ICT-2009-248730

Florence

Multi Purpose Mobile Robot for Ambient Assisted Living

STREP
Contract Nr: 248730

Deliverable: D4.1 State of the Art in human-robotic interaction

Due date of deliverable: (2010-07-31)
Actual submission date: (2010-07-30)

Start date of Project: 01 February 2010 Duration: 36 months

Responsible WP: OFFIS

Revision: final

Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013)		
Dissemination level		
PU	Public	x
PP	Restricted to other programme participants (including the Commission Service	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (excluding the Commission Services)	

0 DOCUMENT INFO

0.1 Author

Author	Company	E-mail
Melvin Isken	OFFIS	Melvin.isken@offis.de
Björn Vester	OFFIS	Bjoern.vester@offis.de
Florian Winkler	NEC	Florian.Winkler@neclab.eu
Leszek Holdenderski	Philips	leszek.holenderski@philips.com
Dietwig Lowet	Philips	Diewig.lowet@philips.com
Mortaza Bargh	Novay	Mortaza.Bargh@novay.nl
Leire Martinez	Fatronik	lmartinez@fatronik.com

0.2 Documents history

Document version #	Date	Change
V0.1	2010-03-01	Starting version, template
V0.2	2010-03-15	Definition of ToC
V0.3	2010-05-24	First complete draft
V0.4	2010-06-27	Integrated version (send to WP members)
V0.5	2010-07-01	Updated version (send PCP)
V0.6	2010-07-05	Updated version (send to project internal reviewers)
Sign off	2010-07-22	Signed off version (for approval to PMT members)
V1.0	2010-07-29	Approved Version to be submitted to EU

0.3 Document data

Keywords	
Editor Address data	Name: Melvin Isken Partner: OFFIS e.V. Address: Escherweg 2 Phone: +49-441-9722-229 Fax: +49-441-9722-111 E-mail: melvin.isken@offis.de
Delivery date	2010-07-30

0.4 Distribution list

Date	Issue	E-mailer
	Consortium members	al_florence_all@natlab.research.philips.com
	Project Officer	Luiz.Santos@ec.europa.eu
	EC Archive	INFSO-ICT-248730@ec.europa.eu

Table of Contents

1 List of figures

2 Introduction

2.1 Objective

This document represents the State of the Art in human robotic interaction. The objective of the work package, where this deliverable belongs to, is to design and implement the Florence Human-Machine-Interface (HMI) for direct and remote communication. The robot will be the connecting element between several AAL services, home infrastructure and the user. It is the single interface for all. Therefore a novel kind of dialog management will assure a consistent way of communication between user and robot independent of the service or technical system behind. Because of the fact that there are high barriers especially for elderly to get in touch with technical systems and to ensure that the robot will be accepted by elderly people, the results of the user studies in work package WP1 will be used as requirements, so the interaction design of the user with the robot will take into account the usability requirements of elderly people.
The goal is to create a component-based architecture for the HMI corresponding to known and proven robot design approaches for service robots. This document tries to cover most of the State of the Art in this design approaches and general human robotic interfaces.

2.2 Scope

As the project´s overall goal is to develop robotic services for elderly, this State of the Art may be more oriented to services usable for elderly (but is not limited to). The State of the Art for human machine interaction and human robotic interaction is rather huge so it is not possible to cover all work that has been done in these areas. This document will show some general interaction techniques and have a closer look on the integration of elderly.

2.3 Outline

This document is structured the following way:
The first part deals with general interaction modalities and combinations of it. It is followed by a section regarding the dialogue management. This section goes more into detail since this topic will be a major aspect of this work package. The document ends with a section about the appearance of robotic systems and how this is related to the acceptance by the user. This will give some advices for the user tests in work package WP1, which can use these results for the outline of their evaluations.
References to web pages are marked as footnotes, references to papers use the standard citation style.

3 Human robotic interaction / Feedback mechanisms

This chapter assembles the State of the Art in human robotic interaction and feedback mechanisms. The different kinds of interaction and feedback components can be divided by their modalities. Information about the State of the Art for each of the different modalities is presented in the next subsections.

The design of human-robot interfaces has been subject to extensive research since the beginning of robot engineering. This is due to the fact that the design of such an interface can make the difference between a robot being perceived as a useful, reliable and powerful assistant to a human being and utter rejection.

As pointed out in [Adams2002], the development of effective, efficient and usable interfaces requires the inclusion of the user's perspective throughout the entire design and development process. In [Raskin2000] the author stresses that human-robot interfaces should be humane, which he defines the following way:

"An interface is humane, if it is responsive to human needs and considerate of human frailties" [Raskin2000, pg. 6]

For example, according to [HRIWI2010] when users encounter proactive behavior on the part of the robot and the robot does not respect a safety distance, penetrating the user space, he or she might express fear. This is dependent on one person to another. Only intensive experiment can permit a more precise model. It has been shown that when a robot has no particular use, negative feelings are often expressed. The robot is perceived as useless and its presence becomes annoying.

Another interesting fact is that people tend to attribute personality characteristics to a robot that were not implemented. This is something which needs to be checked early and by extensive user tests.

There are two types of human-robot interfaces which need to be considered in Florence. There will be a direct interface between a human and the physical robot, e.g. when an elderly person is sitting in front of the robot and interacting with it. This interface might be consisting of voice control, gesture recognition, touch screens, proximity sensors, etc. or any combination of such technologies. The second interface is a remote interface to the robot that allows a robot operator or a care person to remotely control the robot in order to assist, supervise or examine an elderly person from a remote office. This kind of interface poses different challenges to the developers.
In [Kadous2006] the authors elaborate on these aspects while in [Olsen2003] metrics are presented that can be used to evaluate human-robot interactions in a more general way.

Yanco and Drury [Yanco2002] [Yanco2004] defined eight interaction possibilities for human(s)-robot(s) interaction. In the following diagram H stands for human, R stands for Robot and double headed arrows indicate command flows.

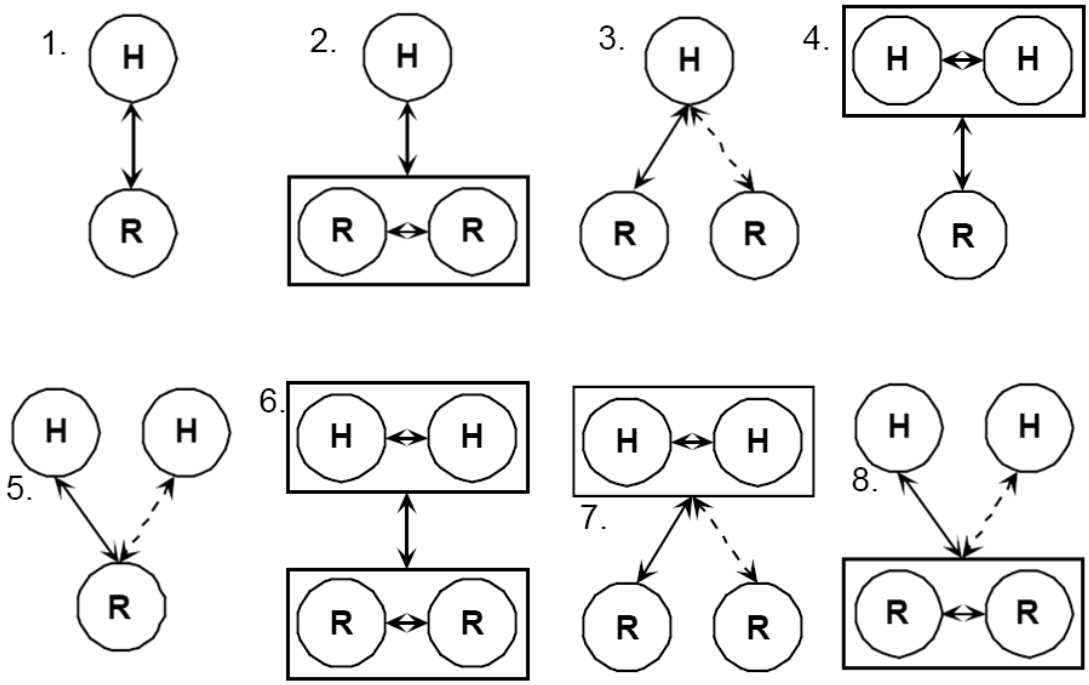


Figure 1: Interaction possibilities for human-robot interaction ([Yanco2002])

1. One human controls one robot.
2. One human controls a group of robots, issuing one command that the robots coordinate among themselves to fulfill the task.
3. One human controls multiple individual robots, issuing multiple individual commands to robots that operate independently.
4. Humans agree on robot commands and issue one coordinated command to a single robot.
5. Humans issue different commands to a single robot that the robot must resolve conflict and/or prioritize.
6. A team of humans issue a command to a team of robots. The robots coordinate to determine which robot(s) performs which portion(s) of the command.
7. A team of humans issues one command per individual robot.
8. Individual humans issue different commands to a team of robots, which the robots must resolve conflict and/or prioritize and divide among themselves.

In the Florence system interactive situation 1 and 5 will mostly take place. For collaborative service development human-robot-human interaction is studied, where the aim is to make the interaction with the robot so intuitive that humans will only need to worry about interacting with the other human to emphasize the collaborative aspect of the activity.

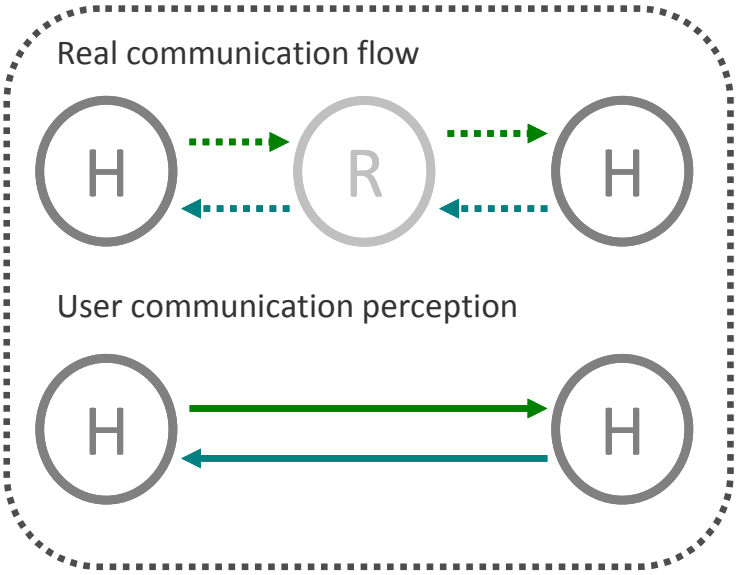


Figure 2: Communication flow and users perception

3.1 Visual

This subsection describes different methods of visual interaction with a user. This includes

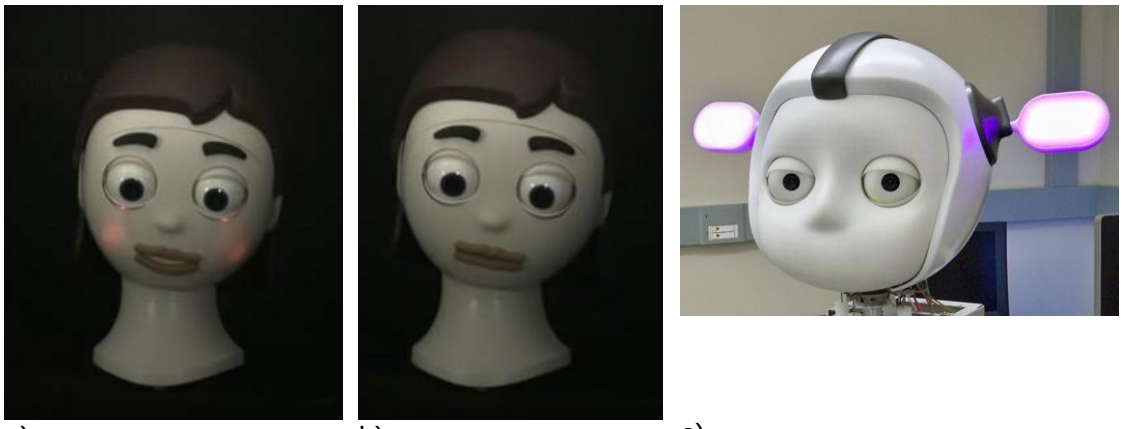
- Displays
- Lights and lighting techniques (indicators, signals)
- Projection systems

Visual interaction is the first modality that is considered at the design of a system. The simplest way is a small light that indicates whether the system is turned on or not. Very often different modes of operation of a system are presented via status lights. So the user is able to quickly identify the status of the system. In terms of robotics, simple on-and-off-lights are insufficient. Modern robots can have a lot of different states and are able to communicate with the user via optical modalities.
To present a suitable set of information computer displays are a common technique. If one has a look at the State of the Art in service robotics (see Florence Deliverable D5.1) nearly all of them carry some kind of display to interact with the user. For the system designer the use of a display has many advantages as it can display a variety of information. These can be in the form of text, simple signs, symbols or just colors.

An extension of displays is video projectors. Projectors extend the size of the display carried by the robot by magnitudes. A robot can be used to project information to special places like a text explanation besides an image on the wall, show the correct way via arrows on the ceiling, walls or ground or simply project complete videos. Because of the size of the projection, it is often better readable and more people can get the information at a time.
A problem with projecting information is the need of projection screens. In most cases the projection screen has to be of bright material, the distance has to be known to

focus on it and the projector has to be in a straight line to the projection area to avoid deformation. Another point is a direct line of sight that is also needed.

Multicolored lights can be used to encourage emotional expressions. This is imitating the human ability to slightly chance the color of the skin e.g. at the face. If someone is angry, sad, happy and so on, this can be underlined by colors. In robotics this is used in different ways. Some systems like FloBi[1] try to imitate direct human behavior by integrating lights behind the cheeks. Other systems like Simon[2] have light sources elsewhere around the head.



a) b) c)

Figure 3: Emotions via colors, a) and b) show FloBi[1] robot (left one with red cheeks), c) Simon[2] with colored "ears"

3.2 Acoustic

Another way to interact is with acoustic systems.

- Sound signals
- Speech recognition and speech output

In [Iwahashi2003] the authors present a way for a robot to acquire language by continuously associating speech input signals with visual and behavioral information. A robot is shown objects (a mug, a toy, etc.) which are moved from one place to another while at the same time the name of the object as well as the action that is performed on it is spoken to the robot's voice input system. Complex algorithms implemented inside the robot allow the robot to not only learn the names of objects and actions but also derive a grammar for the language in which the audio input is given. This allows the robot to continuously acquire a language by learning words, grammar and associating voice signals with visual recordings of objects.
The intention behind this approach is that the authors believe that a useful conversation between two beings (human-to-human as well as human-to-robot) greatly depends on the common experiences of the conversing parties. These experiences

[1] http://aiweb.techfak.uni-bielefeld.de/flobi-icra2010
[2] http://www.simontherobot.com

are taught to the robot by a human in much the same way a child is learning to speak and associate objects and actions with names and expressions.

In case of emergency or guard dog robots the acoustical interaction also includes warning signals via loudspeakers. They can present alarm signs or warn a user not to come closer or leave some place. In the case of emergency robots this is useful because visible signs may not be seen (e.g. through smoke at a fire).

3.3 Tactile

Tactile interaction is very common way to interact with electronically systems (push buttons etc. But also the robotic systems are able to use tactile stimuli to interact with the user.

- Touch(screens)
- Button-based Interfaces
- Vibration Feedback / Alarms

In [Schaefer1999 and Hans2001] the authors describe outlooks of the functionality of a robot system they have built. Among these are helping in household tasks such as fetching or carrying objects, support in grasping, holding and lifting objects, serving drinks and also control of technical home infrastructure. Although the authors do not specifically mention how the system is doing it, the idea of using the robot system to interact with technology at home is an interesting concept. The authors describe a scenario, in which the robot could provide an interface to turn on or off lights, heating systems or roll up and down shutters. Furthermore, they mention the robot as a communication platform supporting the user in making video calls, communicating with a physician, doctor or other authorities and for the supervision of vital signs. A first version of a robot that had this functionality was built in 1998. The project's name was Care-O-bot (Figure 4).[3]

Figure 4: First version of Care-O-bot (Fraunhofer IPA)

[3] http://www.care-o-bot.de/english/ (Fraunhofer IPA)

Systems like Care-O-bot use tactile touch screen interfaces to interact with the user. Most of the state of the art service robots include touch screens and buttons to interact with the user.

3.4 Gesture-based

- Gesture recognition (mechanical or video based)
- Laser-pointing gestures

Gesture-based human-robot interaction is a common research subject since – if done well – it will allow a user to interact with a robot in a rather natural way by signaling the robot what it should do.

Gesture interaction techniques are good providers of intuitiveness; easy to explain, powerful in its implications, impossible to forget. *"Furthermore, gesture interaction brings pleasure to us as it provides something we don't get to do much when working with computers and other interactive tools: Gestures are done from muscle memory, rather than cognitive memory, just like typing on a computer keyboard. Most of the time tools that run on software tax our cognitive capacity but leave the intelligence that lives in our bodies relatively untapped, which makes us East African Plains Apes a little uncomfortable; using those gestures makes us happier animals."* (from Jonathan Korman)[4]

In [Cipolla1996], the authors describe a mechanism to detect the direction indicated by the index finger of a human operator and track the hand in 3D space without the need for a full 3D scene reconstruction. This mechanism can be used to indicate actions and directions to a robot which is equipped with a stereoscopic camera system. The authors prove that the concept is working, however also greatly depends on computational power as well as lighting and contrast conditions. Apparently tracking hand movements in the stereo view system is a major problem of the algorithm presented.

In [Waldherr2000], a gesture-based interface for human-robot interaction with a service robot is proposed. The goal is to instruct a mobile robot both through pose (i.e. static pointing) and motion gestures. To this end, a dual-color tracking algorithm is used that allows motion speeds of up to 30 centimeters per second. For the recognition of actual gestures, a neural network-based approach and a graphical correlation template matching algorithm are applied. The results of this project were quite accurate with an error rate of less than 3%; however, there were limitations in the tracking algorithms and a constraint that people keep a constant distance to the robot in order to allow the detection of gestures properly. The authors conclude that it would be beneficial to not only rely on gestures but also on speech input and combine these two modalities to increase accuracy.

Gesture recognition can be resolved mechanically or through vision based techniques. Mechanical gesture recognition requests wearable sensing technology (position and acceleration sensors), while vision based recognition is based on one or several cameras analyzing images/videos in search for behavior or gesture patterns.

[4] http://www.cooper.com/journal/2007/10/intuition_pleasure_and_gesture.html

Examples of mechanical gesture recognition

Ubi-Finger [Tsukada2002] is a compact gesture input device attached to the finger. It realizes sensuous operations on various devices in real world with gestures of fingers. The devices are selected with an infrared sensor. The gestures are similar to the one used in real life and so easy to remember. The “beginning” of a gesture is initiated by pressing a button on the *Ubi-Finger* to lower ambiguity.

GestureWrist [Rekimoto2001] is a wristwatch-type input device that recognizes human hand gestures by measuring electronic capacity changes in the wrist shape caused by sinews arrangements due to gestures. To do this, a combination of transmitter and receiver electrodes is attached to the back of the watch dial and inside of the wristband. When a gesture is recognized, the *GestureWrist* gives tactile feedback to the user. Inside of the wristwatch dial, a ceramic piezoelectric-actuator is attached to produce the feedback. The *GestureWrist* can be used as a command-input device, with a physical appearance almost identical to today’s wristwatches.

Examples of vision based gesture recognition

A gesture recognition system developed by Karpouzis et al [Karpouzis2004] performs hand and head localization based on skin color detection preceded by morphological operations and gesture classification using Hidden Markov Models (HMM) trained from predefined sequences. The system is especially suited to communication impaired people and people who are not familiar to traditional input/output devices. One big advantage is that it is user-independent and no additional devices are needed because it is vision-based.

The *Gesture Pendant* [Gandy2000] allows ordinary household devices to be controlled, literally, with the wave of a hand. The user wears a small pendant that contains a wireless camera. The user makes gestures in front of the pendant that controls anything from a home theater system, to lighting, to the kitchen sink. The pendant system can also analyze the user's movement as he/she makes gestures. This means that the system can look for loss of motor skill or tremors in the hand that might indicate the onset of illness or problems with medication. It can also observe daily activities to determine, for example, if a person has been eating regularly and moving around. *Gesture Pendant* is a camera-based gesture recognition system that can be worn like a pendant. A user can hand gesture in front of it while it is worn around the neck. The current prototype is still noticeably bigger than an ideal one, and would always have to be worn over their clothes.

Examples of laser based gesture recognition

Another way of presenting gestures to a robot is the way of “point and click” which a lot of users are already familiar with from computer systems.
Kemp et al. [Kemp2008] have designed such a point-and-click interface for a robot system. This system enables the user to show a desired object to the robot and the robot can get this object for the user. “The human points at a location of interest and illuminates it (“clicks it”) with an unaltered, off-the-shelf, green laser pointer. The robot detects the resulting laser spot with an omnidirectional, catadioptric camera with a narrow-band green filter. After detection, the robot moves its stereo pan/tilt camera to look at this location and estimates the location’s 3D position with respect to the robot’s frame of reference.” (pg. 8).
Since this kind of pointing does not require any further learning of objects (from robot side) it can be used instantly in every environment. E.g. this can help users that are

impaired and not able to stand to reach higher places. Or elderly persons that are not able to bend down anymore. They can stay in their seat and just point to the desired object.

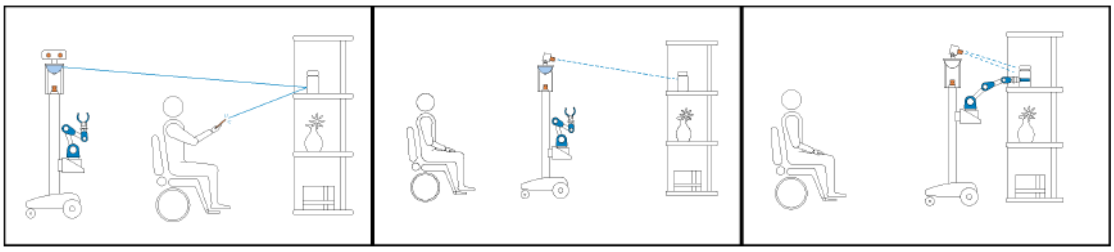


Figure 5: Point and click principle [Kemp2008]

This idea of laser gestures was extended by Ishii et al. [Ishii2009]. They not only used the laser to point at an object, the laser is also used to give a command what to do with the object. They use a robot similar to a vacuum robot[5]. The robot is capable of moving objects that lie on the floor to another location. The laser is used to mark an object (if the user "draws" a circle around it with the pointer) and afterwards it is used to show the desired path with its destination.

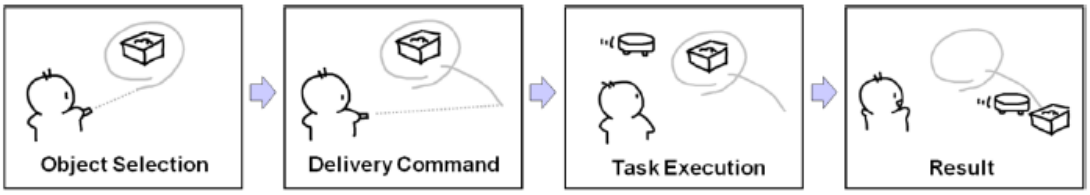


Figure 6: Point and command principle [Ishii2009]

3.5 Comparison of modalities

In this section we analyze input interaction modalities, i.e., methods by which a user of a device can issue commands to the device, or give some feedback to the device. We first present a high-level classification of input modalities, and then analyze the benefits and drawbacks of each class. Our analysis concentrates on ease of use versus feasibility, since we treat the ease of use as the most important benefit (for the user), and (in)feasibility as the most important drawback (for an implementer).

One can roughly group input interaction modalities into five classes:

- brain waves
- natural language
- gestures
- tangible objects
- graphics

[5] Like Roomba from iRobot, http://www.irobot.com; or XV-11 from Neato Robotics, http://neatorobotics.com/

The particular order in which we present the classes reflects the ease of use. The first three classes are easier for the user than the remaining two, since they involve the user's body alone, without any need for carrying around special objects (at least in ideal case).

Brain waves:
The ultimate solution to the problem of controlling devices would be the *read my mind* approach. Imagine a device that can sense your brain waves. Assume the device is so intelligent it can analyze the brain waves, deduce your specific needs, and find a way to adapt its functionality to fulfill them. You would control such a device simply by thinking.

One could even consider a more advanced approach termed *anticipate my mind*. If the device was fast enough, it could guess your needs before you consciously formulated your thoughts, like a good butler who serves your favorite drink before you even realize that this is exactly what you have just wanted. Of course, it is questionable whether this particular solution would please most people, due to a serious invasion of privacy.

Unfortunately, the current technology is far from realizing this kind of ultimate solutions. The state-of-the-art is an interface that uses EEG (Electro-Encephalography) with several electrodes placed on the head, in the form of a headband or headset. The interface works like a small-scale EEG, if compared to clinical devices. It can capture and amplify not only brainwave activity, like alpha- and beta-waves, but also some stronger electric impulses, e. g. from muscle contractions in the face. A sophisticated software can recognize certain learned patterns and generate events.

Most devices on the market, like the NIA[6] (Neural Impulse Actuator) from OCZ, EPOC[7] from Emotive Systems or MindSet[8] from NeuroSky, are intended as input devices for computer games and have an SDK for user applications, because of the very small market and publicity. Some projects use the technology to help physically impaired persons. For example, Brainfinger[9] is a neural interface developed by Brain Actuated Technologies, Inc. which allows paralyzed persons to control a computer and type texts. Another medical use aims at psychological disorders like ADHD. Used in therapy, the EEG and an output device acts as a biofeedback system, helping ADHD patients to learn to focus at a task.

But most of the home-use devices are not sophisticated enough and have to deal with recognition problems through the weak signals of brainwaves and complexity of the patterns.

Natural language:
The second best (as far as the ease of use is concerned) solution would be an interface based on natural language, either spoken or written. A typical approach to processing a natural language consists of three steps:

[6] http://www.ocztechnology.com/products/ocz_peripherals/nia
[7] http://www.emotiv.com/
[8] http://www.neurosky.com/mindset/mindset.html
[9] http://www.brainfingers.com/

- tokenization (i.e., recognizing words, by recognizing particular phonemes and grouping them into sequences)
- syntactical analysis (i.e., parsing according to grammar rules)
- semantically analysis (i.e., extracting meaning from parsed phrases)

Again, the current technology is not mature enough to consider this solution feasible. The main unsolved problem is a reliable semantically analysis of what we say or write. This is a very difficult problem and the Artificial Intelligence research community is nowhere close to solving it in general.

The second serious problem is the way we input natural language phrases to a language processing system. In case of a spoken language, the reliability of the tokenization step still remains problematic. In case of a written language, the problem of recognizing words disappears, but in expense of the keyboard problem. The current ways of entering text are too cumbersome for ordinary users of home devices. A PC like keyboard is just too big, and smaller ones, like in mobile phones, are not convenient for longer texts. Virtual keyboards realized with the help of touch screens are still too expensive (even though getting cheaper). Unless some innovative ways of entering text are found, the keyboard problem will hinder the usefulness of using a written natural language for controlling home devices. Devices like a Laser Virtual Keyboard[10] could be a way to solve it, but these are not working stable enough at the moment.

A good overview of state-of-the-art in natural language processing, with emphasis on spoken languages, is in [Frost2006].

Gestures:
Another class of solutions is based on gestures, as used by deaf people and divers, for example. In order to instruct a device by gestures one would need a camera, or a touch screen, equipped with sophisticated image processing software that would recognize particular movements of your hands and interpret them accordingly. A nice example of how such a system could work in practice is shown in several scenes of "Minority Report"[11] movie where the main hero controls a sophisticated computer system by simply waving his hands.

Although existing technology makes this solution feasible (several demonstrators can be found on YouTube[12]), the cost is still prohibitively high. In addition to the high cost of current gesture recognition systems, approaches based on gestures suffer from a more fundamental problem: the set of gestures with universally accepted meaning is rather small, about a dozen. As a consequence, users would have to learn new sets of specific gestures for controlling devices of different kind. For example, controlling your light system would probably need a different set of gestures than controlling your DVD recorder.

Since the success of Nintendo Wii[13], gesture control seems to be popular in many more consumer products. Examples can be found in CES 2008 [CES2008].

[10] http://www.thinkgeek.com/computing/8193/
[11] http://en.wikipedia.org/wiki/Minority_Report_(film)
[12] "Cynergy Labs: Project Maestro" http://www.youtube.com/watch?v=7CoJGrtVs4c or "Tracking fingers with the Wii Remote" http://www.youtube.com/watch?v=0awjPUkBXOU
[13] http://www.nintendo.com/wii

Tangible objects:
The problem with universally accepted gestures can be alleviated by tangible interfaces. A tangible object can easily convey a meaning, with little effort from a user to guess the meaning. For example, the meaning of the up/down buttons on the traditional remote controller for a TV set is so obvious that we take for granted how easy it is to adjust the volume of a TV set, or select a TV channel, just by pressing the buttons. This value of tangible interfaces has been recognized for a long time. In fact, we configure most devices with the help of tangible interfaces.

Tangible interfaces tend to be costly to design and implement, since they are application specific, which means that they usually have to be done from scratch.

Of course, further research is still needed to find proper (i.e., the most natural for a particular class of users and applications) sets of objects, and ways of manipulating them.

Graphics:
The traditional approach to simplify the way we control complex systems is to hide the complexity behind a fancy GUI.

The main drawback of graphical approaches is that they currently relay on using a PC (i.e., a processor with monitor and mouse). In general, graphical approaches to controlling complex systems work quite well in situations where users are semi-professionals (say, hobbyists and researchers) and a PC is an accepted input device. Recent developments in touch screen technology and tablet PCs allow the increased use of such systems.

The graphical approaches have turned out to be quite successful in some particular application domains, for example in robotics. There, the most prominent are Philips OPPR (the software system used to control iCat)[14], Lego Mindstorms[15], MIT Scratch[16] and Microsoft Robotic Studio[17].

The above mentioned limitations are likely to disappear in several years when the Tablet PC becomes cheap enough to be considered a feasible replacement for the traditional remote controller. Think in terms of a device similar to Smart phones or smart tablets (popular examples: Apple iPhone or iPad, but there are a lot more) with a touch screen on which you can manipulate various graphical elements with your finger.

Conclusions:
User interface technologies can be roughly partitioned into five categories: brain waves, natural language, graphics, gestures and tangible objects. This particular partitioning forms two cascades. The first one is a descending cascade from easier to more difficult usage, and the second one is an ascending cascade from less feasible to more feasible implementation.

[14] http://www.research.philips.com/technologies/projects/robotics/index.html
[15] http://mindstorms.lego.com/
[16] http://scratch.mit.edu/
[17] http://www.microsoft.com/robotics/

In summary, using thoughts to control home devices would be the easiest for end users. Unfortunately, the technologies to sense brain waves, and analyze them, are in very early experimentation phases, and thus not currently feasible for commercial home management applications. The second easiest would be a natural language interface. Although some more or less successful applications of natural language interaction between humans and devices exist, this technology suffers from the fundamental problem of reliable semantically analysis of spoken or written phrases. Currently, only very simple subsets of natural language can be analyzed reliably, and this severely limits the application of this technology in home environment. Graphical user interfaces solve the semantically analysis problem. However, they introduce the PC problem (one needs a monitor and a pointing device to use GUIs). The PC problem is solved by gestures (at least on the pointing device side) but then the set of universally understood gestures is rather small (especially across different cultures) so we are back at the semantics problem. Finally, tangible interfaces seem to solve both the semantics problem (since a tangible object conveys a meaning by itself) and the PC problem simultaneously, so we conclude that tangible interfaces should be considered as the most promising interaction technology for home management.

3.6 Visual vs. Physical interaction and combinations

Some modalities may never be used on their own but mixed/combined with different ones to get a full working interaction.

3.7 Hybrid approaches

Tangible interfaces should be application specific. However, application specific solutions are costly since they need to be developed from scratch for every new application.

To lower the cost, one may consider solutions that combine several approaches. For example, imagine a 3D monitor equipped with some sensors (say, cameras) that can measure the position of your hand in front of the monitor. Assume the monitor shows a scene (say, generated by a 3D gaming engine) that contains visualizations of some tangible objects. The scene appears to you in front of the screen (that's why we need a 3D monitor for) so you can attempt to grab the objects and reposition them. Of course, the objects are virtual, not tangible, so you are not really repositioning them. Instead, the 3D gaming engine creates a sequence of new scenes in response to the sensors that measure the position of your hand. This is very similar to virtual reality but without the cumbersome gloves, helmets and goggles. Notice that this solution combines graphical, gesture and tangible interfaces. A similar combination can be achieved by using a touch screen, as in the Philips Entertaible[18], but without the 3D effect.

Tangible interaction
Tangible interaction refers to a person that interacts with digital information through the physical environment. The initial term used to design such an interaction was *Graspable User Interface* [Fitzmaurice 1996].

[18] http://en.wikipedia.org/wiki/Philips_Entertaible

In [Ishii1997] conclusion was that GUIs fall short of embracing the richness of human senses and skills people have developed through a lifetime of interaction with the physical world. In this work an effort was made to try and illustrate part of this richness in "Tangible Bits." Tangible Bits allows users to "grasp & manipulate" bits in the center of users' attention by coupling the bits with everyday physical objects and architectural surfaces. Tangible Bits also enables users to be aware of background bits at the periphery of human perception using ambient display media such as light, sound, airflow, and water movement in an augmented space. The goal of Tangible Bits is to bridge the gaps between cyberspace and the physical environment, as well as the foreground and background of human activities. In the following figure three prototype systems – the metaDESK, transBOARD and ambientROOM try to identify underlying research issues of the three key concepts of Tangible Bits: interactive surfaces, the coupling of bits with graspable physical objects and ambient media for background awareness.

In the case of the Florence system which integrates robot and home environments, ambient media feedback could be implemented. If the robot included graspable objects for interaction users may find it easier to interact with the services running on the Florence system.

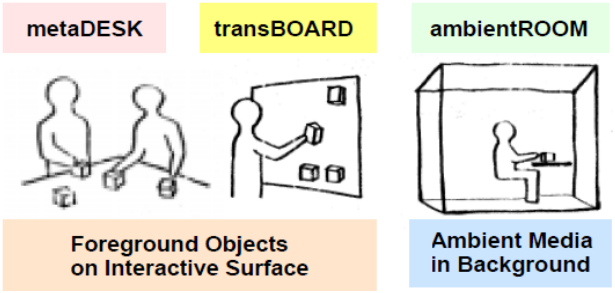


Figure 7: Tangible interfaces [Ishii1997]

Hybrid interaction

In "Designing Hybrid Interactions through an Understanding of the Affordances of Physical and Digital Technologies" PhD Thesis, [Terrenghi2007] states that to understand how different aspects of the physical world can be integrated in the design of interactive systems it is necessary, first of all, to thoroughly examine what specific qualities of physical interaction could be drawn upon as design resources (whether this be consciously or not).

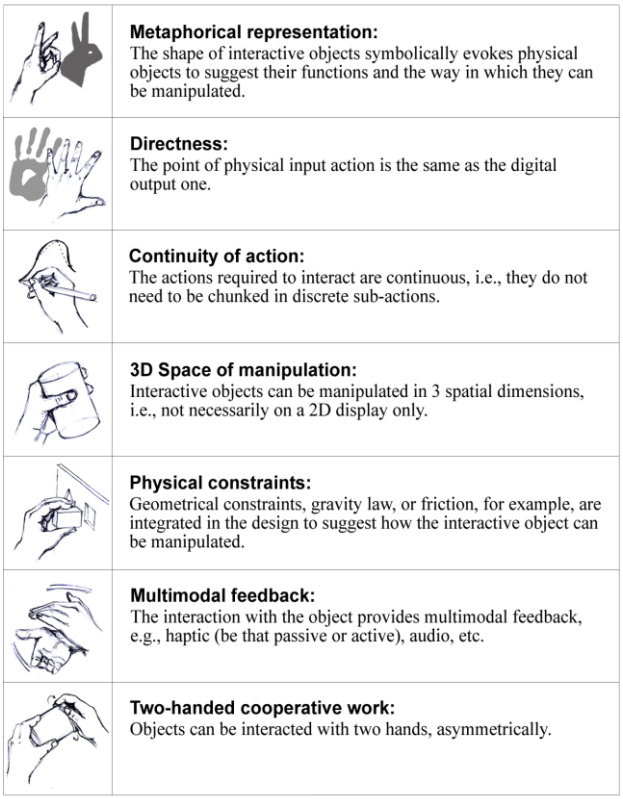


Figure 8: Specific qualities of physical interaction [Terrenghi2007]

In the design of hybrid interactions both input and feedback are considered. They both can be encouraged or represented in the physical object, producing an effect in the virtual world or a mixed interaction can happen, were the input is through the physical object and the feedback is represented in the digital world or vice versa. Physical constraints could be implemented in the Florence system to minimize the possibilities of action perceived by the user at each time or with each service.

Different studies on hybrid interaction have used different technologies and different physical concepts *(metaphors)* to enrich the interaction experience:

- [Maekawa2009] *MADO Interface: a Window like a Tangible User Interface to Look into the Virtual World*- a tangible user interface consisting of a compact touch-screen display and physical blocks. “MADO” means “window” in Japanese, and MADO Interface is utilized as the real window into the virtual world. Users construct a physical object by simply combining electrical blocks. Then, by connecting MADO Interface to the physical object, they can watch the virtual model corresponding to the physical block configuration (shape, color, etc.)

- [Want1999] *Bridging Physical and Virtual Worlds with Electronic Tags*- designing and building a number of new physical prototypes which incorporate key technologies to address issues which limited the use of earlier experimental systems. In particular, they have combined four technologies (RFID identifier tags and readers, RF networking, infrared beacons, and portable computing) in a seamless and tightly integrated way.

- [Rekimoto2001] *DataTiles: A Modular Platform for Mixed Physical and Graphical Interactions*- Tagged transparent tiles are used as modular construction units. These tiles are augmented by dynamic graphical information when they are placed on a sensor-enhanced flat panel display. They can be used independently or can be combined into more complex configurations, similar to the way language can express complex concepts through a sequence of simple words.

In the Florence system the different services should be studied and metaphors of the services' domains should be analyzed in order to provide possible hybrid interaction modes. Interaction enhancement will not guarantee acceptance, but will help bring down the learning curve which substantially contributes to the acceptance of a system.

3.8 Conclusions

Norman [Welie2000] suggests several user interface principles, which gives an indication of the kind of problems and questions (identified by [McKay1999]) users may have when interacting with a system. These guidelines can be used to check the interaction modalities that will be used in the Florence system.

- Visibility: For providing the user the ability to figure out how to use a feature or an interface just by looking at it.
- Affordance: To involve the perceived and actual properties of an object that suggests users how the object is to be used.
- Natural mapping: For creating a clear relationship between what the users wants to do and the mechanism for doing it.
- Constraints: To reduce the number of ways of performing a task and the amount of knowledge necessary to perform it, and making it easier to understand.
- Conceptual models: For making a user understand how something works corresponds to the way it actually works. To enable user to confidently predict the effects of their actions.
- Feedback: To indicate the user that a task is being done or is being undertaken correctly.
- Safety: To protect users against unintended actions or mistakes.
- Flexibility: To allow user to change their mind and perform task differently.

With these guidelines and the overview of the principles mentioned in this section, we think we will be able to select appropriate interaction modalities for the Florence system. Since a lot of research groups around the globe develop new insights in

human robotic interaction, we will continuously have a look on the latest results produced by these research fields.

4 Dialogue Management

Dialogue is a verbal (spoken) and possibly nonverbal interaction between two or more participants. The interaction must convey meaning, i.e., to be understood by all participants, through a gradual expansion and refinement of the common ground [Lison2010]. This section describes the basic principles of dialogue management and gives an overview of the state of the art in research and available tools/products.

4.1 Introduction

Dialogue Management (DM) determines the communicative actions to carry out given a goal, the context, and the history of the interactions. A dialogue manager, usually a dedicated software component in a dialogue system, controls the flow of the interactions and their content (e.g., what to say) [Lison2010].
In a verbal interaction, a dialogue manager should receive a natural language utterance as input and produce another natural language utterance as output, emulating some human conversational skills. Figure 9 illustrates the main components of a generic dialogue system from [Lison2010].

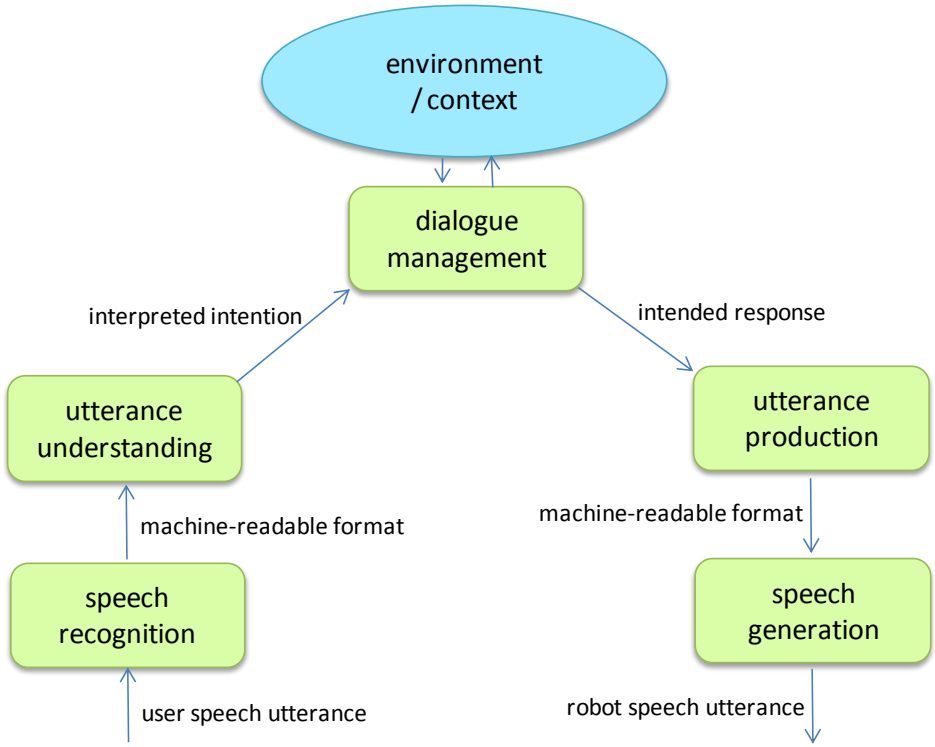


Figure 9: A verbal dialogue system with a dialogue manager [based on Lison2010]

- The speech recognition identifies words and phrases in spoken language and converts them to a machine-readable format, of which the speech understanding component derives the intention as input for the dialogue manager.
- The dialogue manager uses a strategy to emulate human conversation skills like: recognizing dialogue structure, interpreting the context, taking turns,

managing conversational obligations, planning responses and managing uncertainty.

- The dialogue manager produces an intended response, which is converted to a machine readable format by the utterance production component and then to a speech signal by the speech generation component.

4.1.1 Dialogue models

This section gives an overview of main dialogue models.

Finite state dialogue models: a dialogue system based on a finite state dialogue model is very popular in the spoken language processing community. Here the dialogue state is essentially a state variable, from which there are some possible transitions to other states. Dialogue actions that determine transitions from a state may vary widely in such systems from keywords to more abstract sentence mood or illocutionary labels.

Indicating the course of legal conversations, finite state dialogue models have been used for realizing simple tasks such as query answering and template filling. For question answering usually a quite simple model is employed that consists of two states: one in which the question is posed and another in which the answer is provided (or sometimes multiple result states depending on the answer type). There are also some toolkits available for building such systems by wiring together dialogue states with expected actions, see [Bohlin1999] for an overview.

The Information State Update (ISU) model: ISU is a generic, realistic and flexible approach based on a declarative representation of dialogue modeling. The "information state" of a dialogue represents the information necessary to distinguish it from other dialogues. The information state captures the relevant information of previous actions in the dialogue and it motivates a future dialogue action [Larsson2000]. The information state may include also the information state of the participants.

It is important to distinguish information state approaches to dialogue modeling from other structural dialogue state approaches. The latter approaches assume that a dialogue behaves according to some grammar with the states representing the results of performing a dialogue move in previous states and each state allows a set of allowable next dialogue moves. Here the information of the dialogue is implicit in the state. In ISU approach, on the contrary, the information is the dialogue state itself, which can easily be modeled as a register indicating the state number.

Authors of [Georgila2004] integrate reinforcement learning with the ISU approach to dialogue management in order to develop adaptive multimodal dialogue systems. DIPPER[19] [Bos2003] is a collection of software agents for prototyping (spoken) dialogue systems. DIPPER supports building (spoken) dialogue systems, by offering interfaces to speech recognizers, speech synthesizers, parsers and other kinds of agents. DIPPER offers its own dialogue management component (the DIPPER DME), based on the information state approach to dialogue modeling. Authors of [Bohlin1999-

[19] DIPPER Software available at http://www.ltg.ed.ac.uk/dipper

2] defined a dialogue move engine that updates the information state of an agent on the basis of observed dialogue moves and selects appropriate moves to be performed.

Plan or task based approach views dialogue from the perspective of cooperating agents, as its framework is laid down in [Cohen1978]. In this approach each agent has a mental state that also its model of the other participant's mental state. The dynamic mental state usually consists of three basic attitudes:

- Belief that concerns an agent's view of the world including the mental state of other agents. It is also called knowledge especially when regarding the beliefs of another agent thought to be true,
- Desires or goals that indicate the basic motivation for action,
- Intentions or plans that indicate the way the agent has decided to act to achieve its goal.

Markov Decision Processes (MDP) model can be used for generating dialogue strategies as shown in [Young1990] and [Levin1998]. In MDP the system is modeled as a set of dialogue states and a set of speech actions produced by the system. The objective is to maximize the reward obtained for fulfilling a user's request. Representing the dialogue state in MPD is not trivial. A common solution is to indicate the states with respect to a single goal. For example, consider booking a flight in an automated travel agent system, the system state is determined by how close the agent is able to book the flight.

A conventional MDP-based dialogue manager must always know the current state of the system. Such systems perform well under certain conditions, for example, for retrieving e-mails or making travel arrangements over the phone [Levin1998], [Walker1998]. Such application domains are generally low in both noise and ambiguity. Human speech is basically both noisy and ambiguous, and many real-world systems must also be speaker-independent. Any system managing human-machine dialogues must be able to perform reliably even with noisy and stochastic speech input.

Partially Observable PMD (POMDP) model inverts the conventional notion of state in a dialogue and views the world as partially unobservable. The underlying state in POMDP is the intention of the user with respect to the dialogue task, and user utterances are the only observations about the user's state, from which some knowledge about the current state can be inferred. In section 4.2.2 we will describe the basics of POPMD model in detail, therefore this section will just give a brief overview of the related work. As a pioneering work, POMDP is applied in [Roy2000] for the dialogue management problem for the nursing home robot application. In this application, the POMDP states represent the user's intentions; the observations are the speech utterances from the user; and the actions are the system responses. They show that the POMDP dialogue manager operates better than MDP one does with noisy speech utterances (fewer mistakes and automatically adjusting to the quality of the speech recognition). Extending the work of [Roy2000], the model of [Zhang2001] composes the states of the user's intensions and "hidden system states" and infers the observations from the user's utterances and lower-level information of speech recognizer, robust parser, and other input modalities. Later, the authors of [Williams2005] [Williams2006] further extended the model by adding the state of the dialogue from the perspective of the user to the state set. In [Bui2007] the authors extended the POMDP dialogue model by integrating the user's affective states and

divided user's affective states into state and observation spaces. All these approaches focus on spoken dialogue systems.

4.1.2 Issues

In the Natural Language Processing Community dialogue models are based on the idea that dialogues can be modeled as finite-state machines (e.g. [Sitter1992]). In the Formal Methods Community, on the other hand, mathematical logic is used to model technical systems, where the logic is subjected to very powerful analyses using mechanized theorem provers and model checkers. In [Shi2005] the authors combine finite-state machines with formal methods for modeling and controlling dialogue management to ensure robustness and correctness. Their intention is to address particular communication problems of mode confusion and knowledge disparities. *Mode confusions* occur if the human operator loses track of the mode transitions performed by the robot. *Knowledge disparities* occur if the robot's knowledge representation mismatches that of the users. Developing a robot that does according to user expectation requires identification of such situations undoubtedly. Consider a scenario from [Shi2005]: *"wheelchair Rolland is driving down a corridor when a person suddenly steps into its path. Upon seeing the colleague, the user may decide to stop and talk for a moment, uttering please halt. However, unbeknownst to the user, Rolland did not actually acknowledge the user's utterance, but decided to come to a stop of its own accord –having viewed the colleague as an obstacle. Thus, when the colleague moves on, the user will be surprised that the wheelchair continues on its path, despite the user not having instructed it to continue. Such mode confusions become increasingly common with more complex systems and are now well known to have potentially extremely serious consequences."*

Other issues involved in dialogue management are: dealing noisy sensory information and speech recognition, dealing with synchronous/asynchronous dialogues, learning of dialogue policies, having flexibility to learn and adapt to users' preferences, context, content, etc; and dealing with user interrupt-ability delicately.

4.2 Planning under uncertainty

Due to importance of managing a dialogue under uncertainty, this section describes the basics of dealing with uncertainty using the POPMD model.

4.2.1 Planning problem

Dialog management can be seen as a planning & control problem where an agent controls its action in order to achieve a long term goal. This process should take place in an uncertain environment with multiple paths to achieve the objective. This process is complex and requires dealing with tradeoffs.

The long term *goal* of dialogue manager can be to achieve a successful dialogue with the user. The success here is strongly task dependent, for example, when a robot is instructed to move to a particular room, the successful dialogue is to understand which room the user commands to move into. A goal can comprise multiple and competitive objectives and therefore often trade-offs between objectives are needed. In addition to goal and actions, the planning problem includes a set of states, representing a particular view of the environment. The states summarize the interaction history and

allow keeping only the information that is directly relevant to the dialogue. These states can be related by e.g. probabilistic dependencies or independencies. The agent uses actions to achieve its goals. In human computer interfaces the set of possible actions include producing utterances and doing physical movements (e.g., moving eyebrows).

In real circumstances there are two types of uncertainties that a dialogue manager encounters: partially observable states and stochastic dialogue actions. The real state of the dialogue cannot be recognized directly because it is observed through noisy sensory data. Moreover, the course of actions of a particular dialogue cannot be determined in advance. The dialogue management should be considered as “planning under uncertainty” [Lison2010].

4.2.2 Modeling of uncertainty

Partially Observable Markov Decision Processes (POMDPs) provides a powerful modeling framework to deal with uncertainties like: partially observable states, uncertain actions, incomplete knowledge of environment dynamics, and multiples potentially conflicting objectives. POMDPs use probabilistic distributions to encode uncertainties and use a real-valued utility to specify the objectives. Via reinforced learning, moreover, it is possible for POMDPs to learn optimal action policies, or dialogue policies, using both exploration and exploitation strategies [Lison2010].

A POMDP is defined by the tuple <S,A,Z,T,O,R>, see [Bui2007] and [Doshi2007], where

- S is the set of states (of the environment). The states represent the user’s intent (e.g., in a spoken dialog manager it is what the user wants to say and in case of a robotic wheelchair, see [Doshi2007], it is the places the user would like to go),
- A is the set of the agent's actions, which include queries to the user and physical movement.
- Z is the set of observations from the environment that agent can experience. Unlike many conventional dialog managers, the POMDP model assumes that the current state, i.e., the user’s intent, is hidden and infers it from a set of probabilistic observations. In the spoken dialog management setting, the observations correspond to the utterances heard by the dialog manager ears.
- T is the transition model that gives the probability $P(s'|s, a)$ of transitioning from state s to s' if the robot takes action a, given s is the state they were.
- O is the observation model that gives the probability $P(o|s, a)$ of seeing observation o from state s after taking action a.
- R is the reward model that as represented by $R(s, a)$, specifies the immediate reward for taking action a in state s.

In a dialogue management context, the dialogue manager uses a State Estimator (SE) to compute its internal belief about the user's current state s and a policy π to select actions.

- SE inputs are the previous belief state, the most recent action and the most recent observation. SE returns an updated belief state.
- The policy π selects actions based on the system's current belief state.

POMDPs are suitable in designing affective dialogue models because [Bui2007]:

- POMDP allows for realistic modeling the user's hidden states (e.g., user intention) by incorporating them into the state space

- Recent research for spoken dialogue management systems (see [Zhang2001] and [Lison2010, Bohlin1999, Larsson2000, Georgila2004] therein) shows that a POMDP-based dialogue model copes well with the uncertainty arising at different levels of speech recognition, natural language understanding, etc.
- Dynamic Bayesian Networks can usually represent the transition model and observation model of a POMDP. These networks are suitable for simulating the behavior of the user.

A POMDP based dialogue manager can specify various types of actions ranging from spoken language utterances [Bui2007] to movement instructions to relocate a robot [Doshi2007].

4.3 Florence specific aspects

This section describes those DM functions that are specific to the Florence project and will possibly be researched further in the project.

4.3.1 Interactions modes

Within the setting of the Florence project one can recognize the following interaction modes between the elderly user and the robot:

- Direction interaction: in this mode the robot is used as the communication medium. In this mode, the other communicating endpoint can be:
  - The robot itself (when for example asking or reminding something),
  - A remote user such as a caretaker, family member, friend, or …
  - A group of users such as buddies, community members, or …
- Indirect interaction: in this mode the robot is not directly involved in the communication path, but the communication is under control of the robot. In this mode the communication occurs through other medium present in the home environment (e.g., television and mobile phone). In this mode, the other communicating endpoint can be:
  - A remote user such as a caretaker, family member, friend, or …
  - A group of users such as buddies, community members, or …

Human-system interaction modes can also be defined from the perspective of human interaction channels. In [Munoz2009] different interaction channels are distinguished based on five basic physiological senses (visual, auditory, haptic (touch), olfactory (smell), taste). For interacting with users by different services one can distinguish the following possible interaction modes per channel:

- Icons and graphical elements as visual interaction,
- voice and sounds for auditory interaction,
- gestures recognition and tactile displays as haptic interaction,
- taste and smell for olfactory interaction.

Furthermore, a set of additional options such as tangible user interfaces, avatar based interaction, smart objects, multimodality, and adaptive graphical user interfaces, have been also recognized under a so-called spanning channel.

4.3.2 User interrupt-ability

Interacting with users (in e.g. connectedness service), giving feedback to users (in e.g. coaching service) or asking users' feedback (for e.g. supervised learning explicitly)

requires interrupting users by the robot or the system. Managing user interruptions and notifications is an important and well-known problem in human computer interaction. The decision whether to interrupt a user depends on the user context [Kern2003], [Avrahami2007] and [Hofte2007].

For non-time-critical messages the interruption of users (i.e., user notification) can be postponed to an appropriate time [Ho2005]. There is little work in literature aimed at predicting or determining the suitable timing for user interruption. Horvitz and Apacible [Horvitz2003] presented methods for modeling a user's state of being interruptible based on events available in office settings. They used a prediction method based on Bayesian networks to predict the expected interruption cost of users in the current time and in the near future. Their study exclusively focused on fine-grained physical activity changes (like sitting-to-walking/standing and vice versa) by using accelerometer sensors attached to users. Our study in Florence will consider more coarse activity changes (like eating, watching TV, etc) being detectible in robot@home environments. Kern et al [Kern2004] proposed to distinguish between the social and personal interrupt-ability of a user in wearable and mobile settings. They presented a method to estimate the social and personal interrupt-ability of a user from wearable sensors. Chen et al [Chen2007] studied the use of physiological measurements like Heart Rate Variability (HRV) and Electromyogram (EMG) signals for predicting user interrupt-ability status. They found high correlations for both HRV and EMG with user self-reports and they made a linear interrupt-ability model based on these two measures. All these authors proposed models of the user state for being interruptible across users rather than per individual. Further, they solely relied on sensors attached on users, an assumption which does not necessarily hold in Florence settings.

Florence will study the interruption and notification of elderly users and carers by robots and Florence system. The following advances beyond the state of the art will be provided:

- Most of existing works do not attempt to predict or determine the suitable timing for user notifications. Our objective is to fuse sensory data from heterogeneous sources and determine or predict the most appropriate situation (e.g., time and place) to interrupt users. To this end, the criterion is to optimize user experience.
- In principle, our approach will be similar to that of [Horvitz2003]. However, we will focus on robot-user interactions for modeling a user's state of being interruptible based on context information. To best of our knowledge, this issue is not addressed in literature.
- Asking explicit questions (e.g., through the robot) to probe the cost of interruption is another topic that its suitability needs to be investigated within the Florence setting. Our eventual aim, nevertheless, is to converge to a solution whereby no user involvement is required.
- Unlike all works done so far that derive models of the user state for being interruptible across users, our approach will focus on per individual models.

4.4 Existing solutions

4.4.1 Tools and platforms for "human-humanoid" interaction

In this section we give an overview for dialogue management platforms used in the research community for dialogue systems between humans and virtual humanoid characters (avatars), also referred to as conversational agents.

A well-known conversation agent is Greta[20]. Greta is a real-time three dimensional embodied conversational agent with a 3D model of a woman. She is able to communicate using a variety of verbal and nonverbal behaviors. The Greta avatar can talk and simultaneously show facial expressions, gestures, gaze, and head movements. The Greta avatar is used in various European projects: e.g. CALLAS, SEMAINE[21], HUMAINE, and national French projects: ISCC Apogeste, ANR MyBlog3D.

The behavior of "Greta" is specified by means of two standard XML languages: the FML (Function Mark-up Language) and BML (Behavior Modeling Language) allow the user to define her communicative intentions and behaviors. Greta can be used with different external Text To Speech software.

Figure 10: Greta[20]

FML and BML are part of a multimodal behavior generation framework SAIBA[22] for Embodied Conversational Agents (ECAs). This framework is based on a three stage model where the stages represent intent planning, behavior planning and behavior realization. A Function Mark-up Language (FML), describing intent without referring to physical behavior, mediates between the first two stages and a Behavior Markup Language (BML[23]) describing desired physical realization, mediates between the last two stages.

[20] http://perso.telecom-paristech.fr/~pelachau/Greta/
[21] http://www.semaine-project.eu/
[22] http://www.mindmakers.org/projects/SAIBA
[23] http://wiki.mindmakers.org/projects:BML:main and http://www.mindmakers.org/projects/BML

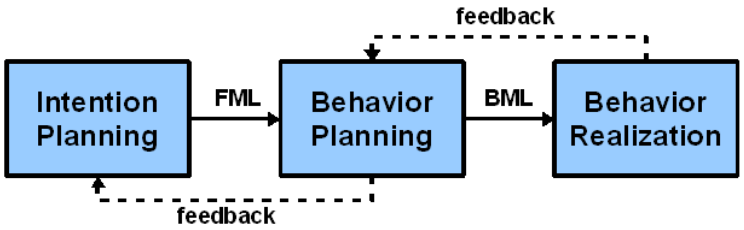


Figure 11: Overview SAIBA[22]

BML descriptions can be embedded in a larger XML message or document by starting a <bml> block and filling it with behaviors that should be realized by an animated agent. This block coordinates speech, gesture, gaze, head and body movement by including a set of corresponding behavior elements inside a single <bml> element. Other possible behavior elements include torso, face, legs, lips and a wait behavior. Every behavior is divided into six animation phases. Each phase is bounded by a sync-point that carries the name of the motion transition it represents. The seven sync-points are: start, ready, stroke-start, stroke, stroke-end, relax and end. Synchrony between behaviors is achieved by assigning the sync-point of one behavior to a sync-point of another, causing the two to align at that point in time. In the example above, the stroke (most effortful part) of the head nod occurs exactly when the speech starts. More information about BML can be found at [Vihjalmsson2007].

Elckerlyc[24] is another BML compliant behavior "realizer" for generating multimodal verbal and nonverbal behavior for Virtual Humans. It is designed specifically for continuous, as opposed to turn-based, interaction with tight temporal coordination between the behavior of a Virtual Human and its interaction partners. Animation in Elckerlyc is generated using a mix between the precise temporal and spatial control offered by procedural motion and the naturalness of physical simulation.

EMBR[25] is another important real time character animation engine that offers a high degree of animation control via the EMBRScript language. In EMBR the viewpoint is that the FML and BML layer of the SAIBA framework are too limited to specify very detailed animation. EMBR therefore provides for an extra layer, the animation layer, to specify the animations in more detail. This is done using the EBMRScript language

4.5 Conclusions

As mentioned in this chapter, "dialog management can be seen as a planning & control problem where an agent controls its action in order to achieve a long term goal. This process should take place in an uncertain environment with multiple paths to achieve the objective. This process is complex and requires dealing with tradeoffs." Within this section some kinds of dialog models are presented as well as approaches to deal with planning problems and modeling of uncertainty. Florence has two different interaction modes (direct / indirect) which have to be considered. With this overview of the state of the art in dialog management we think we will have a good starting point to design the Florence dialog management (also by possibly using the presented existing solutions). The dialog management development will be in close collaboration with the decision making part from WP3.

[24] http://hmi.ewi.utwente.nl/showcase/Elckerlyc
[25] http://embots.dfki.de/EMBR/

5 Interaction / Relation / Acceptance to Robotic Systems based on Appearance

This section concentrates on how robots are designed to fit into different applications. The outer look very much influences how the robot is recognized and how the people try to interact with it.

A major part of a robot (or any other device a human gets in contact with) is the outer appearance of it. The first emotional impression defines the way a person looks at a thing. In case of robots most people think of scary futuristic devices first. Since in lots of movies (starting in the very beginning of the 20th century) the robots are “bad” and trying to demolish the human race, most people look askance at robots. So a robotic system has always to overcome this obstacle before people get used to it.
Robots designed for industrial applications don’t need to look friendly, harmless or the like. They are designed as economical as possible for their special task. Workers in fabrication system don’t need to establish a relationship to the robot which is regarded as a tool.
Robots look very different based on their use as toys, assistants or something else. The user establishes a relationship to this device as it would be an animal or something similar. The stronger the relationship is the more the user will deal with the device. It is reported that many families “adopt” the famous vacuum cleaner robot “Roomba” as a family member [Forlizzi2006]. A lot of research is going on in the field of social robotics e.g. how robots can adopt human behavior to interact with humans [e.g. Breazeal1999, Breazeal2000].
There are also special designed robots to interact with a special kind of clientele, like people with mental disabilities. For example the Keepon Robot is designed for work with autistic children. Such children have problems with social interaction normally. The robot reduces its interaction modalities mainly on eyes and body movement to train the children on concentration.

5.1 Human robotic acceptance basics: the uncanny valley

The *uncanny valley* is a theory in robotics technology. It deals with the user acceptance based on the appearance of robotic systems that are designed to look like humans. It states that at a point where the robot looks and acts nearly but not exactly like a human the robot is more rejected than a system that doesn’t look exact human-like.
It is one of the most poetic, ingenious terms in all of robotics: the uncanny valley. Even without any explanation, it's evocative. Dive deeper into the theory and it gets better. In a 1970 paper in the journal Energy, roboticist Masahiro Mori proposed that a robot that's too human-like can veer into unsettling territory [Mori1970], tripping the same psychological alarms associated with a dead or unhealthy human. "This," writes Mori, "is the Uncanny Valley.”[26]

[26] http://www.popularmechanics.com/technology/engineering/robots/4343054

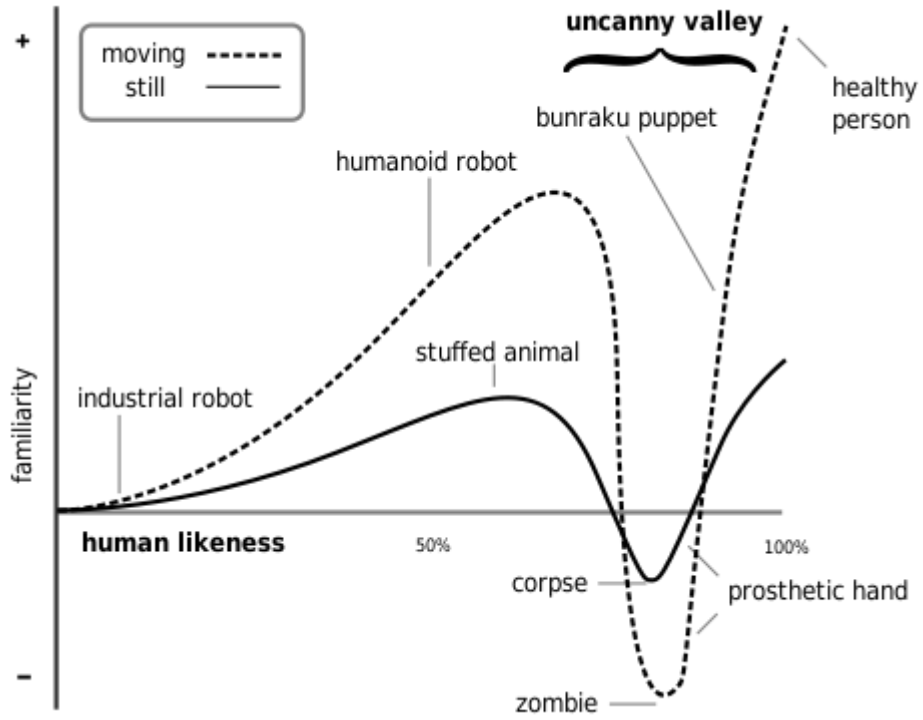


Figure 12: Uncanny valley[27]

Mori's hypothesis states that as a robot is made more humanlike in its appearance and motion, the emotional response from a human being to the robot will become increasingly positive and empathic, until a point is reached beyond which the response quickly becomes that of strong revulsion. However, as the appearance and motion continue to become less distinguishable from a human being, the emotional response becomes positive once more and approaches human-to-human empathy levels.

This area of repulsive response aroused by a robot with appearance and motion between a "barely human" and "fully human" entity is called the uncanny valley. The name captures the idea that a robot which is "almost human" will seem overly "strange" to a human being and thus will fail to evoke the empathic response required for productive human-robot interaction.[27]

A number of theories try to explain this phenomenon (taken from[27]):

- **Mate selection.** Automatic, stimulus-driven appraisals of uncanny stimuli elicit aversion by activating an evolved cognitive mechanism for the avoidance of selecting mates with low fertility, poor hormonal health, or ineffective immune systems based on visible features of the face and body that are predictive of those traits.
- **Mortality salience.** Viewing an "uncanny robot elicits an innate fear of death and culturally-supported defenses for coping with death's inevitability.... Partially disassembled androids... play on subconscious fears of reduction, replacement, and annihilation: (1) A mechanism with a human facade and a mechanical interior plays on our subconscious fear that we are all just soulless machines. (2) Androids in various states of mutilation, decapitation, or

[27] http://en.wikipedia.org/wiki/Uncanny_valley

disassembly are reminiscent of a battlefield after a conflict and, as such, serve as a reminder of our mortality. (3) Since most androids are copies of actual people, they are Doppelganger and may elicit a fear of being replaced, on the job, in a relationship, and so on. (4) The jerkiness of an android's movements could be unsettling because it elicits a fear of losing bodily control."

- **Pathogen avoidance.** Uncanny stimuli may activate a cognitive mechanism that originally evolved to motivate the avoidance of potential sources of pathogens by eliciting a disgust response. "The more human an organism looks, the stronger the aversion to its defects, because (1) defects indicate disease, (2) more human-looking organisms are more closely related to human beings genetically, and (3) the probability of contracting disease-causing bacteria, viruses, and other parasites increases with genetic similarity." Thus, the visual anomalies of android robots and animated human characters have the same effect as those of corpses and visibly diseased individuals: the elicitation of alarm and revulsion.
- **Sorites paradoxes.** Stimuli with human and nonhuman traits undermine our sense of human identity by linking qualitatively different categories, human and nonhuman, by a quantitative metric, degree of human likeness.
- **Violation of human norms.** The uncanny valley may "be symptomatic of entities that elicit a model of a human other but do not measure up to it." If an entity looks sufficiently nonhuman, its human characteristics will be noticeable, generating empathy. However, if the entity looks almost human, it will elicit our model of a human other and its detailed normative expectations. The nonhuman characteristics will be noticeable, giving the human viewer a sense of strangeness. In other words, a robot stuck inside the uncanny valley is no longer being judged by the standards of a robot doing a passable job at pretending to be human, but is instead being judged by the standards of a human doing a terrible job at acting like a normal person.
- **Western, Middle Eastern, and religious constructions of human identity.** The existence of artificial but humanlike entities is viewed as a threat to the concept of human identity, as constructed in the West and the Middle East. This is particularly the case with the Abrahamic religions (Christianity, Islam, and Judaism), which emphasize human uniqueness. An example can be found in the theoretical framework of psychiatrist Irvin Yalom. Yalom explains that humans construct psychological defenses in order to avoid existential anxiety stemming from death. One of these defenses is "specialness", the irrational belief that aging and death as central premises of life apply to all others but oneself. The experience of the very humanlike "living" robot can be so rich and compelling that it challenges humans' notions of "specialness" and existential defenses, eliciting existential anxiety.

Since it cannot be exactly defined how the effect of the uncanny valley is created, it should be at least considered in the stage of design of a robotic system.

5.2 Humanoid robots

Building robots that look like humans is one of the oldest challenges in robot research ever. Since humans try to project their capabilities into the designs they develop, they always tried to build robots that look like themselves, human. In [Takanashi2007] it is stated *"By constructing anthropomorphic/humanoid robots that function and behave like a human, we are attempting to develop a design method of a humanoid robot having human friendliness to coexist with humans naturally and symbiotically, as well*

as to scientifically build not only the physical model of a human but also the mental model of it from the engineering view point."
For long times it has been a nearly impossible to design a human-like robot because the movement is extremely complex and no hardware was able to control such a robot. Nowadays it is getting more and more possible to create and control such systems because the computing hardware is getting more powerful. Robots like the ASIMO[28] from Honda (Japan) are already quite good in walking even while carrying something. Currently big problems are obstacles and unforeseen events. A really challenging problem is the ability of climbing stairs.

The design of robots like ASIMO is kind of clean and futuristic. They should look like modern machines with a human touch. Other robot designs have also a human touch (head, two arms, two legs) but don't look human at all. This design is often used in movies to show the "bad guy" machines. Such a design could be accepted by kids as a toy but not by people using a service robot.

Figure 13: Comparison of different appearances, extreme example[28,unknown]

A recently published matrix[29] of "robot babies" gives also a quick overview of different designs of robots. It shows different approaches to build baby-like robots. Each system has a similar intention but still quite different design.

[28] http://world.honda.com/ASIMO/
[29] Erico Guizzo, http://spectrum.ieee.org/automaton/robotics/humanoids/invasion-of-the-robot-babies-infographic

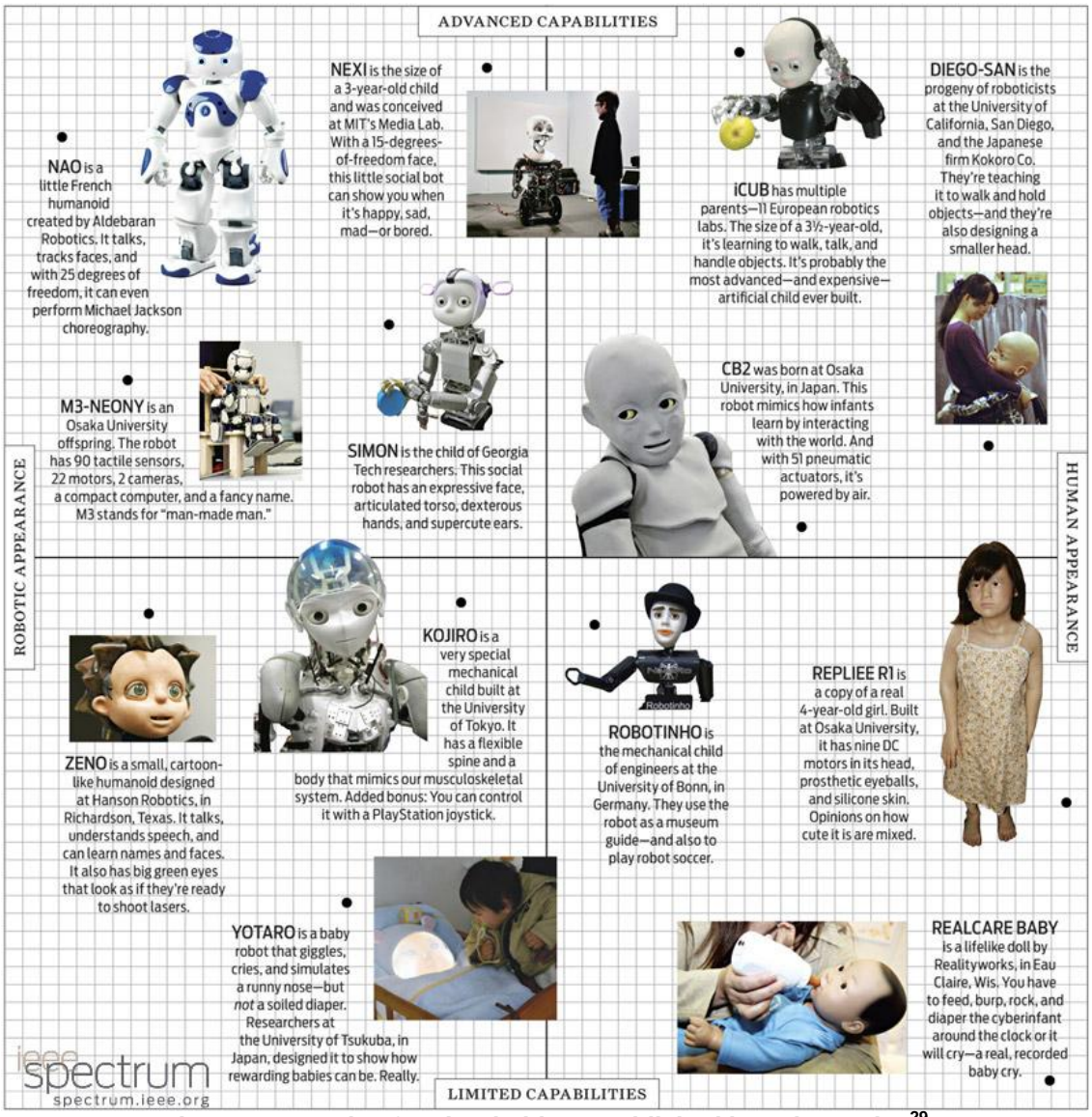


Figure 14: Matrix of "robot babies", published by Erico Guizo[29]

5.3 Animal like robots

Another approach is not to design human-like robots but robots that are looking like pets or animals. This often enables a quicker or easier acceptance because the demands on the intelligence of animals are not as high as on humans. So the people are not disappointed by missing capabilities (e.g. speech) because they are used to not being able to talk to animals. A lot of animal-like robots were designed for medical or therapeutically issues.

PARO[30] is an advanced interactive robot developed by AIST, a leading Japanese industrial automation pioneer. It allows the documented benefits of animal therapy to

[30] http://www.parorobots.com/

be administered to patients in environments such as hospitals and extended care facilities where live animals present treatment or logistical difficulties.

Paro has been found to:

- reduce patients' stress and their caregivers'
- stimulate interaction between patients and caregivers
- have a Psychological effect on patients, improving their relaxation and motivation
- improve the socialization of patients with each other and with caregivers

Figure 15: Paro Robot[30]

Paro has five kinds of sensors: tactile, light, audition, temperature, and posture sensors, with which it can perceive people and its environment. With the light sensor, Paro can recognize light and dark. It feels being stroked and beaten by tactile sensor, or being held by the posture sensor. Paro can also recognize the direction of voice and words such as its name, greetings, and praise with its audio sensor.

Paro can learn to behave in a way that the user prefers, and to respond to its new name. For example, if he strokes it every time he touches it, Paro will remember the previous action and try to repeat that action to be stroked. If he hits it, Paro remembers its previous action and tries not to do that action. By interaction with people, Paro responds as if it is alive, moving its head and legs, making sounds, and showing your preferred behavior. Paro also imitates the voice of a real baby harp seal.

The Paro robot was examined by Kidd et al. [Kidd2006] (and also others) for the social interaction with elderly. The robot was brought in contact with the elderly in a special setting (discussion groups) and the result of interaction was observed. "We observed that the robot Paro, like My Real Baby[31], has features that can increase social interactions. This effect is increased in the presence of caregivers or experimenters who are willing to participate in the interactions. These kinds of interactions provide not only pleasing, feel-good experiences, but also provide evocative experiences for the residents. The robot, for all its limitations, communicates that it enjoys being handled, thus its robotic nature can set up a connection based on the attributes of dependence and nurturance that are read from interactions with it."

Leonardo[32] is a robot developed by Professor Cynthia Breazeal of the Massachusetts Institute of Technology Media Lab in conjunction with Stan Winston Studio and DARPA. Physically it appears to be anthropomorphic, covered in synthetic fur and having a vaguely humanoid body about two and a half feet tall. The robot has a highly

[31] Robotic Baby Doll, http://www.generation5.org/content/2001/mrb.asp
[32] http://robotic.media.mit.edu/projects/robots/leonardo/overview/overview.html

mobile face and arms, but cannot walk. Its purpose is to serve as a "sociable robot" capable of emotional expression, vision, and "socially guided learning." Breazeal believes that the embodiment of Leonardo makes it more capable of forming emotional attachment with humans: "Breazeal has conducted -experiments demonstrating that individuals have a deeper, more intense emotional reaction to Leonardo than to a high-resolution two-dimensional animation of Leonardo on a computer screen."[33]

Figure 16: Leonardo Robot[32]

Leonardo has 69 degrees of freedom - 32 of those are in the face alone. As a result, Leonardo is capable of near-human facial expression (constrained by its creature-like appearance). Although highly articulated, Leonardo is not designed to walk. Instead, its degrees of freedom were selected for their expressive and communicative functions. It can gesture and is able to manipulate objects in simple ways.

5.4 Other special forms

Keepon[34] is a small creature-like robot designed to interact with children by directing attention and expressing emotion. Keepon's minimal design makes its behaviors easy to understand, resulting in interactions that are enjoyable and comfortable.
Keepon has soft rubber skin, cameras in its eyes, and a microphone in its nose:

Figure 17: Keepon Robot[34]

[33] http://en.wikipedia.org/wiki/Leonardo_%28robot%29
[34] http://beatbots.net

Keepon has four degrees of freedom. Attention is directed by turning +/-180° and nodding +/-40°, while emotion is expressed by rocking side-to-side +/-25° and bobbing up to 15mm:

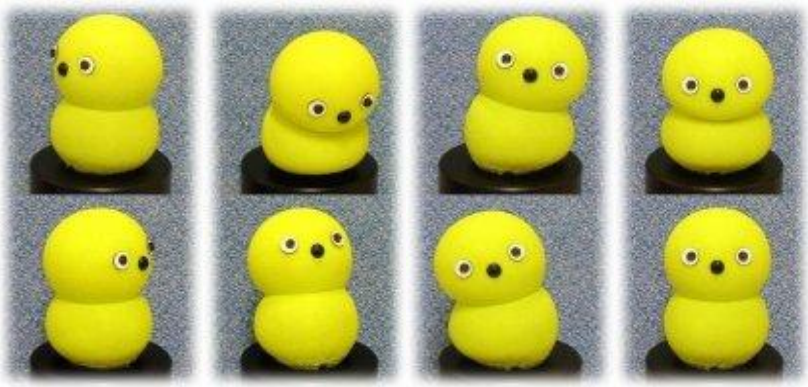


Figure 18: Movements of Keepon robot[34]

Keepon has been used since 2003 in research on social development and communication. Studies concentrated on behaviors such as eye-contact, joint attention, touching, emotion, and imitation between Keepon and children of different ages and levels of social development. In the case of children with autism and other developmental disorders, the results were encouraging with the use of Keepon as a tool for therapists, pediatricians, and parents to observe, study, and facilitate social interactions.

5.5 Mimics/Facial expressions for (social) interaction

Another important aspect is not only the complete outer appearance but also the facial expressions of the robot. A robot with these capabilities will be able to give emotional feedback to a user. For interaction between humans, this is a very important for how the speech is interpreted.

"Robots are slowly learning to take advantage of nonverbal cues, which are an integral part of natural human communication. Things like eyebrow movement can be used to convey information about emotional states, but subtle nonverbal cues can also provide more pointed information. Think about how much can be said with a quick glance. Professor Bilge Mutlu [Mutlu2009] from the University of Wisconsin is trying to teach robots to communicate information with glances, by seeing how well effectively they can "leak" the answer to an object guessing game played with a human.
[…]
There is one caveat at the end of the paper: "We found that the leakage cue affected the performance of only pet owners and not others, which might suggest that pet owners become, through their interaction with their pets, more sensitive to nonverbal behavior."
[…]"[35]

Cynthia Breazeal from MIT Artificial Intelligence Lab (they also created the Leonardo robot mentioned above) studied "How to build robots that make friends and influence people" [Breazeal1999]. It is stated that "In order to interact socially with a human, a

[35] http://www.botjunkie.com/2009/11/03/robots-learn-to-look-shifty

robot must convey intentionality, that is, the human must believe that the robot has beliefs, desires, and intentions." To achieve that, the robot has to be able to display facial expressions.

5.6 Conclusions

The robotic user acceptance is an issue targeted by many researchers. Improvement on user acceptance is considered in the Florence project a key issue of its philosophy.

The user acceptance has been studied from various perspectives including general appearance, social abilities and service dependent appearance. *"Acceptance is defined as the healthcare robot being willingly incorporated into the person's life. For acceptance of robots to occur, there are three basic requirements: a motivation for using the robot, sufficient ease of use, and comfort with the robot physically, cognitively and emotionally."* [Broadbent2009]

Results from a user tests published in the 2003 Special issue on Socially Interactive Robots stated that: *"the majority of the participants preferred to view a service robot as a smart appliance, although other perspectives, such as a personal assistant, were also rated as acceptable.[…] When asked about the degree of independence of a robot, the majority of subjects preferred a robot that does only what it has been instructed to do, and does not act independently."* [Sverinson2003]. In the Florence project the user will be in control for all services on demand and for in those services that need to be permanently active for long term assessment or vigilant services, users will always be informed when they are active and will be given the option of switching them off.

The robot's physical appearance has shown to be related to the capabilities the users relate to them. *"Visual appearance of social robots is a significant predictor of the participants' attributions of applications. Particularly the distinction of human-likeness and animal-likeness led the naïve participants to have distinctive expectations of what applications match the robot's capabilities. Human-like robots were expected to be used in the application fields of security, business, research, healthcare, personal assistance, teaching, transport, care giving, and public assistance. Animal-like robots were expected to serve as companions, entertainers, toys, and robotic pets."* [Hegel2009]. The Florence robot is a multipurpose robot that aims at target in both AAL and lifestyle services. The consortium needs to approach the robot appearance in such a way that it is neither perceived as a mere companion nor as a functional device that might produce rejection. If the hardware appearance is kept neutral enough, variations of the touch screen's visual image (colours, change in avatar representation) as well as variations in the interaction mode (e.g.: different voice tone) could help to define different states or profiles for the robot.

So the goal of the Florence system design (appearance) will be to design a system that is very easy to use and also creating a kind of relation to the user so that the user feels comfortable with it. There has been a lot of research in this field and we will try to use as much experiences of other groups as possible.

6 Human robotic interaction Frameworks

This section collects human robotic interaction frameworks (in difference to D2.1 which deals with robotic platforms). So a list of interaction frameworks is provided here.

Frameworks specialized on human robotic interaction are rare. Since most robotic frameworks include more than interaction itself they are mentioned in Deliverable 2.1. There is no “ready-to-go” and “downloadable” framework available, although some papers are published in this field. Lee et al. [Lee2005] propose a human-robot interaction framework for home service robots. Their framework is divided to three interaction modules: multimodal, cognitive and emotional. Figure 19 shows the general outline of the framework.

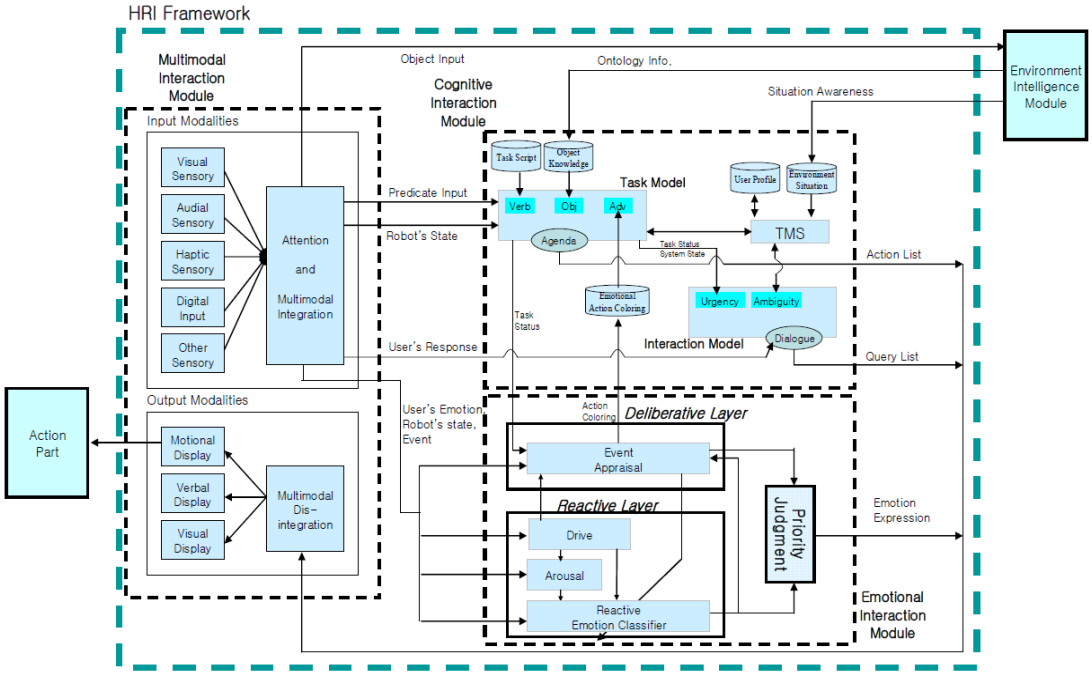


Figure 19: Homan robot interaction framework from [Lee2005]

The multimodal module provides different communication channels to the outside world. Since the cognitive and the emotional modules are closely related, the distinction is made by the orientation of the modules. The cognitive module is task-oriented and the emotional model is human-oriented. Parts of this infrastructure may also be suitable for the Florence robot.

Another specialized approach is **ROILA**[36] the “Robotic Interaction Language” developed by Bartneck et al. The following paragraph is taken from the website www.roila.org where the function and history of ROILA is explained.

“The number of robots in our society is increasing rapidly. The number of service robots that interact with everyday people already outnumbers industrial robots. The easiest way to communicate with these service robots, such as Roomba or Nao, would be natural speech. But current speech recognition technology has not reached a level

[36] http://roila.org/

yet at which it would be easy to use. Often robots misunderstand words or are not able to make sense of them. Some researchers argue that speech recognition will never reach the level of humans.
Palm Inc. faced a similar problem with hand writing recognition for their handheld computers. They invented Graffiti, an artificial alphabet, that was easy to learn and easy for the computer to recognize. ROILA takes a similar approach by offering an artificial language that is easy to learn for humans and easy to understand for robots. An artificial language as defined by the Oxford Encyclopedia is a language deliberately invented or constructed, especially as a means of communication in computing or information technology.
We reviewed the most successful artificial and natural languages across the dimensions of morphology and phonology (see overview in the form of a large table) and composed a language that is extremely easy to learn. The simple grammar has no irregularities and the words are composed of phonemes that are shared amongst the majority of natural languages. The set of major phonemes was generated from the overview of natural languages. Moreover, we composed a genetic algorithm that generated ROILA's words in a way that they are easy to pronounce. The same algorithm makes sure that the words in the dictionary sound as different from each other as possible. This helps the speech recognizer to accurately understand the human speaker.
Most previously developed artificial languages have not been able to attract many human speakers, with the exception of Esperanto. However, with the rise of robots a new community on our planet is formed and there is no reason why robots should not have their own language. Soon there will be millions or robots to which you can talk to in the ROILA language. In summary, we aim to design a "Robotic Interaction Language" that addresses the problems associated with speech interaction using natural languages. Our language is constructed on the basis of two important goals, firstly it should be learnable by the user and secondly, the language should be optimized for efficient recognition by a robot."

The approach of creating a new language for interaction is not feasible for the Florence project. Elderly people will have a lot of difficulties in learning a new language and the barrier to use the system will be very high.

Omek Interactive[37] is working on a so called "**Shadow SDK**" which is a motion and gesture recognition framework. This is known from systems like the EyeToy[38] camera from Sony Playstation or the upcoming Project Natal[39] from Microsoft.
The difference is that Omek is distributing an open SDK that can be used for other occasions than gaming.

[37] http://www.omekinteractive.com/index.htm
[38] http://www.eyetoy.com/
[39] http://www.xbox.com/projectnatal/

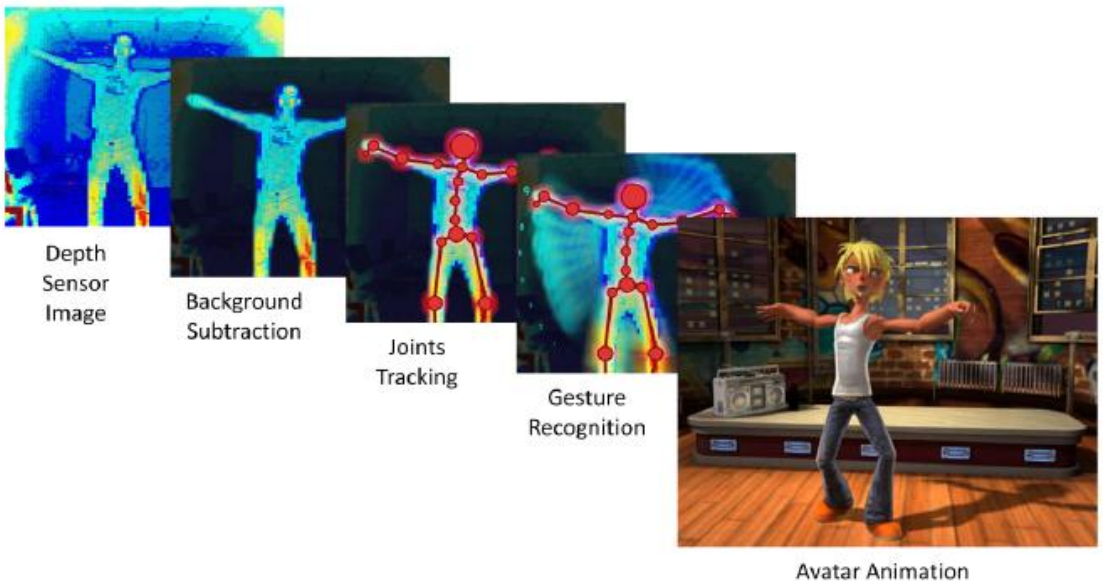


Figure 20: Different working steps of the Shadow SDK framework[37]

It contains a Tracking API, Gesture Recognition Framework, Automatic Gesture Learning Tool Suite and a Motion Gaming Framework. Some of the features include:

Gesture Control:
Simple gesture control interface to register gestures, turn gestures on, off, impose delays and refractory periods.

Hierarchical Structure:
Omek Shadow implicitly uses a hierarchical state machine to manage the dependency structure of the different gestures, enabling the developer to construct complex behaviors from many simple gesture building blocks, such as conditioning the recognition of one gesture on previous detections of other gestures.

Different Gesture Creation Options:
Omek provides a library of pre-generated gestures for developers to immediately integrate into their applications. In addition, developers can manually write their own gestures, or they can generate them using the Automatic Gesture Learning Tool Suite:

- Record people performing the desired gestures and manage the recorded sequence data.
- Mark both positive and negative examples for the desired gesture.
- Train an unsupervised learning classifier on the marked data to generate a new gesture that can be easily integrated into the game – Learning.
- Immediately view the results in real-time to decide if more data should be added to the automatic learning process.

This framework might be useful for Florence because the way of gesture recognition is a very easy way for people to interact with robots. It would be no problem for elderly to control the robot if they just could point to a direction and the robot will move there.

An open source framework for audition is developed by the Kyoto University, Japan. It is called **HARK**[40]. HARK enables a robot to recognize where a sound comes from and even how many people are talking. It uses multiple microphones to be able to detect the direction of sounds. A robot can then turn to that direction to have a look at the sound source.

Figure 21: Sound separation example of HAWK framework[40]

Another feature is the ability of separating sounds even with background noise (e.g. music).
Since the framework is open source so it would be possible to integrate it in a system like Florence and the detection of sound direction and multiple speakers could be a very useful thing. If the person calls the robot it will turn to it and the person is able to have a look at the robot's screen. If there is background noise or more speakers in one room the robot will still be able to understand it's commands.

6.1 Conclusions

The market of specialized human robotic interaction frameworks is not as big as other framework markets but there are quite some usable frameworks available. Florence will try to make use of them (at least the open source frameworks) and add the specific Florence functions afterwards. There are also some other APIs (not complete frameworks) like the Skype API[41] which will also be used to extend the Florence system (the Skype API is not released by the time of writing so it is not included in the list above).
With the help of such frameworks the Florence system will be able to provide functionality like sound recognition and videoconferencing without the need of developing the functions from scratch. This will speed up the development process of the whole system.

[40] http://winnie.kuis.kyoto-u.ac.jp/HARK/
[41] http://developer.skype.com

7 Overall conclusions for the Florence robot

Conclusions for every topic covered by this document can be found at the end of every section. Overall it can be stated that the interaction technologies that will be used by the Florence system have a major influence on how the people will judge the system. First of all the outer appearance will set the first impression for the user. If the user doesn't like the appearance he will probably reject the system even if the functions are useful. To overcome this risk we will use former experiences from literature and also questionnaires and wizard-of-oz-tests. In cooperation with WP3 the decision making and dialog management unit will be developed which will be the "heart" of the communication with the user. On top of this, different interaction modalities will enable the Florence robot to give and receive input to/from the user.

Wakamaru[42] describes that *"Making home robots more socially intelligent can contribute to acceptance".* In user tests carried out with iCat the following behaviours were shown [Heerink2006]:

- listening attentively, for example by looking at the participant and nodding
    - In Florence we could translate this into trying to approach the user and direct the screen towards it.
- being nice and pleasant to interact with, for example by smiling and being helpful)
    - Florence may not be able to smile, it would depend if we include and avatar or not, but it should still provide hints in a friendly manner; either through a friendly voice or through a friendly visual message.
- remembering little personal details about people, for example by using their names
    - We think this could be implemented in Florence.
- being expressive, for example by using facial expressions
    - Maybe we should not take this into account as we are providing more of a service robot than a social robot. Although relevant for user acceptance, in service robotics, functionality is usually preferred over facial expression.
- -admitting mistakes
    - Not only is admitting mistakes a way of creating a more natural interaction, but it is also a way of having systems that can learn much more easily thanks to user feedback when they act incorrectly.

A survey realized to better understand the need of the people for having a service robot at home explained that *"most participants preferred speech (82%) followed by touch screen (63%), gestures (51%), and command language (45%)"* [Green2010]. Even if speech gets high percentages the consortium is aware that when asked, users most of the times relate speech interaction to perfect natural language communication and that is not feasible nowadays. Different modalities will be taken to focus groups for discussion.

"Research suggests that a robot's ability to adapt its behaviour to the user's preferences and personality can improve acceptability"[42]. The WP3 decision making module and learning flow will provide the Florence system with the ability to adapt.

The user acceptance can be increased in several ways, one of them being the robotic appearance, but not the only one and not always the most important; Dautenham [Dautenham2005] mentioned that "Humanlike communication was desirable for a robot

[42] http://www.mhi.co.jp/kobe/wakamaru/english/about/technology.html

companion, whereas humanlike behaviour and appearance were less essential". Improvement on social capabilities and appropriate means of interaction combined with a robotic appearance that suits the services provided should set a well grounded starting point for increased user acceptance. The Uncanny Valley hypothesis shows that human resemblance raises high human like expectations in users, while pet liked are perceived as companions. The approach suggested by the consortium is to avoid both a human like and pet like appearance. An extra action to take into account is to give users the control over their own life, giving them control of the services that are provided to them.

8 SUMMARY

This document shows the state of the art in human robotic interaction. The list mentioned in this document cannot be complete as the area of human robotic interaction is very broad and there is a lot of research going on. We wanted to show major aspects of this research and especially aspects that are of relevance for Florence. This document will help to decide which modalities and which existing solutions to use during the development of the Florence system.
Since this document does not cover every system that is available, we may use or add technologies for Florence which may be discovered later. The developments in WP2 and 3 are also relevant for the human robotic interaction, so some aspects may be changed upon experiences gathered there. WP1 will help to define the bases for the overall design like the appearance of the Florence robot.

9 REFERENCES

[Adams2002] Critical Considerations for Human-Robot Interface Development, J. Adams, AAAI Fall Symposium: Human Robot Interaction Technical Report FS-02-03, November 2002, pp.1-8

[Avrahami2007] D. Avrahami, J. Fogarty, and S.E. Hudson, “Biases in Human Estimation of Interruptability: Effects and Implications for Practice,” In Proceedings of CHI 2007, April-May 2007.

[Bohlin1999] Bohlin P., Bos J., Larsson, S., Lewin, I., Matheson, C., and Milward, D., “Survey of existing interactive systems, Deliverable D3.2., Trindi Project, 1999.

[Bohlin1999-2] Peter Bohlin, Robin Cooper, Elisabet Engdahl, and Staffan Larsson. Information states and dialogue move engines. In Jan Alexandersson, editor, Proc. IJCAI-99 Workshop on Knowledge and Reasoning in Practical Dialogue Systems, August 1999.

[Bos2003] J. Bos, E. Klein, O. Lemon, and T. Oka, “DIPPER: Description and Formalisation of an Information-State Update Dialogue System Architecture”, in 4th SIGdial Workshop on Discource and Dialogue, Sapporo, 2003.

[Breazeal1999] Breazeal, C. & Scassellati, B. (1999b), How to build robots that make friends and influence people, in ‘Proceedings if IROS99’, Kyonju, Korea, pp. 858–863.

[Breazeal2000] Cynthia L. Breazeal, 2000, Sociable Machines: Expressive Social Exchange Between Humans and Robots, PhD thesis, Massachusetts Institute of Technology

[Broadbent2009] Broadbent, E. and Stafford, R. and MacDonald, B., Acceptance of Healthcare Robots for the Older Population: Review and Future Directions, International Journal of Social Robotics, Vol. 1, No. 4, pp. 319-330, Springer, 2009

[Bui2007] Bui, T.H. and Poel, M. and Nijholt, A. and Zwiers, J. (2007) A POMDP approach to Affective Dialogue Modeling. In: Proceedings workshop on The Fundamentals of Verbal and Non-verbal Communication and the Biometrical Issue, 2-12 Sep 2006, Vietri sul Mare, Italy. pp. 349-355. IOS Press. ISBN 978-158603-733-8.

[CES2008] CES 2008 website, http://ces.cnet.com/

[Chen2007] D. Chen, J. Hart and R. Vertegaal, “Towards a Physiological Model of User Interruptability,” In Proceedings of INTERACT 2007.

[Cipolla1996] Human-Robot Interface by Pointing with Uncalibrated Stereo Vision, R. Cipolla, N.J. Hollinghurst, Elsevier, 1996, http://mi.eng.cam.ac.uk/~cipolla/publications/article/1996-IVC-pointing.pdf

[Cohen1978] Cohen, P.R., "On Knowing What to Say: Planning Speech Acts", Ph.D. thesis, University of Toronto, 1978.

[Dautenhahn2005] Dautenhahn, K. and Woods, S. and Kaouri, C. and Walters, M. and Koay, K. and Werry, I., What is a Robot Companion – Friend, Assistant or Butler?, IROS 2005

[Doshi2007] Finale Doshi , Nicholas Roy, Efficient model learning for dialog management, Proceedings of the ACM/IEEE international conference on Human-robot interaction, March 10-12, 2007, Arlington, Virginia, USA.

[Fitzmaurice1996] Fitzmaurice, G.W., 1996. *GRASPABLE USER INTERFACES*. University of Toronto.

[Forlizzi2006] Jodi Folizzi and Carl DiSalvo, 2006, Service Robots in the Domestic Environment: A Study of the Roomba Vacuum in the Home, HRI'06, March 2–4, 2006, Salt Lake City, Utah, USA

[Frost2006] Advances in Speech Technology, Frost & Sullivan, Technical Insights, 2006.

[Hans2001] Hans, M., Graf, B., and Schraft, R.D. 2001. Robotic home assistant Care-O-bot: Past-present-future. In Proc. IEEE InternationalWorkshop on Robot and Human Interactive Communication (RoMan), Paris, France, pp. 407–411.

[Gandy2000] Gandy, M., Starner, T., Auxier, J. & Ashbrook, D. (2000). The Gesture Pendant: A Self-illuminating, Wearable, Infrared Computer Vision System for Home Automation Control and Medical Monitoring. In: Fourth International Symposium on Wearable Computers (ISWC'00), p. 87.

[Georgila2004] K Georgila and O Lemon, "Adaptive multimodal dialogue management based on the information state update approach", W3C Workshop on Multimodal Interaction, 2004.

[Heerink2006] Heerink, M. and Krose, B. and Wielinga, B. and Evers, V., Human-robot user studies in eldercare: Lessons learned, Smart homes and beyond, pp. 31-38, 2006

[Hegel2009] Hegel, F. and Lohse, M. and Wrede, B., Effects of Visual Appearance on the Attribution of Applications in Social Robotics, IEEE International Symposium on Robot and Human Interactive Communication, ROMAN 2009

[Ho2005] J. Ho and S.S, Intille, "Using Context-Aware Computing to Reduce the Perceived Burden of Interruptions from Mobile Devices," In Proceedings of CHI 2005, April 2005.

[Hofte2007] G.H. ter Hofte, “Xensible interruptions from your mobile phone,” In Proceedings of ACM MobileHCI 2007, Singapore, pp. 176-179, September 2007.

[Horvitz2003] E. Horvitz and J. Apacible, “Learning and Reasoning about Interruption,” In Proceedings of the 5th International Conference on Multimodal Interfaces (ICMI2003), November 2003.

[HRIWI2010] Human Robot Interaction, Wikipedia, http://en.wikipedia.org /wiki/Human-robot_interaction

[Ishii1997] Ishii, H. & Ullmer, B., 1997. Tangible bits: towards seamless interfaces between people, bits and atoms. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. Atlanta, Georgia, United States: ACM, pp. 234-241. Available at: http://portal.acm.org/citation.cfm?doid=258549.258715 [Accessed February 23, 2009].

[Ishii2009] Ishii, K., Zhao, S., Inami, M., Igarashi, T., and Imai, M. Designing Laser Gesture Interface for Robot Control. In Proc. INTERACT2009, 479–492.

[Iwahashi2003] Language Acquisition Through a Human-Robot Interface by Combining Speech, Visual and Behavioral Information, N. Iwahashi, Elsevier, March 2003, http://linkinghub.elsevier.com/retrieve/pii/S0020025503001671

[Kadous2006] Effective User Interface Design for Rescue Robotics, M.W. Kadous, R.K. Sheh, C. Sammut, HRI ’06, March 2006

[Karpouzis2004] Karpouzis, K., Raouzaiou, A., Drosopoulos, A., Ioannou, S., Balomenos, T., Tsapatsoulis, N. & Kollias, S. (2004). Facial expression and gesture analysis for emotionally-rich man-machine interaction. In: N. Sarris, M. Strintzis, (Eds.), 3D Modeling and Animation: Synthesis and Analysis Techniques, pp. 175-200. Hershey, PA: Idea Group Publ.

[Kemp2008] Kemp, C.C., Anderson, C.D., Nguyen, H., Trevor, A.J., Xu, Z.: A Point-and-Click Interface for the Real World: Laser Designation of Objects for Mobile Manipulation. In: 3rd ACM/IEEE International Conference on Human-Robot Interaction, pp. 241--248, (2008)

[Kern2003] N. Kern and B. Schiele, “Context-Aware Notification for Wearable Computing,” In Proceedings of the IEEE International Symposium on Wearable Computing (ISWC 2003), IEEE Computer Society, Washington, DC, USA, October 2003.

[Kern2004] N. Kern, S. Antifakos, B. Schiele, and A. Schwaninger, “A Model for Human Interruptability: Experimental Evaluation and Automatic Estimation from Wearable Sensors,” In Proceedings ISWC 2004, pp. 158-165, October-November 2004.

[Kidd2006] Kidd, C. D., Taggart, W., & Turkle, S. (2006). A Sociable Robot to Encourage Social Interaction among the Elderly. ICRA 2006, pp. 3972-3976. Piscataway: IEEE.

[Larsson2000] S. Larsson and D. Traum, “Information state and dialogue management in the TRINDI Dialogue Move Engine Toolkit”, Natural Language Engineering, vol. 6, no. 3-4, pp. 323-340, 2000.

[Lee2005]: KangWoo Lee, Hyoung-Rock Kim, Wan Chul Yoon ,Young-Sik Yoon, Dong-Soo Kwon (2005): Designing A Human-Robot Interaction Framework For Home Service Robot, 2005 IEEE International Workshop on Robots and Human Interactive Communication, pp. 286-293, 0-7803-9275-2/05

[Levin1998] Esther Levin, Roberto Pieraccini, andWieland Eckert. 1998. Using Markov decision process for learning dialogue strategies. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)

[Lison2010] Pierre Lison and Geert-Jan M. Kruijff, “Human-Robot Interaction: Dialogue Processing & Management”, presentation 2010, retrieved from http://www.dfki.de/~plison/lectures/FSLT/HRI.FSLT2010-Feb3.pdf on 15 May 2010.

[Loije2005] Looije R., Cnossen F., and Neerincx, M.: “Incorporating guidelines for health assistance into a socially intelligent robot”, The 15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN06), Hatfield, UK, September 6-8, 2006

[Maekawa2009] Maekawa, T. et al., 2009. MADO interface: a window like a tangible user interface to look into the virtual world. In *Proceedings of the 3rd International Conference on Tangible and Embedded Interaction*. Cambridge, United Kingdom: ACM, pp. 175-180. Available at: http://portal.acm.org/citation.cfm?id=1517704 [Accessed December 23, 2009].

[McKay1999] McKay, E. N. (1999), *Developing User Interfaces for Microsoft Windows*, Microsoft Press

[Mori1970] Mori, Masahiro (1970). Bukimi no tani the uncanny valley. Energy, 7, 33–35. (In Japanese)

[Munoz2009] Cecilia Vera-Muñoz, Mercedes Fernández-Rodríguez, Patricia Abril-Jiménez, María Fernanda Cabrera-Umpiérrez, María Teresa Arredondo, and Sergio Guillén, “Adaptative User Interfaces to Promote Independent Ageing”, Springer, Volume 5615/2009, Book title: Universal Access in Human-Computer Interaction, Intelligent and Ubiquitous Interaction Environments, Pages 766-770, Computer Science, Springer, July 14, 2009.

[Mutlu2009] Bilge Mutlu, Fumitaka Yamaoka, Takayuki Kanda, Hiroshi Ishiguro, Norihiro Hagita (2009): Nonverbal Leakage in Robots: Communication of

Intentions through Seemingly Unintentional Behavior, HRI'09, March 11–13, 2009, La Jolla, California, USA.

[Olsen2003] Olsen, D., and Goodrich, M., Metrics for evaluating human-robot interactions. In Proc. NIST Performance Metrics for Intelligent Systems Workshop, (2003). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.71.4292&rep=rep1&type=pdf

[Raskin2000] The Humane Interface, J. Raskin, New Directions for Designing Interactive Systems. Reading Mass: Addison-Wesley, 2000

[Rekimoto2001] Rekimoto, J., Ullmer, B. & Oba, H., 2001. DataTiles: a modular platform for mixed physical and graphical interactions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM New York, NY, USA, pp. 269-276.

[Roy2000] N. Roy, J. Pineau, and S. Thrun. Spoken Dialogue Management using Probabilistic Reasoning. In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, 2000.

[Schaefer1999] Care-O-bot: A System for Assisting Elderly or Disabled Persons in Home Environments, C. Schaeffer, T. May, In Assistive technology on the threshold of the new millenium. IOS Press, Amsterdam, 1999

[Shi2005] Shi, H. & Bateman, J. "Developing Human-Robot Dialogue Management Formally", In Proc. of the Symposium on Dialogue Modelling and Generation, 2005.

[Sitter1992] Sitter, S., Stein, A., "Modeling the illocutionary aspects of information-seeking dialogues", Information Processing and Management 28, (1992) 165-180.

[Sverinson2003] Severinson-Eklundh, K. and Green, A. and Huttenrauch, H., Social and collaborative aspects of interaction with a service robot, Robotics and Autonomous Systems, Special issue on socially Interactive Robots (2003)

[Takanashi2007] Atsuo Takanishi, Yu Ogura and Kazuko Itoh: Some Issues in Humanoid Robot Design, Department of Mechanical Engineering, Waseda University, Japan, 2007

[Tsukada2002] Tsukada, K. & Yasumura, M. (2002). Ubi-Finger: Gesture Input Device for Mobile Use. In: Proceedings of the APCHI 2002, 1, pp. 388 – 400.

[Terrenghi2007] Terrenghi, L., 2007. Designing Hybrid Interactions through an Understanding of the Affordances of Physical and Digital Technologies. Available at: http://edoc.ub.uni-muenchen.de/7874/ [Accessed October 29, 2009].

[Vilhjalmsson2007] Hannes Vilhjalmsson, Nathan Cantelmo, Justine Cassell, Nicolas E. Chafai, Michael Kipp, Stefan Kopp, Maurizio Mancini, Stacy

Marsella, Andrew N. Marshall, Catherine Pelachaud, Zsofi Ruttkay, Kristinn R; *The Behavior Markup Language: Recent Developments and Challenges*, 2007;

[Waldherr2000] A Gesture-Based Interface for Human-Robot Interaction, S. Waldherr, R. Romero, S. Thrun, Kluwer Academic Publishers, 2000

[Walker1998] Marilyn A. Walker, Jeanne C. Fromer, and Shrikanth Narayanan. 1998. Learning optimal dialogue strategies: a case study of a spoken dialogue agent for email. In Proc. ACL/COLING'98

[Want1999] Want, R. et al., 1999. Bridging physical and virtual worlds with electronic tags. In *Proceedings of the SIGCHI conference on Human factors in computing systems: the CHI is the limit*. Pittsburgh, Pennsylvania, United States: ACM, pp. 370-377. Available at: http://portal.acm.org/citation.cfm?id=303111 [Accessed January 11, 2010].

[Welie2000] Welie M. and Trætteberg H., 2000, Interaction Patterns in User Interface, *The Proceedings of 7th Pattern Languages of Program Conference*, Illinois, USA http://jerry.cs.uiuc.edu/plop/plop2k/proceedings/Welie/Welie.pdf

[Williams2005] J.D. Williams, P. Poupart, and S. Young. Factored Partially Observable Markov Decision Processes for Dialogue Management. In 4th Workshop on Knowledge and Reasoning in Practical Dialog Systems, 2005.

[Williams2006] J.D. Williams and S. Young. Scaling POMDPs for Dialog Management with Composite Summary Point-based Value Iteration (CSPBVI). In AAAI Workshop on Statistical and Empirical Approaches for Spoken Dialogue Systems, July 2006.

[Yanco2002] Yanco H. A., and Drury J. L., 2002, A Taxonomy for Human-Robot Interaction, AAAI Fall Symposium on Human-Robot Interaction, AAAI Technical Report FS-02-03, pp. 111 – 119.http://www.cs.uml.edu/~holly/papers/yanco-drury-taxonomy-fss02.pdf

[Yanco2004] Yanco H. A., and Drury J. L., 2004, Classifying Human-Robot Interaction: An Updated Taxonomy, IEEE International Conference on Systems, Man and Cybernetics, Vol 3, pp 2841 – 2846 http://ieeexplore.ieee.org/iel5/9622/30423/01400763.pdf?tp=&isnumber=&arnumber=1400763

[Young1990] Sheryl Young. 1990. Use of dialogue, pragmatics and semantics to enhance speech recognition. Speech Communication, 9(5-6), Dec.

[Zhang2001] B. Zhang, Q. Cai, J. Mao, and B. Guo. Spoken Dialog Management as Planning and Acting under Uncertainty. In Proceedings of Eurospeech, 2001.