

5 Conclusion

The goal of the exploratory user study was to gain insights on the credibility and persuasive effect of the humanoid toy robot RS Media. Our basic assumption was that the higher the detail level of the advice, the higher the persuasive power of the robot would be in a situation of uncertainty. In other words we expected that children participating in the user study under the experimental condition 3 (the robot says: "I would take the other card, because the chance to win is 66%") would result in a high credibility rating and opinion change rate. However, we found out that this was not the case

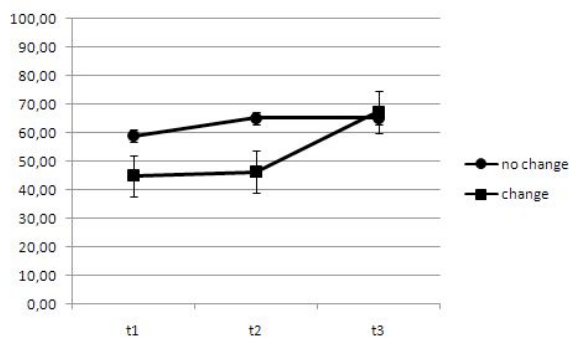


Figure 7. Interaction of Opinion Change and Level of Certainty

for the children. Although the third condition hint indeed included the highest amount of information, the formulation was probably not child-adequate.

We assume that the younger children could not fully grasp the knowledge the robot made available. Therefore they did not produce a higher opinion change rate than condition 2 (the robot says: “I would take the other card, because the chance to win is higher”). This assumption needs to be tested again with a homogeneous sample to exclude age effects.

It is interesting to note that children, who were unsure in the specific situation, gained a higher level of confidence through the advice of the RS Media robot. This finding raises the suggestion that robots could be incorporated in situations which are complex and lead to uncertainty. Furthermore, we could show that younger children were more often persuaded by the RS Media robot. A direction for future research is to test, if the persuasive effect is even higher for pre-school and kindergarten children. However, one has to consider the side-effect that as soon as the prediction of the robot was wrong, the credibility dropped immensely, in other words: “a liar is not believed even if he speaks the truth”. These findings should be taken into account, if robots were integrated into the education of primary school children. For a purposeful integration, it will be necessary to consider two ethical guidelines for persuasive technology: (1) Persuasion should be based on prior (real or counterfactual) consent - The teacher should not use the persuasive robot to persuade children of something he/she him/herself would not consent to be persuaded to do. (2) Persuasion should, if possible, favor factual feedback over other means of persuasion - The content that the persuasive robot is teaching the child, should be objective and demonstrated by reproducible facts. Moreover, replacing human teachers by artificial persuasive robots could lead to so called “post social relationships”, in which children tend to prefer being engaged in asymmetric, conflict-free artificial interaction than putting effort in dealing with humans and interpret their actions.

In general this study showed that the RS Media robot can serve as persuasive actor and thereby supports the media equation: media equal real life [30]. This means that persuasive effects, which would have occurred when a human advised a child, also occurred when the human was replaced by the robot. However, for future research it will be relevant to identify if the same persuasive effect can be revealed, if the robot replaces the role of a human adult or the role of a human child. To increase the validity of the experimental results, a test condition, without any robot replacement, needs to be introduced.

Moreover, possible differences between screen agents and embodied robots need to be examined. Furthermore, the variability of shapes (antropomorphic, zoopomorphic, and cartoon-like) must be taking into account. Even the sound of the voice and if it matches the robot’s embodiment could have an influence on the robot’s persuasive power and should be therefore experimentally controlled.

A further aspect is the choice of the game. All children told us that they did not know the “Monty Hall problem”, when asked at the end of the study. However we can not rule out the possibility that the children assumed that the robot would trick them into choosing the wrong card. This may be especially the case as finding the treasure underneath a card is known as a game of con artists. Research in these areas will help us to derive more concrete design implications to build more persuasive robots. However, the most important implication of design for this case study is that the RS Media robot can serve as persuasive actor and that the persuasive power lies in the comprehension to the message and is not a matter of detail level.

ACKNOWLEDGEMENTS

The authors would like to thank all participating children.

REFERENCES

- [1] Parade Magazin. <http://www.parade.com/>. last time retrived: February 2009.
- [2] Wowwee robotics. <http://www.wowwee.com/>. last time retrived: January 2009.
- [3] R. A. Abdulla, B. Garrison, M. Salwen, P. Driscoll, and D. Casey, ‘The Credibility of Newspapers, Television News and Online News’, in *Journalism and Mass Communication*, (2002).
- [4] A. Powers and S. Kiesler, ‘The advisor robot: tracing people’s mental model from a robot’s physical attributes’, in *HRI ’06: Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pp. 218–225, New York, NY, USA, (2006). ACM.
- [5] C. Bartneck, ‘Affective expressions of machines’, in *CHI ’01: CHI ’01 extended abstracts on Human factors in computing systems*, pp. 189–190, New York, NY, USA, (2001). ACM.
- [6] C. Bartneck, E. Croft, and D. Kulic, ‘Measuring the anthropomorphism, animacy, likeability, perceived intelligence and perceived safety of robots’, in *Metrics for Human-Robot Interaction Workshop in affiliation with the 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI 2008), Technical Report 471*, pp. 37–44, Amsterdam, (2008). University of Hertfordshire.
- [7] D. K. Berlo, J. B. Lemert, and R. J. Mertz, ‘Dimensions for evaluating the acceptability of message sources.’, *Public Opinion Quarterly*, **46**, 563–576, (1969).
- [8] C. C. Bracken, ‘Perceived source credibility of local television news: The impact of television form and presence’, *Journal of Broadcasting & Electronic Media*, **50**, 723–741, (2006).
- [9] A. M. Cook, M. Q. Meng, J. J. Gu, and K. Howery. Development of a robotic device for facilitating learning by children who have severe disabilities. In *Proceedings of the IEEE Transactions on Neural Systems and Rehabilitation Engineering*, pages 178–187, Edmonton, Canada, (2002).
- [10] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg, ‘Wizard of oz studies: why and how’, in *IUI ’93: Proceedings of the 1st International Conference on Intelligent User Interfaces*, pp. 193–200, New York, NY, USA, (1993). ACM.
- [11] Munjal Desai, Kristen Stubbs, Aaron Steinfeld, and Holly Yanco, ‘Creating trustworthy robots: Lessons and inspirations from automated systems’, in *Proceedings of the AISB Convention: New Frontiers in Human-Robot Interaction*, (May 2009).
- [12] B. Fogg, *Persuasive Technology*, Morgan Kaufmann, San Diego, 2002.
- [13] B. J. Fogg and H. Tseng, ‘The elements of computer credibility’, in *CHI ’99: Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 80–87, New York, NY, USA, (1999). ACM.
- [14] C. Gaziano and K. McGrath, ‘Measuring the concept of credibility.’, *Journalism Quarterly*, **63**(3), 451–462, (Aut 1986).

- [15] Jaap Ham and Cees Midden, 'A robot that says bad!: using negative and positive social feedback from a robotic agent to save energy', in *HRI '09: Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pp. 265–266, New York, NY, USA, (2009). ACM.
- [16] M. Heesacker and R.E. Petty, 'Field dependence and attitude change: Source credibility can alter persuasion by affecting messagerelevant thinking', *Journal of Personality and Social Psychology*, **51**, 653–666, (1983).
- [17] C. I. Hovland, I. L. Janis, and H. H. Kelly, 'Communication and Persuasion', *New Haven, Conn.: Yale University Press*, (1953).
- [18] M. Imai and M. Narumi, 'Immersion in interaction based on physical world object', in *The 2005 International Conference on Active Medial Technology (AMT2005)*, pp. 523–528, (5 2005).
- [19] C. D. Kidd, *Sociable Robots: The Role of Presence and Task in Human-Robot Interaction*, Master's thesis, Massachusetts Institute of Technology, 2003.
- [20] A. Kohut, S. Keeter, C. Doherty, M. Dimock, K. Parker, J. M. Horowitz, R. Suls, S. Neidorf, L. Christian, J. Kiley, K. Holzward, A. Tyson, and J. Albrittain. Key News Audiences Now Blend Online and Traditional Sources. <http://people-press.org/reports/pdf/444.pdf>, August 2008. last time retrived: February 2009.
- [21] T. Komatsu and M. Nambu, 'Effects of the agents appearance on peoples estimations about the agents abilities: Questionnaire investigations for liberal arts and informatics students', in *Proceeeding of the 17th IEEE International Symposium on Robot and Human Interactive Communication*, pp. 142–147., (2008).
- [22] S. Krauss and X. T. Wang, 'The psychology of the monty hall problem: discovering psychological mechanisms for solving a tenacious brain teaser', *J Exp Psychol Gen*, **132**, 3–22, (Mar 2003).
- [23] C. Lathan, A. Brisben, and C. Safos. Cosmobot levels the playing field for disabled children. *Interactions*, 12(2):14–16, (2005)
- [24] J. C. McCroskey, P. R. Hamilton, and A. N. Weiner, 'The effect of interaction behavior on source credibility, homophily, and interpersonal attraction', *Human Communication Research*. v1, 42–52, (1974).
- [25] F. Michaud, J.-F. Laplante, H. Larouche, A. Duquette, S. Caron, D. Letourneau, and P. Masson. Autonomous spherical mobile robot for child-development studies. *IEEE Transactions on Systems, Man, and Cybernetics*, 35:471– 480, (2005)
- [26] P. R. Mueser and D. Granberg, 'The monty hall dilemma revisited: Understanding the interaction of problem definition and decision making', *Experimental 9906001*, EconWPA, (June 1999).
- [27] Christoph Obermair, Wolfgang Reitberger, Alexander Meschtscherjakov, Michael Lankes, and Manfred Tscheligi, 'perframes: Persuasive picture frames for proper posture', in *PERSUASIVE2008: Proceedings of the 3rd International Conference on Persuasive Technology*, pp. 128–139, Berlin, Heidelberg, (2008). Springer-Verlag.
- [28] B. Prazak, A. Hochgatterer, G. Kronreif, and M. Furst. Robot supported play - New possibilities for physically handicapped children?! In *Proceedings of the Association for the Advancement of Assistive Technology in Europe*, Dublin, Ireland, (2003)
- [29] Kohei Ogawa, Christoph Bartneck, Daisuke Sakamoto, Takayuki Kanda, T. Ono, and Hiroshi Ishiguro, 'Can an android persuade you?', in *18th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN2009*, pp. 553–557, Toyama, (2009). IEEE.
- [30] B. Reeves and C. Nass, *The media equation: How people treat computers, televisions, and new media like real people and places*, Cambridge University Press, New York, 1996.
- [31] J. Reichenbach, C. Bartneck, and J. Carpenter, 'Well done, robot! - the importance of praise and presence in human-robot collaboration', in *RO-MAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication*, ed., K. Dautenhahn, pp. 86–90, Hatfield, (2006). IEEE.
- [32] B. J. Shields, T. M. Palermo, J. D. Powers, S. D. Grewe, and G. A. Smith, 'Predictors of a child's ability to use a visual analogue scale', *Child: Care, Health and Development*, **29**(4), 281–290, (2003).
- [33] K. Shinozawa, B. Reeves, K. Wise, S. Lim, and H. Maldonado, 'Robots as new media: A cross-cultural examination of social and cognitive responses to robotic and on-screen agents', *International Communication Association*, (2002).
- [34] Shyam S. Sundar and Clifford Nass, 'Source orientation in human-computer interaction', in *Communication Research*, volume 27, pp. 683–703. Sage Journals, (2000).
- [35] Amos Tversky and Daniel Kahneman, 'Judgment under uncertainty: Heuristics and biases', *Science, New Series*, **185**(4157), 1124–1131, (1974).
- [36] A. Weiss, R. Buchner, T. Scherndl, and M. Tscheligi, 'I would choose the other card: humanoid robot gives an advice', in *HRI '09: Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pp. 259–260, New York, NY, USA, (2009). ACM.
- [37] J. Weizenbaum, 'Eliza—a computer program for the study of natural language communication between man and machine', *Commun. ACM*, **9**(1), 36–45, (1966).
- [38] F. Yamaoka, T. Kanda, H. Ishiguro, and N. Hagita, 'Lifelike behavior of communication robots based on developmental psychology findings', in *Humanoids 2005*, pp. 406–411, (2005).