



Deliverable 3.1

"Report on first experiments on different modalities of interaction with companions"

Contract number: **FP7-215554 LIREC**

Living with Robots and intEractive Companions

Start date of the project: 1st March 2008

Duration: 54 months

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 215554.



Identification sheet

Project ref. no.	FP7-215554
Project acronym	LIREC
Status & version	[Final] "D3.1"
Contractual date of delivery	30 th of November 2008
Actual date of delivery	
Deliverable number	D3.1
Deliverable title	"Report on first experiments on different modalities of interaction with companions"
Nature	Report
Dissemination level	PU
WP contributing to the deliverable	WP3/WP2
WP / Task responsible	WP3/T3.1.1
Editor	Ana Paiva
Editor address	INESC-ID / Instituto Superior Técnico - Tagus Park Av. Prof. Dr. Cavaco Silva, 2780-990 Porto Salvo, Portugal
Author(s) (alphabetically)	Ginevra Castellano (QM), Ricardo Chaves (INESC-ID), Luísa Coheur (INESC-ID), Secundino Correia (CNOTINFOR), Pedro Cuba (INESC-ID), João Dias (INESC-ID), Martin Diruf (UB), Sibylle Enz (UB), Marco Estanqueiro (CNOTINFOR), Marta Gacsi (EOTETO), Mário Rui Gomes (INESC-ID), Mathias Jacobsson (SICS), Mariusz Janiak (WRUT), Eniko Kubinyi (EOTETO), Iolanda Leite (INESC-ID), Sara Ljungblad (SICS), Carlos Martinho (INESC-ID), David Matos (INESC-ID), Adam Miklosi (EOTETO), Robert Muszynski (WRUT), Diana Oliveira (CNOTINFOR), Luís Oliveira (INESC-ID), Ana Paiva (INESC-ID), André Pereira (INESC-ID), Guilherme Raimundo (INESC-ID), Krzysztof Tchon (WRUT), Isabel Trancoso (INESC-ID), Michael Walters (UH), Andreas Wichert (INESC-ID), Marek Wnuk (WRUT). Checkers: Nik Gaffney (FOAM), Dorothee Loziak (QM)
EC Project Officer	Pierre-Paul Sondag
Keywords	Communication, perceiving the user, affective sensing, modalities of expression, experiments.
Abstract (for dissemination)	Overview of the different types of modalities for communication with companions from two perspectives: perceiving the user and expressions in the companion, and report on the first experiments with companions.

CONTENTS

1	Introduction	4
2	Communication with Social Robots and Social Virtual Agents	5
2.1	Perceiving the user	5
2.2	Expressions in Social Robots and Virtual Agents	15
3	Experiments with real companions	26
3.1	Human-dog experiments	26
3.2	Human-human experiments	30
4	Experiments with artificial social agents and robots	35
4.1	Robot-house experiments	35
4.2	Pleo experiments	37
4.3	iCat experiments	38
4.4	Mozart experiments	44
4.5	AIBO experiments	46
5	Concluding remarks and future work	53
6	References	54
7	Appendices	62
7.1	“Pleo – first contact” experiment	62

1 Introduction

One of the main goals of LIREC is to develop ground research that will help and promote the building of artificial companions. To do that, the communication between the user and the artificial companion needs to be considered and deeply researched. In general we will consider that this communication involves the generation of expressive behaviour of the companion as well as the perception of the expressive behaviour from the user. Indeed, we believe that social robots and virtual characters must have the ability of perceiving the user, respond, react and express the appropriate behaviour, so that interactions with humans can be more believable, natural and enjoyable, and lead to long-term relations.

The aim in WP3 is to value the social behaviour of the companion avoiding the shallow, often pre-canned or reactive approaches that have been taken quite often. Furthermore, we will look at the communication by focusing on the speech, facial and body interactions. As such, a set of research questions drive the development in this workpackage:

1. Visual human affect analysis by combining face and body gestures and analysis of speech
2. Selecting appropriate expressive behaviour for the current embodiment and the current interaction with the current users
3. Integration of non-verbal behaviour with domain-specific language and speech capabilities
4. Integration of expressive behaviour with task-related behaviour

Further, as a companion may take different forms, e.g. it may present itself as a robot, a mobile handheld device, a graphical virtual character etc, these different forms, will determine the modalities, forms and semantics of such communication. Furthermore, aware of the large dimension of the problem at hand, in this workpackage we will focus primarily on the communication functions that are associated with long term relations, and thus will support the development of such relationships.

However, to address these goals we needed to understand and explore some basic aspects of the communication with real and artificial companions, in order to focus the research during the rest of the project. Thus, to set up the scene, we have conducted a set of experiments with real and some preliminary artificial companions, in order to understand some specificities of the communication with them.

This deliverable reports those preliminary experiments, describing some of the main issues and findings. We expect that the results we obtained will impact forthcoming research in LIREC.

This deliverable is organised as follows. In the next Chapter we will make an overview of the different types of modalities for communication with companions and look at it from two perspectives: perceiving the user and expressions in the companion. This overview allows us to understand the techniques and issues related to the perception and expression in our companions. In Chapter three we present two experiments conducted with people and “real” companions (dogs and humans). These experiments were conducted to find out particularities of the relations established between humans and their “real” companions. Chapter four goes a bit further and presents some very preliminary experiments with “artificial” companions (in this case we have explored only robotic companions with the “Pleo”, AIBO, iCAT and the Robot-house experiments). Although different in the research questions addressed, these experiments focus on trying to find out some particularities of the communication between users and artificial “companions” and extract some elements important to drive the research on companions. Finally, in Chapter five some conclusions and drawn taking into account the experiments performed.

2 Communication with Social Robots and Social Virtual Agents

Communication with our artificial companions will involve two different aspects:

- Perceiving the user's expressions and trying to understand them in the context of the task at hand;
- Generating expressive behaviour by the companion.

These two facets lead to a communication loop that we aim at achieving with our companions. In both of these two aspects, different modalities can be used. In LIREC we will focus on speech, natural language; facial expressions and body expression.

This Section reviews some fundamental work in these areas focusing primarily in works that may impact directly or indirectly our development of the artificial companions.

2.1 Perceiving the user

Affect sensitivity refers to the ability of analysing the verbal and non-verbal behaviour of users in order to understand their affective states.

Most previous studies focused primarily on the design of systems able to recognise basic emotions (e.g., joy, sadness, disgust, surprise, fear and anger), and were largely based on acted affective expressions (Zheng et al., to appear). While few studies have so far addressed the problem of finding methods for inferring more complex states, the design of artificial companions requires an affect sensitivity which goes beyond the ability to recognise prototypical emotions, and is able to capture spontaneous and more variegated affective signals conveying more subtle states such as, for example, boredom, interest, frustration, agreement, willingness to interact, etc.

Affective expressions are multimodal. An important issue in relation to affect sensitivity is the need for multimodal systems that are able to analyse different modalities of expression, as well as to fuse different channels of information in order to achieve a better understanding of the affective message communicated by the user. In the next Sections an overview about the state of the art in affect recognition from face, body gesture and speech is provided, and some examples of multimodal systems are reported.

The specifications of the verbal and non-verbal behaviours that our companions will be sensitive to, as well as of the systems to analyse such behaviours, will depend on the design of the specific scenarios for user-companion interaction and will be defined step by step along with the test of technical issues and requirements in such scenarios.

2.1.1 Facial Affect Recognition

Research in psychology extensively investigated the type of messages conveyed by the face and many researchers share the belief that facial expressions communicate emotions and affective states (see, for example, Ekman & Friesen, 1969; Ambady & Rosenthal, 1992). Given the relevance of affective communication through facial expressions during everyday life, the potential applications of automatic analysis of facial expressions are, not surprisingly, numerous. In the research on artificial companions, the ability for a robot or a virtual agent to be able to detect affective signals conveyed by the face is of extreme importance.

Automatic analysis of facial expressions can be used to detect human affect at different levels. Affect can be either detected directly from changes in the facial expressions or can be inferred after recognition of facial muscle actions (or action units) is performed (Pantic & Bartlett, 2007). Several systems were proposed in the literature for the recognition of facial actions units and for the detection of facial affect (see, for example, surveys by Pantic, 2006;

Pantic & Rothkrantz, 2000). These systems use either geometric facial features or appearance-based facial features. Geometric features are used for the detection and tracking of facial characteristic points. Vukadinovic and Pantic (Vukadinovic & Pantic, 2005), for example, developed a fully automatic facial point detector using Gabor wavelets and GentleBoost-based point detectors. Gabor filters, Haar-like filters and learned image filters (e.g., from principal component analysis) are examples of appearance-based features. Many systems were developed for facial affect recognition based on appearance features (see, for example, Anderson & McOwan, 2006; Littlewort et al., 2006). Anderson and McOwan developed an automatic system for real-time recognition of facial expression which can operate effectively in dynamic scenes and is robust to head motion. The system extracts facial motion from frontal views of facial expressions and is capable of recognising six basic emotions.

2.1.1.1 FACS - Facial Action Coding System

Facial Action Coding System (FACS) (Eckman & Friesen, 1975; Eckman & Friesen, 2002) is a common standard to systematically categorise the physical expression of emotions. The FACS model of emotions provides a description of human emotions on the basis of Action Units (AUs) which represent the deformations of selected regions of the face caused by shrinking or relaxing of particular muscles. Image processing techniques have been proposed for extraction of some selected features of the human face in order to automatically recognise AUs and thus the human facial expression of emotions.

The FACS system is frequently used in many projects dealing with facial expression recognition (Kapoor et al., 2003; Kanade & Cohn) and face animation (Filmakademie; Valve). It has been used in many socially interactive robots as a model for facial expression of emotions (Head; Fredslund; Orlov) FACS defines 32 AUs (Action Units), which are a contraction or relaxation of one or more muscles. Several examples, particularly useful for the basic emotions coding:

- AU1 Inner Brow Raiser – Frontalis (pars medialis)
- AU2 Outer Brow Raiser – Frontalis (pars lateralis)
- AU5 Upper Lid Raiser – Levator palpebrae superioris
- AU6 Cheek Raiser – Orbicularis oculi (pars orbitalis)
- AU7 Lid Tightener – Orbicularis oculi (pars palpebralis)
- AU10 Upper Lip Raiser – Levator labii superioris
- AU12 Lip Corner Puller – Zygomaticus major
- AU20 Lip stretcher – Risorius / platysma
- AU25 Lips part – Depressor labii inferioris or relaxation of Mentalis, or Orbicularis oris
- AU26 Jaw Drop – Masseter, relaxed Temporalis and internal pterygoid

Facial expression is coded by assigning several AUs as its label. In general, the labelling of expressions requires trained experts. Some efforts have been taken to design a vision system for specific face features extraction (Namysl, 2008a; Namysl, 2008b). It will be used to automatically identify FACS codes, and thus quickly identify emotions.

The basic emotions can be described by the following combinations of AUs (Filmakademie; Vanger et al.):

- Happiness 6 + 12 + 25
- Sadness 1 + 4 + 15
- Disgust 4 + 10 + 17

- Anger 4 + 5 + 7 + 24
- Surprise 1 + 2 + 5 + 26
- Fear 1 + 2 + 4 + 5 + 20 + 25

For example, surprise is labelled by the following Action Units: Inner Brow Raiser (AU1) + Outer Brow Raiser (AU2) + Upper Lid Raiser (AU5) + Jaw Drop (AU26), which is reflected in the following features of the face image (see Figure 2.1):

- widely open eyes (AU5)
- raised eyebrows (AU1+AU2)
- open mouth (AU26)
- horizontal wrinkles on the forehead (AU1+AU2)

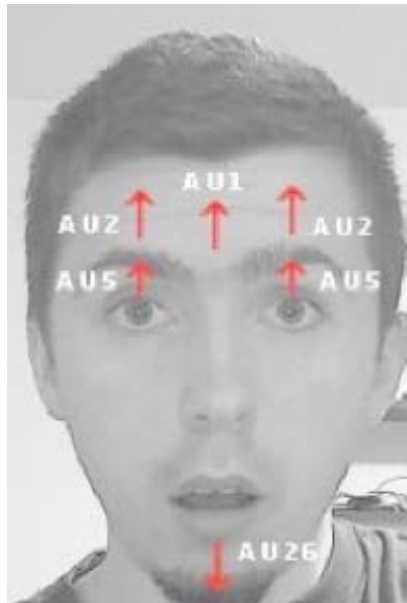


Figure 2.1 Action Units in case of surprise expression (Namysl, 2008a)

2.1.1.2 Selection of the face features

The comparison of different descriptions of facial expression of emotions (Darwin, 1872; Devon, 2006; Givens; Filmakademie; Valve) resulted in the observation that several face features are commonly used in most of them (Namysl, 2008a). The proposed vision system makes use of the following features:

- wrinkles on the forehead
- opening of the eyes
- position and shape of the eyebrows
- shape of the lips

The features are described by the following numerical parameters:

- number of horizontal wrinkles in the centre of the forehead
- eyebrows bend and declination angles
- relative distance between the eyelids
- relative distance between pupils and eyebrows

- aspect ratio of the lips
- relative positions of the lips corners
- relative area of the visible teeth

2.1.1.3 Implementation issues of the vision system

The vision system (Namysl, 2008a) has been implemented with OpenCV library (Intel) in linux environment. The main processing stages are as follows:

- image acquisition
- face detection (main ROI)
- face sub-regions extraction (eyes, mouth, nose, forehead)
- face features parameterization

The image acquisition stage should allow to detect the face on a wide image of the robot neighbourhood. The resolution of the face ROI must be reasonably good for the further processing. The envisioned (not implemented yet) solution is using a PTZ camera and implementing an active face tracking system.

Face detection based on Haar classifier (Reimondo) resulted in very precise (but time-consuming) ROI definition. A small trade-off in the precision allows for a considerable increase of the processing speed. The method based on skin colour detection (Huang & Chiang, 2006) is much faster, but requires that there are no objects of the skin-like colour in the field of view in order to avoid erroneous hits.

Detection of the face sub-regions have been implemented both with Haar classifiers and a simple, but effective method of horizontal and vertical projections (Chen et al., 2005). The examples are presented in Figure 2.2.

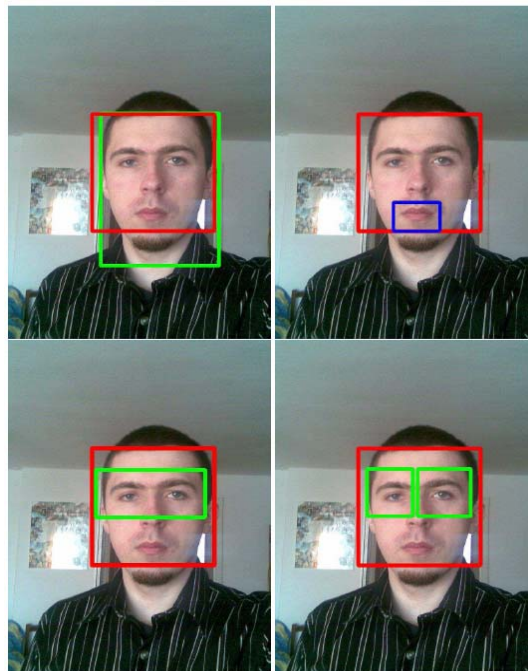


Figure 2.2 ROI detection of the face and its sub-regions (Namysl, 2008a)

In the last processing stage the face features are extracted in appropriate sub-regions and the numerical parameters are computed. Several image segmentation methods have been used:

- dual thresholding (eyebrows, teeth)
- pattern matching (eyes)
- morphological opening and propagation (eyebrows, lips)
- Hough transform (eyebrows, pupils)
- Haar classifier (eyes, lips)

The results of this stage are presented in Figure 2.3.

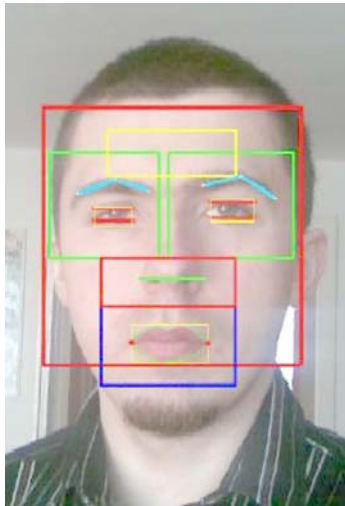


Figure 2.3 The face features detection (Namysl, 2008a)

2.1.2 Body Gesture and Posture recognition

Humans, while interacting with each other, continually communicate their feelings, thoughts, intentions and ideas through body movements. Different body movements can be associated with different purposes. Body movements, for example, can be used, consciously or unconsciously, to communicate feelings and emotions and may accompany important points in social interaction. Gestures of this type can be called expressive or *emotional gestures* (Cowie, 2007). Facial expressions, postures, hand gestures, and full-body movements characterising an affective, emotional state belong in this category. This Section provides an overview of the role of gestures in affective, emotional human-human and human-machine interaction. Psychological studies on emotional gestures are reported, as well as computational models and approaches adopted in computer vision aiming to recognise affect from body gesture and movement.

2.1.2.1 Psychological studies on movement and emotion

Several studies from psychology focus on the relationship between emotion and movement qualities, and investigate expressive body movements (see, for example, Boone & Cunningham, 1998; Boone & Cunningham, 2001; De Meijer, 1989; Pollick et al., 2001; Wallbott, 1998). While some studies have found evidence for characteristic body movements accompanying specific emotions (e.g., Wallbott, 1998), others argue that movements may be only indicative of the intensity of emotion, but not of its quality (Ekman and Friesen, 1974).

Wallbott (Wallbott, 1998) analysed body movements and postures of actors portraying different emotions, elicited via a scenario approach. Results showed some distinctive patterns of movement and postural behaviour associated with some of the emotions studied. When specific patterns did not emerge, few distinctive features that allow for reliable distinction between emotion categories were found. For example, lifting the shoulders seems to be typical for joy and hot anger while moving the shoulders forward is frequent for disgust, as well as for despair and fear. Further, anger and joy appeared to be characterised by high movement activity and dynamics and expansive movements.

Other researchers have mainly investigated the relationship between emotions and movement qualities. De Meijer (De Meijer, 1989) found relationships between emotion categories and amount, type and weight of movement features in a rating study where a group of subjects evaluated videos with body movements performed by three actors. Boone and Cunningham (Boone & Cunningham, 1998) identified six expressive cues involved in the recognition of four basic emotions (anger, fear, grief, and happiness) and further tested the ability of children in recognising emotions in expressive body movement through these cues. These six cues were “frequency of upward arm movement, the duration of time arms were kept close to the body, the amount of muscle tension, the duration of time an individual leaned forward, the number of directional changes in face and torso, and the number of tempo changes an individual made in a given action sequence”. Based on previous findings (Boone & Cunningham, 1998; De Meijer, 1989), Boone and Cunningham (Boone & Cunningham, 2001) demonstrated children’s ability to encode the emotional meaning of an underlying music excerpt by moving a teddy bear to indicate one of four emotions (happiness, sadness, anger, fear). Results showed that children characterised their expressive movements with respect to force, rotation, shifts in movement pattern, tempo, and upward movement: children used more force, rotation and shifts in movement patterns and faster, more upward movement to encode happiness and anger more than fear and sadness.

While these studies show that it is possible to associate characteristic movement qualities with specific emotions, another area of research focuses on whether it is possible to discriminate affect observing the same movement performed in different ways depending on the underlying emotion. Pollick et al. (Pollick et al., 2001), for example, investigated perception of affect from point-light animations of arm movement in everyday actions (e.g., drinking, knocking, etc.) and found a significant correlation between the movement kinematics (arm velocity) and the activation axis in the two-dimensional space characterising affect with respect to the dimensions of activation and valence as proposed by Russel (Russel, 1980). In particular, greater arm velocities appeared to be associated with higher activation levels.

2.1.2.2 Machine analysis of movement and gesture expressivity

Machine analysis of expressive movement can be useful for recognising affective, emotional state of users.

Camurri et al. (Camurri et al., 2005) proposed a layered conceptual framework to model the analysis and interpretation of human movement expressivity. This framework is based on a layered approach ranging from low-level physical measures (e.g., position, speed, acceleration of body parts, etc.) toward descriptors of overall gesture features (e.g., motion fluency, impulsiveness, etc.) and high-level information describing semantic properties of gestures (meaning, affect, emotion, expressiveness, etc.). Camurri et al. also developed the EyesWeb Expressive Gesture Processing Library (Camurri et al., 2004), which contains a collection of software modules running in EyesWeb XMI (Camurri et al., 2007) for the automated analysis of human movement and extraction of expressive motion cues.

Other studies focus on analysis of hand and arm gestures expressivity. Zhao (Zhao, 2001) used an approach based on “Laban Movement Analysis” to acquire motion qualities of

communicative gestures. Motion features were extracted from live performances using motion capture and video-based techniques and neural networks were trained to estimate the relationships with the four dimensions of Laban's Theory of Effort (Laban & Lawrence, 1947). Caridakis et al. (Caridakis et al., 2006) proposed a framework for video-based tracking and automated extraction of expressivity parameters of hand gestures. This approach includes the parameters Overall activation, Spatial Extent, Temporal, Fluidity, Power/Energy and Repetitivity, inspired from the expressive behaviour module implemented by Hartmann et al. (Hartmann et al., 2005) to synthesise expressive gestures in the Greta Embodied Conversational Agent.

2.1.2.3 Affect recognition by body movement and gesture analysis

While psychologists have stressed the importance of the visual channel carrying information about body movement and gesture for people to successfully recognise human affect, most of the studies on affective computers have focused on automated affect recognition based on facial expressions and audio data. Nevertheless, some attempts in the research community were made towards the design of systems capable of analysing expressive body movement and using this information to recognise affect.

Studies on automated emotion recognition from body movement and gesture analysis can be distinguished into a number of separate categories. First of all, some studies investigated emotions conveyed by gestures, while others have focused on emotions communicated by postures. As it will be shown in more detail in the following text, they differ in terms of gesture acquisition (some of them use video-based analysis, others motion-captured data, and others data collected from different types of sensors), features used (e.g., low-level indicators such as points' coordinates or high-level dynamic features), number of gesture and posture classes analysed (i.e., one or more different classes used to express different emotions), and the presence of gesture recognition before prediction of emotions.

Camurri et al. (Camurri et al., 2003) classified expressive gestures in dance performances using motion cues extracted using video analysis techniques. They found relationships between emotions communicated by the full-body movement of dancers and motion cues such as quantity of motion and contraction/expansion of the body.

In further studies, Volpe (Volpe, 2003) used decision trees to automatically recognise emotions in dance fragments using statistical measures (e.g., average, standard deviation, etc.) applied to the temporal profile of several motion cues (e.g., quantity of motion, contraction/expansion of the body, kinematical indicators, etc.).

Castellano et al. (Castellano et al., 2007; Castellano, 2008) proposed a computational approach for affect recognition based on the expressivity of human movement. In this approach expressive motion cues (e.g., movement qualities such as the quantity of motion, fluidity, etc.) are considered as a source of affective information. They proposed a model that defines features retaining information about the temporal dynamics of expressive motion cues is proposed. The approach was tested on two corpora of affective expressions portrayals and results showed how features describing temporal aspects of the expressivity of human movement are successful in the discrimination of affect.

Bernhardt and Robinson (Bernhardt & Robinson, 2007) proposed a framework for the recognition of affect in knocking motions from a motion-captured database. They computed motion cues (e.g., velocity and acceleration of the hand, etc.) over motion primitives obtained using an approach based on segmentation by motion energy and clustering of motion segments. Support Vector Machines were then trained for each motion primitive using statistical measures of the extracted motion cues as features.

Kapur et al. (Kapoor et al., 2005) used full-body skeletal movements' data obtained with a technology based on the Vicon motion capture system to automatically classify four emotional states. They acquired different gestures for each of the emotions considered and

3D positions of fourteen reference points were recorded. Further, they trained different classifiers using statistical measures of kinematical indicators (e.g., velocity and acceleration) computed for each reference point.

Other studies show that the integration of body movement with other modalities (e.g., facial expressions) increases the performances of automated emotion recognition systems.

Balomenos et al. (Balomenos et al., 2005) used gestures to support the output of the facial expression analysis in a bimodal emotion recognition system. Their approach is based on Hidden Markov Models, and uses the position of the centroid of the head and hands, computed using video-based techniques, to recognise gesture classes. Recognised gestures are then associated a posteriori with specific emotions based on predefined mappings.

El Kaliouby and Robinson (El Kaliouby & Robinson, 2005) proposed a vision-based computational model to infer mental states from head movements and facial expressions. Their approach is based on Hidden Markov Models for the real-time recognition of head and facial actions and on Dynamic Bayesian Networks to model mental states over time.

Gunes and Piccardi (Gunes & Piccardi, 2007) fused facial expression and expressive body gesture information at different levels for bimodal emotion recognition. They found that using expressive body information improves the accuracy of the emotion recognition based on face only. As for recognition based on body gesture, several types of gestures associated with different emotions were classified using image features (e.g., the centroid and area of the upper body, the centroid and orientation of the hand, etc.) computed with video analysis.

Valstar et al. (Valstar et al., 2007) combined multimodal information conveyed by facial expressions, head and shoulders movement to discriminate between posed and spontaneous smiles.

Castellano et al. (Castellano et al., 2008) proposed a framework for multimodal emotion recognition in which facial expressions, body gesture and speech data is fused at the feature and decision level to predict eight emotional states in a speech-based interaction.

Other researchers investigated the relationships between posture and affective dimensions.

Mota and Picard (Mota & Picard, 2003) for example showed how sequences of postures can be used to predict affective states related to a child's interest level during a learning task performed with the computer. They collected postures' data through pressure sensors mounted on a chair and they used Hidden Markov Models to predict the affective state related to sequences of postural behaviour.

Bianchi-Berthouze and Kleinsmith (Bianchi-Berthouze & Kleinsmith, 2003) proposed a model that can selforganise postural features into affective categories to provide robots with the ability to incrementally learn to recognise affective human postures through interaction with human partners.

Other studies have used Mixed Discriminant Analysis to identify nuances of affective states (De Silva et al., 2005) and affective dimensions (Kleinsmith & Bianchi-Berthouze) starting from low-level descriptors of human postures.

2.1.3 Speech Recognition and understanding

Spoken language understanding has two basic components: the automatic speech recognition (ASR) module and the natural language understanding (NLU) module. These modules may be highly integrated or quite separate, if an off-the-shelf recognition system is adopted. The next Section covers the state-of-the-art in NLU systems. The current Section addresses the ASR module. See Rabiner (Rabiner, 1989) and Young (Young, 2002) for excellent overviews on this topic.

ASR is the task of finding the most likely set of words for a given acoustic signal (Gilbert, 2008). This probabilistic model can be represented as computing $\arg \max_W P(W | a)$, where W is a string of words and a is a set of features that are extracted from the acoustic signal. Most often, these features encode the spectral characteristics of the signal, the most typical being the cepstrum and energy along with their first- and second-order time derivatives.

The basic approach to speech recognition is to apply Bayes' rule to convert the problem into computing $\arg \max_W P(a | W) P(W)$, where $P(a | W)$ corresponds to the acoustic model, representing the probability of the acoustics given the word string, and $P(W)$ is the language model, representing the probability of the string of words that are under consideration.

$P(W)$ is typically modeled as a Markov process. For a string of N words, the joint probability can be expressed as $P(W) = P(w_1 w_2 \dots w_N) = P(w_1) P(w_2 | w_1) \dots P(w_N | w_1, \dots, w_{N-1})$, which may be simplified as an n -gram model by truncating the history to $n - 1$ words.

One can distinguish two basic approaches: the recognition of isolated words, with a small vocabulary, and the recognition of continuous speech, with a large vocabulary. In the first scenario, a different acoustic model is built for every word in the vocabulary (typically less than 200 words). The most frequent approach for training the acoustic models is hidden Markov models (HMM). When a large vocabulary is needed, however, it becomes much more efficient to train models for each sub-word unit (very frequently context-dependent phones), and to model each word as a sequence of sub-word units. This lexical model can also admit multiple pronunciations (sub-word sequences) for each word.

Competitive acoustic models can also be built using hybrid models which combine the temporal modelling capabilities of HMMs with the pattern matching capabilities of artificial neural networks (ANN).

For relatively simple command and control (C&C) applications, isolated word recognizers (IWR) of limited vocabulary may seem like the most adequate solution. This solution, however, has a major limitation. It needs a specific database for training, with several examples of each word to be recognized, typically recorded in the same environment, and therefore limits the expansion to new words. For dictation or broadcast news subtitling systems, on the other hand, the use of large vocabularies continuous speech recognizers (LVCSR) with sub-word acoustic models is obviously the most adequate solution. An intermediate solution is keyword spotting (KWS), which aims at recognizing a keyword in the middle of a continuous audio stream. KWS approaches are broadly classified into two categories (Szöke, 2005): one based on the acoustic match of the audio data with keyword models in contrast to a background model, and the other one based on LVCSR. The acoustic solution can be based on either word or sub-word models. The LVCSR solution typically takes advantage of the lattice of recognized words, containing several hypothesis in parallel, thus allowing improved performances compared to searching in the raw output result.

The last component of the ASR system, besides the feature extraction stage, and the acoustic, lexical and language models, is the decoder, which searches through all possible recognition choices. The Viterbi algorithm is one of the most common approaches for decoding.

ASR systems can be trained for a specific user (speaker-dependent) or a large variety of users (speaker-independent). One can also start with a speaker-independent recognizer and adapt its acoustic models to a specific speaker. The performance of ASR systems is much better for speaker-dependent or adapted systems than for speaker-independent systems.

The performance is also strongly dependent on the type of input. For read speech, nowadays, one can obtain a very satisfactory performance for a number of applications. For spontaneous speech, however, the performance seriously degrades. This may be attributed to the frequent presence of hesitations, repetitions, filled pauses, fragments and other disfluences that characterize spontaneous speech and which are not present in the large

quantities of text that are typically used for training n-gram language models, and also to the articulation style and pronunciation variability, which bring additional challenges in terms of acoustic and phonological modelling.

A final consideration in this necessarily brief review is the performance of recognizers under noisy environments. This performance can be critically dependent on the presence of noise, especially if the system has not been trained for these conditions.

Several scenarios for ASR can be envisaged in the LIREC project. For some scenarios such as the robot house, the C&C isolated-word recognizer may be appropriate. Even in the simplest scenario, however, one must take into account that the recognition system must be activated in order to recognize keywords. This activation may be done by pronouncing a specific activation keyword (such as the robot name) which is spotted in the continuous input stream. Alternatively, the user may adopt a push-to-talk strategy. This strategy is very effective for PDA interactions, where the user may use the pen to signal beginning and end of speech. Better performances can be expected with the use of this type of interaction, than with the use of activation by specific keywords.

Much more sophisticated scenarios can be contemplated in this project, especially if one allows combined recognition and understanding modules. The combination of multiple knowledge sources may be done at different levels. In this context, the use of the confidence scores produced by the recognition module may be quite effective.

2.1.4 Natural Language Understanding

Natural Language Understanding (NLU) is the process of converting natural language samples into a data structure the computer can deal with. Typically, this structure is a logical form or a frame, but it can also be an utterance in natural language that the computer already understands. Although some systems implement hybrid approaches, involved techniques may classify NLU as:

- Basic NLU;
- Linguistically Motivated NLU;
- Statistical NLU.

Basic NLU include keyword detection, pattern matching and the use of simple algorithms capable of associating new input to already understood utterances. The classical example of a system based on pattern matching is ELIZA (Weisenbaum, 1966), invented by Joseph Weizenbaum in the early 1960's, aiming at emulating a psychologist. Although the obvious limitations of this approach, when ELIZA was released many people believe it to be human. Some psychologists even thought that ELIZA could be used in place of a real psychologist, proving that in some particular situations a pattern matching approach should not be neglected.

Linguistically Motivated NLU uses some level of linguistic information. Typically, systems implementing this paradigm base their performance on a syntax/semantics interface, where each syntactic rule is associated with a semantic rule and logical forms are generated in a bottom-up, compositional process. Variations of this approach are described in (Allen, 1995) and (Jurafsky & Martin, 2008) and implemented, for instance, in systems such as Edite (Reis et al., 1997). This was a conventional natural language interface, developed in the mid 95's in order to be integrated in a multimedia kiosk and answer questions about touristic resources. Like most of the systems following this approach, besides the need of linguistic information, Edite had the following problems:

- new syntactic rules could easily lead to unexpected semantic values;
- syntactic elements, significant for semantic analysis, could be spread in different syntactic rules;

- rules causing over-generation or errors were difficult to identify due to the recursive character of the whole process.

In what concerns Statistical NLU, there are many techniques being explored (see, for instance, [Bhagat et al., 2005] or [Leuski et al., 2006]). Some of these techniques came from the Machine Learning framework. In order to illustrate the approach, consider (Bhagat et al., 2005), where four of these techniques are applied to a small training set constituted by 477 sentence/frame pairs, and compared. Considering that each frame is a set of attribute/value pairs, the first technique computes a model that represents the probability of producing a certain attribute/value pair being given a particular n-gram as input; both second and third techniques cast NLU as a classification task. Maximum Entropy is used in one experiment and Support Vector Machines in the other. Finally, in the last technique a language model for the attribute/value pairs is estimated allowing to build the frame for a given utterance as the set of the most likely pairs. Results from this evaluation run from a 0.75 F-score to a 0.83 F-score. It should be noticed that these results derived from the fact that the domain was limited and that it was possible to develop a kind of named entity recognizer, that replaced each entity of the domain by its correspondent class name.

Considering LIREC scenarios, in order to choose the NLU approach to implement, it will be taken into account:

- a) the type of the recognizer output;
- b) the linguistic variation of the domain utterances.

If the recognizer returns keywords, the only possibility is to perform keyword detection. If the recognizer returns utterances involving a larger vocabulary, several approaches can be adequate. Nevertheless, the fact that the NLU module will have to deal with the error-prone output from the speech recognition module, adds a new complexity to the already hard task of NLU, and narrows the hypothesis. For instance, the pattern matching approach is no longer a good choice, as many unexpected utterances can be returned by the recognizer.

Our first experiments regarding interaction with companions used the Basic NLU approach. The experiments were conducted using “Duarte Digital”, a conversational agent that answers questions about Custódia de Belém, a famous work of Portuguese jewellery. When “Duarte Digital” receives a question, it searches for the most similar question in the knowledge base. This search is based on the Levenshtein distance algorithm applied to words. First results show that Duarte is able to correctly understand around 45% of the input received, although in an extrinsic evaluation 65% of the utterances result in a possible answer (for instance, “I don’t know” or “Could you please repeat what you have just said?”). Even if several improvements can be done to this technique, the fact is that Duarte operates in a very limited domain, which leads us to conclude that systems operating in a larger domain should invest in a different approach.

Being so, statistical methods like the ones presented above, are a good solution. However, one should also have in mind that these solutions involve the creation of a training corpus and that the linguistic variation of the domain utterances will decrease the capacity of producing good results.

2.2 Expressions in Social Robots and Virtual Agents

While perceiving the user is extremely important for attaining some degree of responsiveness in our companions, the generation of expressive behaviour is also fundamental for the communication loop between user and companion to be established. In this Section we will focus now on the expression of companions.

2.2.1 Facial Expressions

Human face is the most important channel for non-verbal communication. There are approximately 600 muscles in the human body (it is difficult to say the exact number because different anatomists group muscles differently), and about 50 of them are located in the face (Standring, 2004), which makes this part of our body extremely expressive. The face therefore plays a central role in our social lives by providing an efficient, and frequently honest, way of communication (Ekman, 1997). Among other things, it serves to coordinate social interactions and to express emotions.

Regarding social interaction, our facial expressions can, for instance, give out cues that let the other interlocutor know that we are waiting for a reply or that we did not understand what s/he said (Keltner, D. & Kring, 1998). From facial expressions we can infer motivations and what action someone is about to carry out. The range of information we convey with our face is so vast that the brain developed dedicated mechanisms for its recognition during mankind's evolution. Since an early age, infants can recognize familiar faces.

As for emotions, our face can show the world what we are feeling. But this ability can also betray us when we try to lie: facial expression of emotions involve involuntary muscle actions that people cannot produce when they are not really experiencing that emotion (Oatley *et al.*, 2006). The human brain is so well trained in the task of facial expression recognition that we are able to tell even the subtlest of changes in a face. Research shows that some basic emotions (happiness, sadness, surprise, anger, disgust and fear) can be universally distinguished through facial expressions, even when they are being expressed by people from a different culture (Ekman & Friesen, 1971).

Given the relevance of facial expression in human expressivity, it is no surprise that researchers devote so much attention to it when developing embodied characters. However, due to its complexity and the brain's natural ability to detect false expressions, creating convincing artificial expressions is not an easy task. Nonetheless, over the years many techniques and systems have been created to not only enhance the final result but also to facilitate the process of achieving it. The remaining of this Section presents some relevant work in this field, both for virtual agents and social robots.

2.2.1.1 Facial expression in virtual agents

Pandzic & Forchheimer (2003) argue that the field of computer facial animation is "inhabited by two kinds of people: the researchers and the artists". While artists aim at producing high quality animations for specific films or games, researchers are more interested in technical aspects, studying mechanisms that can be applied widely rather than to produce good effects in a particular case.

Given the recent advances in computer graphics, facial animations in movies are close of achieving photo-realism (see for example the facial animations produced by Image Metrics¹), and highly believable characters (Pixar movies are a good example). In movies, the end result is not subject to real-time constraints, and therefore 3D animators can make use of complex graphical techniques that yield a high-quality outcome. However, this focus on the outcome usually leads to an ad hoc solution for the particular facial animation of a given character. This one time solution means that none or little work from that developed solution can be reused in the future.

The gaming industry tends to follow a similar approach to the movie industry except that it is bound to real-time restrictions. Each character usually has a library of fixed animations which can be played efficiently within the game. Therefore, interactions between the character and

¹ <http://www.image-metrics.com/>

the user tend to be restricted, being limited to a combination of previously determined behaviours.

As we depart from scripted environments to more emergent scenarios with embodied agents, higher demands are set upon facial animation systems. The increased autonomy in agents is generally no longer satisfied by predetermined and fixed animations. Here, the animation controls are as important as, or in some cases more important than, the final graphic result.

To overcome this challenge, over the years deformation techniques were developed to model facial expressions. These techniques can be broadly separated in two categories: image manipulation, which usually consists of changing texture images (cf. Pighin, 1998) and geometry manipulation, which tries to model the anatomical nature of facial muscles and the human skin (cf. Parke & Waters, 1996). At the same time, many researchers have been focused on creating parameterization standards that allow developers to generate facial expressions by manipulating certain parts of the face independently. Parke (1974) was the first to introduce parameterization models, but many others followed him. Paul Ekman's FACS (described earlier in Section 2.1.1.1) and MPEG-4 (Pandzic & Forchheimer, 2003) are the most common parameterizations used nowadays in virtual characters.

Embodied Conversational Agents (ECAs) provide a good case of non-scripted interactions. ECAs try to mimic the communication channels used by humans when engaged in a conversation, being one of them facial expression. Due to the rich and complex nature of such task it is required that the ECA possess a structured control over its face. This must not only provide precise and accurate control of the face but also grant a meaningful interface for the processes that manipulate it.



Figure 2.4 GRETA displaying different facial expressions

GRETA (Pasquariello & C. Pelachaud, 2001) is an ECA that follows MPEG4 animation standard. It is a 3D animated character that can talk to users while displaying facial expressions and body movements (see Figure 2.4). Many experiments have been performed using this agent. Recently GRETA was used to evaluate user's perception of expressivity of complex facial expressions, gaze movements and gestures (Bevacqua *et al.*, 2007), to study strategies of the politeness theory by analyzing if subjects can infer social context from GRETA'S facial expressions (Niewiadomski & Pelachaud, 2007) and to evaluate facial expressions in terms of empathy (Niewiadomski *et al.*, 2008).

As one can see from the diversity of studies using GRETA there are benefits in having a standard parameterization. For instance, one of the advantages is the variety of facial expressions that developers can produce in short periods of time. Nevertheless, some ECAs still use predefined animations. This happens mostly in agents more focused in dialogue, such as REA, a "Real Estate Agent" (Cassel, 2000) that plays the role of a real estate salesperson that interacts with users (potential buyers) to determine their needs and attempts to sell them a house. REA is capable of speech with intonation, facial display and hand gestures (synchronized with the speech acts). Its behaviours, a mix of facial expressions, body gestures and speech are employed to coordinate social interaction. For example, in the beginning of the interaction REA looks at the user and smiles, when she

wants to give turn in the conversation she raises her eyebrows, and nods her head to give feedback to the user.

2.2.1.2 Facial expression in social robots

There is still much more to explore in facial expressions of social robots than in virtual agents perhaps due to the difficulty in obtaining good hardware platforms that will allow for facial expressions. Indeed, robots with some expressivity in its “face” only started to appear about a decade ago and they are usually very expensive. Consequently, there are no standards in this field yet, and it is difficult to control and manipulate those “faces”. This means that little work can be extensible to other robots. Moreover, whereas in virtual agents we can already simulate almost all the muscles of the human face in a very realistic manner, in robots this is still a difficult task.

Due to the reasons mentioned above, the first robotic faces that appeared were very simplistic and contained few degrees of freedom. One of the first ones was Felix (Figure 2.5), a robot built from commercial LEGO Mindstorms (Cañamero & Fredslund, 2000). The robot was developed to explore believable emotional exchanges and credibility in human-robot interaction. Felix reacts to tactile stimulations expressing emotions through its face. Its degrees of freedom are very limited: there is only one degree on the eyebrows and three on the lips. Felix’s facial expressions are based on approximations (due to the restrictions in degrees of freedom) of Action Units from FACS. Despite its simplicity, in a study performed to evaluate how well humans recognize the facial expressions displayed by the robot, good results were obtained, especially on recognizing expressions such as anger, happiness and sadness. The study also revealed that humans tend to empathize with Felix in a natural way on spontaneous interactions.

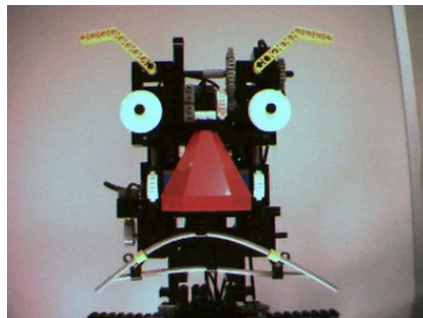


Figure 2.5 Felix, the LEGO robot.

One landmark in expressive and engaging robots with the ability of displaying facial expressions is Kismet (Breazeal, 2002). Kismet was developed with the purpose of engaging people in face-to-face interactions. Its face contains 15 degrees of freedom, which allows it to display social signals such as emotional expressions and gaze. Each ear has two degrees of freedom, each eyebrow can lower and furrow, elevate or slant, and its eyelids can open and close independently. Kismet also has four lip actuators, one at each corner of the mouth, and a single degree of freedom jaw. Kismet’s facial expressions are generated using an interpolation-based technique over a three dimensional space (arousal, valence and stance). Figure 2.6 shows different emotive expressions that Kismet can display to users through its face.

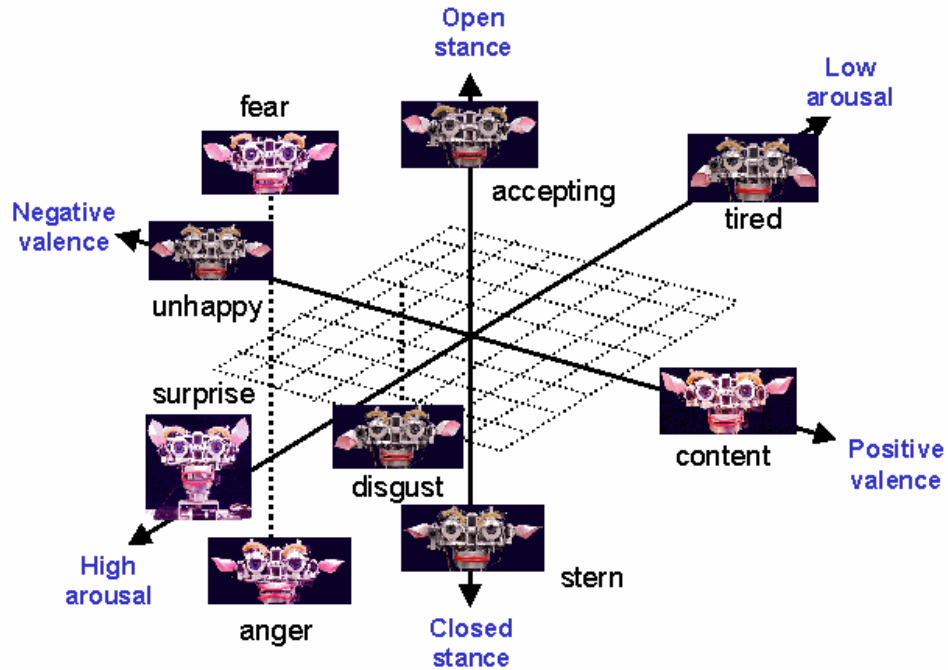


Figure 2.6 Kismet displaying different emotive expressions.

More recently Philips released the iCat Research Platform, a platform for researchers to study human-robot interaction. It includes a user-interface robot called iCat and a software platform called Open Platform for Personal Robotics (OPPR) which enables the rapid development of new applications for the iCat. The cat appearance was chosen because it is familiar to humans. The robot is 38 cm tall and is equipped with eleven RC servos and two DC motors that control different parts of the face such as the eyebrows, eyes, eyelids, mouth and head position. With this setup the iCat can generate several facial expressions such as happiness, sadness, surprise or disgust (Figure 2.7).

The iCat is one of the robots that will be used in the LIREC scenarios. In fact, several studies were already performed by a LIREC partner using this robot. One of the studies (Leite *et al.*, 2008) was performed to evaluate the effects of the robot's affective behaviour (which was displayed to the user using facial expressions) in the user's perception of the chess game. The results indicated that, when displaying the emotional reactions in agreement to a developed emotion model, user's perception of the game increased. Also a long-term experiment (described in Section 4.3 of this document) was recently conducted by INESC-ID using the iCat robot.



Figure 2.7 iCat exhibiting several facial expressions.

Until now we have presented robots with faces that reassemble some aspects of the human face. However, robotic faces with higher levels of anthropomorphism have already been constructed. Impressive work in such robots (usually named androids) comes from Hanson Robotics² and Osaka University's Intelligent Robotics Laboratory lead by Hiroshi Ishiguro. The latter developed the android face depicted in Figure 2.8.



Figure 2.8 One of the android faces developed by Hiroshi Ishiguro.

Several researchers support the idea that emotional expressions of social robots are abstractions of human expressions and therefore robots must be as human-like as possible (Duffy, 2003). On the other hand, as we may see from the studies with Felix, Kismet or the iCat, total anthropomorphism is not always necessary. In fact, it might have negative consequences such as users' frustrated expectations and lack of credibility towards the robot. Mori's theory of uncanny value (Mori, 1970) explores this issue. He argues that people's acceptance of robots gradually increases as realism increases, but only up to a certain limit. When crossing this limit, even the smallest imperfections frighten people and elicit sensations of strangeness.

Then why many researchers devote their attention into building androids? Some argue that "if androids are more likely to fall in the uncanny valley than mechanical-looking robots, the reason may be that our brains are processing androids as human" (MacDorman & Ishiguro, 2006), and therefore androids are the only type of artificial character capable of eliciting in humans the exactly same kind of social responses that human-human interaction does. Therefore, this type of robots will play an important role in cognitive science research.

From the studies presented so far, it is clear that facial expression does impact human-robot interaction, even with low degrees of naturalism. In LIREC, we will extend this work into longer term interactions and investigate the role of companion expressive behaviour in creating a long-life relationship.

2.2.2 Body Expression

2.2.2.1 Body expression in virtual agents

Several computational models of gesture synthesis have been described in the literature for the design of human-like virtual agents (see, for example, Cassell et al., 1994; Cassell et al., 2001; Chi et al., 2000). Nevertheless, generating "natural" movements remains a difficult task. Gestures in virtual agents, in fact, often appear awkward and synthetic and do not provide realistic behaviour.

Kipp et al. (Kipp et al., 2007) proposed a data-driven approach to gesture synthesis for virtual agents focusing on gesture units (i.e., the excursions starting from the rest position and lasting until the release of movement (Kendon, 2004)). They show that using gesture

² <http://www.hansonrobotics.com/robots.html>

units increases the naturalness of the virtual agent, since it allows for the production of smooth and fluid streams of gestures. A key problem in gesture generation is making agents express themselves in order to demonstrate emotional behaviour. The synthesis of gesture expressivity is of great importance for the design of affective virtual agents, that is, agents endowed with emotional behaviour.

Synthesis of gesture expressivity

Adding expressivity to the performance of gestures of an agent implies both the generation of expressive movements and the attachment of coherent motion qualities. Techniques for the generation of expressive movements can be divided into four categories (for a survey of the approaches taken in literature, see Zhao, 2001). One approach consists of adding expressivity to neutral motions, using methods such as Fourier function models or signal processing. Bruderlin and Williams (Bruderlin & Williams, 1995), for example, manipulated neutral motions, applying multiresolution filtering techniques used in the signal processing domain. In this approach, motion parameters are treated like sampled signals: by using filters with different settings, movement can be exaggerated or constrained. Other approaches consist of making the motion fit some constraints, adding secondary movements to the original movement of virtual agents and generating and controlling behaviours. A different problem is to explore which motion qualities must be added to movement to make it convey emotional, expressive information.

The same gesture performed with different motion qualities can convey different expressive content and be perceived in different ways. Some researchers have started to put efforts into the design of computational models of gesture expressivity based on the synthesis of motion qualities. Chi et al. (Chi et al., 2000) built the EMOTE (Expressive MOTion Engine) system, a computational model to generate movements for the torso and the limbs of a virtual agent to create communicative gestures that convey expressivity. The EMOTE system allows one to synthesise gestures based on Effort and Shape components from LMA. Specifically, EMOTE is “a 3D character animation system that allows for the specification of Effort and Shape parameters to modify independently defined arm and torso movements” (Chi et al., 2001). Zhao (Zhao, 2001) designed a computational framework for gesture acquisition and representation to be used with the EMOTE system. Gesture representation is conducted based not only on motion forms, but also on Effort and Shape elements, and it is used to generate gestures by manipulating its motion qualities.

Hartmann et al. (Hartmann et al., 2005) implemented a module for the animation of expressive arm gestures based on six expressivity parameters: Overall Activation (i.e., the quantity of movement during a conversational turn), Spatial Extent (i.e., the amplitude of movement), Temporal Extent (i.e., the duration of movement), Fluidity (i.e., the degree of smoothness of movement), Power (i.e., the energy of movement) and Repetition (i.e., the degree of rhythmic repeats of movement). These parameters act on the gesture stroke and can vary along a scale ranging from 0 to 1 (where 0 corresponds to the absence of movement) for the Overall Activation and from -1 to +1 for the others (where zero corresponds to the movement performed in absence of expressivity modulation). Caridakis et al. (Caridakis et al., 2006) proposed a framework to synthesise gesture expressivity using expressivity parameters automatically extracted from real video sequences and inspired from the model proposed by Hartmann et al. (Hartmann et al., 2005).

Martin et al. (Martin et al., 2005) proposed a copy-synthesis approach for the creation of affective Embodied Conversational Agents (ECAs). This approach can be summarised in two steps. First, a real-life non-acted emotional corpus is manually annotated at different levels. Multimodal behaviour and displayed emotions as well as their temporal evolutions are encoded. At this stage, the animation of the ECA is performed at the face and body levels using the annotation performed during the first step as input.

Castellano and Mancini (Castellano & Mancini, In press) modelled a bidirectional communication between an ECA and a user based on movement and gesture. They

designed a real-time system for analysis and synthesis of emotional gesture expressivity. The system is capable of acquiring input from a video camera, processing information related to the expressivity of human movement and generating expressive copying behaviour: for each gesture performed by the user, the agent responds with a gesture that exhibits the same quality, i.e., the same movement expressivity. The system is formed by the integration of two different software platforms: EyesWeb XMI (Camurri et al., 2007) for video tracking and analysis of human movement, and the Greta ECA for behaviour generation (Pelachaud, 2005). A mapping between the expressive motion cues analysed in humans and the correspondent expressive parameters of the agent is defined, so that the agent is capable of generating gestures with the same expressivity as those performed by the user.

2.2.2.2 Body expression in robots

Body expression in robots is still an under-explored research topic. Nevertheless, note that the methodology for synthesis of gesture expressivity described in the previous Section may be applied to robots as well.

Some attempts towards the design of robots capable of performing expressive movements are reported in the literature.

Kismet (Breazeal et al., 2001) is a sociable robot endowed with expressive behaviour. Expressive features displayed by the robot include facial expressions and neck and eye orientation. The latter is important for displaying expressive postures and directing the robot's cameras towards external, relevant stimuli.

Mertz is an active-vision head robot which can display expressive behaviour by using facial expressions and head and neck movements (Aryananda & Weber, 2004).

Shibata et al. (Shibata et al., 2003) proposed an approach for avoidance planning of robots sensitive to human emotions. Shinozaki and colleagues (Shinozaki et al., 2007) built a robot system in which a humanoid robot is capable of performing dance based on the concatenation of short dance motions. Maeda and Tanabe (Maeda & Tanabe, 2006) proposed a method based on Laban's Theory for evaluating emotional behaviour in a pet-type robot endowed with bodily motion. Nakata et al. (Nakata et al., 2001) proposed a set of features based on Laban Movement Analysis to explain impression produced by a robot's bodily expressions.

Robotic companion hand gestures

Due to the abilities of the human hand, the hand gesture is the most numerous category of body expressions. Regarding their semiotic function, one can give the gestures classification distinguishing command gestures, co-verbal gestures, and sign language gestures (S. Marcel, 2002). Within first two categories symbolic, emblematic, and illustrator gestures can be associated to verbal communication channel. Looking for a compact collection of revealing, cross-cultural hand gestures, representing the verbal channel associated gestures, one can end up with the following gestures set suitable for a robotic companion:

- greeting (rising the opened right hand with the palm directed toward a person)
- farewell (waving up and down the opened right hand placed horizontally)
- warm greeting, disapproval (moving the opened hands sideways with the palms directed up)
- stop (moving the flat right hand ahead with the palm directed toward a person)
- come (waving the fingers of the right hand placed ahead with the back directed toward a person)

- go away (waving the opened right hand placed ahead with the back directed toward a person)
- pointing to objects, directions (with index finger or all fingers)
- silence (rising the right index finger to the mouth)
- thinking, listening (placing the closed right hand with the thumb up in front of the jaw)
- OK (moving the closed right hand ahead with the thumb up)
- tapping forehead (with index finger)
- rejecting, prohibiting (the right index finger up moving left and right)
- shaping of simple imagined objects (hands tracing out simple curves – a circle, a rectangle.)

Contemporary robotic hands and arms constructions are extremely expensive and difficult to control, thus they are no suitable for fast prototyping experiments. Moreover, they outperform gesticulation task requirements creating a need for designing a simple robotic tool aimed to gesticulation. Hence, to experiment with robot hand gestures perception and their temporal synchronisation with the co-expressive part, and the rhythm of speech, a unique hand design has been elaborated (Tchon et al., 2008.)

The construction (see Figure 2.9a) is characterised by the following parameters:

- complete arm with hand,
- bearing joints,
- 5 high torque and high speed digital servos (Dynamixel), 3 lightweight micro servos,
- trajectory tracking, predefined gestures,
- Physical parameters: total length 60cm, weight 1kg,
- DOF: arm 5, hand 3,
- Carrying components: carbon fibre tubes, aluminium elements,
- Control system: distributed, PC/104 based, with limited force control,
- Communication: internal RS485, external Ethernet powered by YARP,

The designed arm consists of 2 links connected via a single 1DOF joint (Figure 2.9b). The arm is to be screwed to a robot body via 3DOF joint (Figure 2.9c) and is endowed with another 1DOF joint, to which a hand can be mounted. The hand (Figure 2.9d) is formed of four 1DOF fingers. The thumb and the index finger are driven by two separate microsensors. The other two fingers are driven by one, shared microsensors.

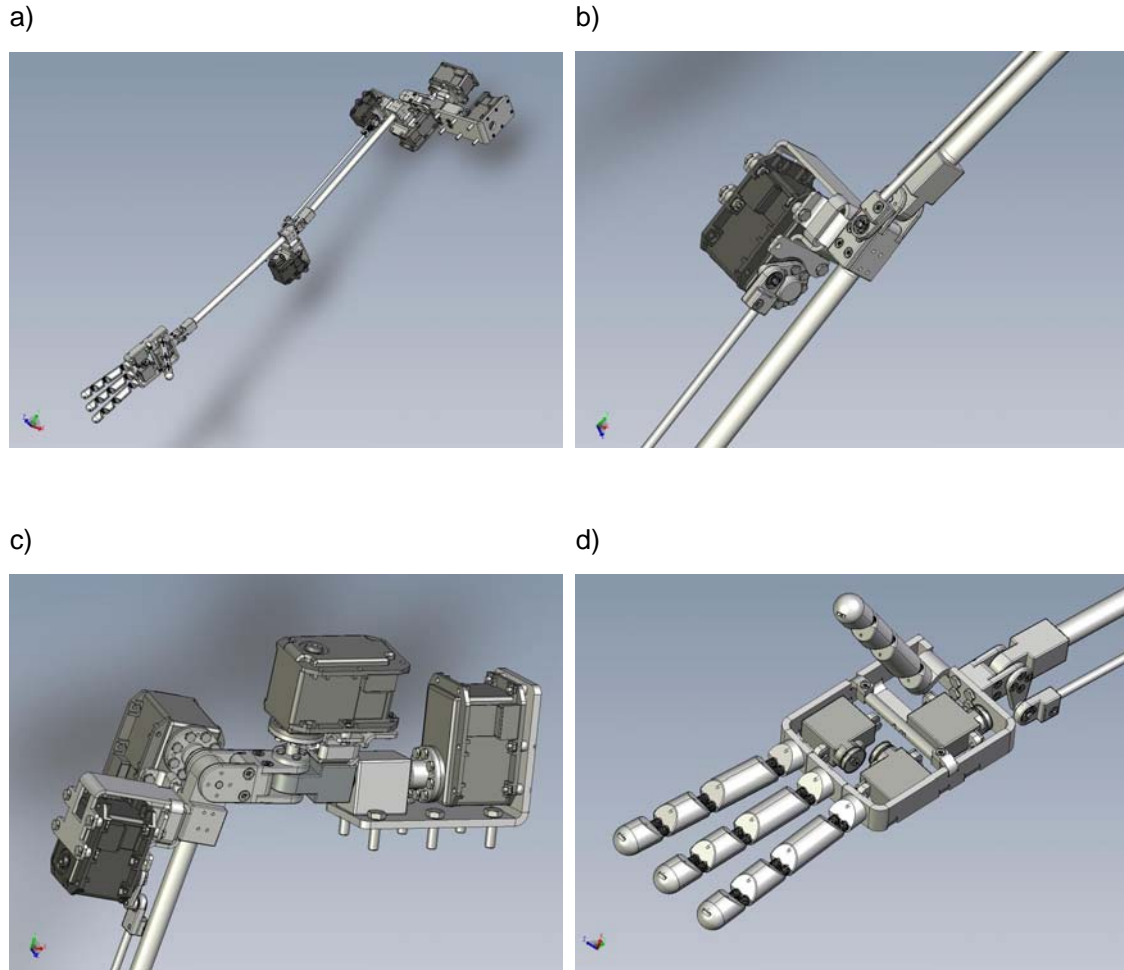


Figure 2.9 Design of robotic hand and arm for gesticulation

2.2.3 Speech and Natural Language Generation

Natural language generation (NLG) is the process of producing natural language from, in general, non-linguistic information. Generation can be defined according to the complexity associated with the processes underlying the production of the final output. These systems may go, in shallow approaches, from simple “canned” text, suitable for simple interactions, to template-based generation, which allows for some sophistication. Shallow approaches, nevertheless, may be unsuitable for general language production, where more fluent communication is required: these situations may warrant the cost of the so-called deep generation, which involves issues ranging from content determination, to referring expression generation, to grammar-based language production. It should be noted that some template-based approaches may also be used as stages in deep generation.

There have been engineering efforts in order to define a general/common architecture for NLG systems (Reiter, 2000; Mellish, 2004). These architectures define a three-stage pipelined architecture which handles the following tasks: document planning (content and structure), text/sentence planning (micro-structure), and surface realization (format and presentation). Actual implementations may vary, depending on the approach being followed at each step in the pipeline (e.g., deep vs. shallow approaches), and on the breakdown of each component in the architecture (depending on the specific task, there may be more to

do at some points). For instance, StoryBook (Callaway, 2001), while mostly following the base architecture, features components for handling application-specific matters (narrative generation).

Considering NLG and speech synthesis, the first task corresponds roughly to deciding what to say and the second to how to say it. Regarding the possible LIREC platforms, and considering that some of them may exhibit restrictive hardware features (memory, processing power), the NLG task may have to be scaled down, so that the machine is able to produce speech in reasonable time (a few seconds). Deep approaches to NLG may take several minutes to produce output, although, of course, this depends on the complexity of the planning behind what is being said. Thus, for LIREC, and as a first approach, either pre-defined “canned” text (or even pre-recorded speech), or template-based approaches are appropriate. The speech synthesiser may further process the output of the NLG component, in order to satisfy its own requirements.

Depending on the specific platforms, on the available information, and on the specific interaction requirements, the NLG part of the interaction could be extended to, included in, or replaced by, a dialog system. Current examples include systems used in projects involving human-robot interaction, such as COMPANIONS (Hakulinen, 2008) or JAST (Foster, 2008), or general dialog/interaction frameworks, such as CMU’s Olympus (Bohus, 2007) or Collagen (Rich, 2001; Sidner, 2005).

Regarding speech output, several approaches can be taken. The simplest one is to use a standard text-to-speech (TTS) system with an available software development kit like the ones provided by CereProc (Andersson, 2008) or by Loquendo (Balestri, 1999), for example. However, the limitations of general-purpose TTS for human-computer dialogs are now clear. Much subtlety and complexity of meaning in natural language dialogs is conveyed by prosody; how something is said is often as important as what words are spoken. Current TTS systems cannot handle the specific prosodic and expressive features required to convey the most common speech acts required in dialogues (Syrdal, 2008):

- Imperative: directive, request, wait, repeat, warning
- Interrogative: question-why, question-yes/no, question-multiple choice
- Assertive: informative-general, informative-detail
- Affective: apology, exclamation-positive, exclamation-negative, greeting, good-bye, thanks
- Others: confirmation, disconfirmation, back-channel, cue phrase

This is mostly due to the technology behind the systems: the concatenative based systems require that the inventory includes examples of utterances conveying the required speech acts (Strom, 2006) and the HMM based speech synthesis systems require examples of such utterances in the training databases (Krstulovic, 2007). Another difficulty common to both approaches is in how to encode the information regarding the speech acts so that it can be used during the voice building and speech synthesis processes.

Given that the focus of the Lirec project is not on developing a full fledge expressive text-to-speech synthesizer we will mostly rely on the use of currently available synthesis systems. In previous projects we have addressed the issue of expressiveness in synthesized speech by using a limited domain speech synthesizer, that is, a synthesizer with a restricted vocabulary (Weiss, 2007). In this case that approach is unfeasible given the requirement of a more flexible system. For this reason a hybrid approach is being followed: a limited-domain approach for common utterances associated with speech acts and an unrestricted domain synthesizer for the remaining sentences. In order to have the same voice for both approaches, the limited domain voice is in fact produced using the unrestricted domain system together with an acoustic transformation to modify its prosodic and expressive features (Kawahara, 2008).

3 Experiments with real companions

In this Section we will describe some experiments conducted with real companions focusing on understanding the types of modalities used in the communication between companions, and what are the specific characteristics that constrain and support that communication.

3.1 Human-dog experiments

Most dogs and dog owners do not need special training in order to develop and maintain communicative interaction. Many experience this as an effortless, automatic process that just “happens”. This often encourages writers for the general public to overstate the dog-human relationship and argue for a co-evolution between the two species. The reality is less exciting because the communicative interaction is based on utilising homologue (similarity because of common ancestry) and analogue (similarity because of common selective factors) signals. The additional difference in comparison to other inter-specific communication between humans and another species is that during domestication dogs acquired skills that helps them to express human-like communicative pattern, and that dogs have a great flexibility in both decoding and emitting various types of signals. In dogs (and in humans) mutual learning about the other's communicative abilities plays a very important role. We think that this mutual learning process is a key to the development of idiosyncronic communication between human and dog and it gives the feeling to the human of having a unique relationship with that companion. Thus the design of robotic companion could draw many inferences from the communicative interaction between humans and dogs.

3.1.1 Human-initiated communication

Social understanding is defined as a complex cognitive process in which the subject is able to integrate contextual and social information, and modify his/her behaviour accordingly. In case of dogs, the owners' verbal commands accompanying gestural and contextual cues could operate as information which facilitates the understanding process.

3.1.1.1 Verbal communication

In humans talk is a dominant way of establishing social contact, thus it is not surprising to find that humans also talk to dogs (see Figure 3.1).

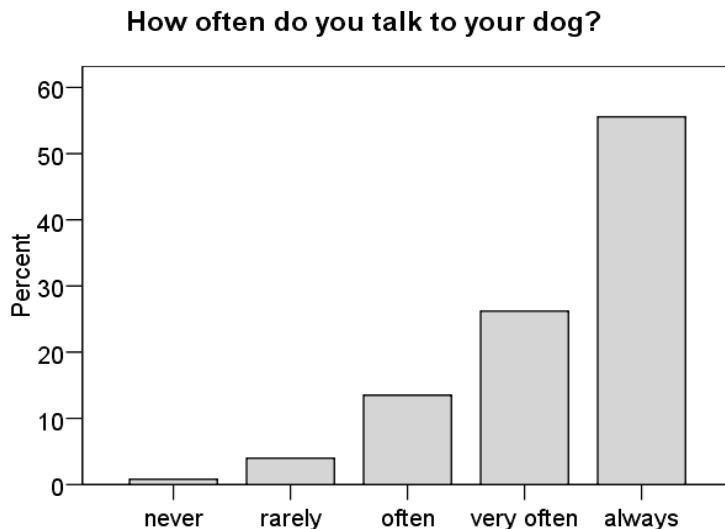


Figure 3.1 The graph represents the answers of 126 pet dog owners' to a questionnaire on human-dog communication (Mirkó et al., 2008. in prep)

The role of attention

Humans' preference to think in terms of language (our preferred but not only system used for communication) makes the analyses of our communication with non-human animals rather complex. Visual attention is an especially important feature of even verbal communication because it seems that dogs use visual cues (eye-contact and directed talk) provided by humans to infer whether they are "addressed" in a particular situation. Dogs rely on behavioural cues of humans to discriminate between attention and inattention, and are particularly sensitive to behavioural signals in commanding situations (Virányi et al., 2004). We have found that dogs learn much better in some social learning context if the human uses eye contact and verbal signals to get the attention of the dog (Pongrácz et al., 2004).

When do owners talk to their dogs?

Owners communicate verbally with their dogs in many different cases, even if they believe that their dogs do not (entirely) understand what they are talking about (see Figure 3.2).

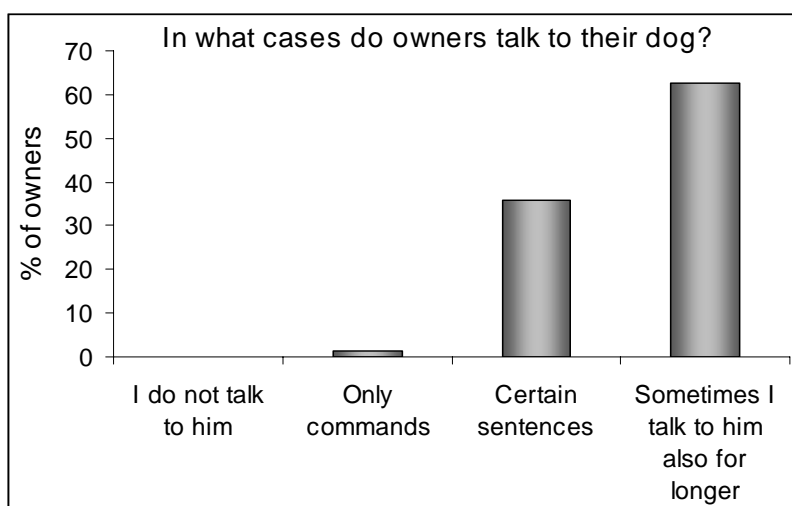


Figure 3.2 The answers of 140 owners provide convincing data that most dog owners talk to their pets as with human companions

How much do dogs understand?

We asked Hungarian pet dog owners to fill out a questionnaire about their verbal communication toward their dogs (Pongrácz et al., 2001). The 37 owners listed 430 different utterances (30 on average), which they thought their dogs knew. Utterances could be ranged into categories of actions: Invitation, Posture, Permission, Disallowance, Referring to object/person, Information giving, Question, and Unique. The age of the owners or dogs, breed of dogs, and the educational status of owners did not strongly affect the utterance structure. According to the owners the verbal communication an average dog understands consists of one-word (69.73 %), two-word (20.73 %), three-word (6.37 %), and four or more word (3.16 %) utterances. Many actions were believed to be executed only in adequate situations, supporting our idea that the communication between dogs and owners could be described as a form of social understanding. According to the owners' opinion, most actions of the dogs were executed only in contextually adequate situations, but around one third of the actions were reported to be performed by the dogs on command independent of the context.

Reaction of the dog	<i>M</i>	<i>SD</i>
Every time	31.80	± 9.78
Contextually adequate situations	51.47	± 16.45
Occasionally	16.86	± 12.88

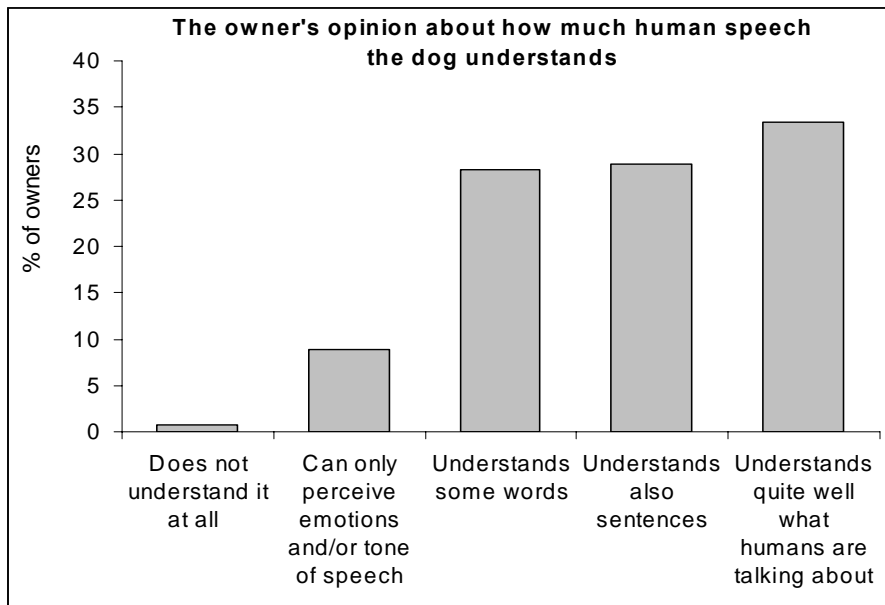


Figure 3.3 In an investigation of 500 dog-owner pairs we found that from the reactions of their companion many owners believe that dogs understand what is told to them

The style of speech

Investigations, based on the experience and views of dog owners, have found that in most families dogs are regarded as members with child-like rights. Affiliative aspects of dog-human relationship have been most often interpreted as a form of social attachment. Humans even seem to use a modified speech for verbal communication with the dog described as “doggerel” (Hirsch-Pasek & Treiman, 1981). Several similarities were observed to the so called “baby talk” which is used by mothers toward infants (using the speech register at higher frequencies, they talk slower and in simpler sentences, rely on a smaller vocabulary, express affection and talk also from the perspective of the infant). Most of these observations were supported in detailed comparison of “doggerel” and “baby talk” (Mitchell, 2001).

3.1.1.2 Visual (gestural) communication

From the behavioural point of view communicative interaction of visual signals can be divided into four stages. First, the sender produces signals for initializing the interaction, next it recognizes that the receiver is in the state to observe the signalling. This state, which often referred to as “attention”, offers the sender to send further signals, and finally the sender might receive a response from the receiver.

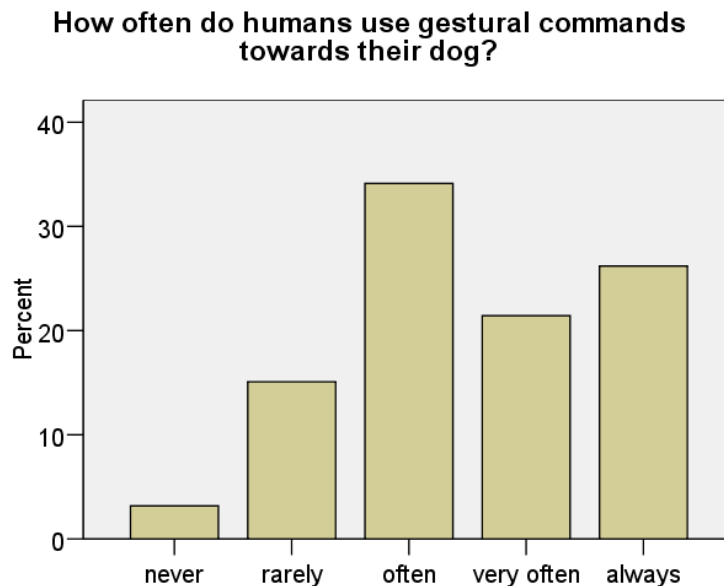


Figure 3.4 Represents the answers of 126 pet dog owners to a questionnaire on human-dog communication (Mirkó et al., 2008 in prep)

The role of attention

One general point in the case of dog-human interaction is that dogs seem “to live” in the visual field of the human. This means that the direction which is in the focus of the human becomes also significant for the dog.

In the case of visual signals attention can be recognized by the sensitivity to certain cues which reliably predict gaze direction and visual awareness (body and head orientation, open eyes etc). In a series of experiments we have found that dogs are sensitive to such behavioural cues signalling attentiveness (Gácsi et al., 2004). Dogs are able to discriminate humans' face orientation (forward or backward) because they approached the person mostly from the direction of the face when retrieving an object on command. Importantly, this sensitivity is context dependent since dogs show no such discrimination in the context of play but only if they are asked to perform a retrieving task.

Comprehension of human gestures

Dogs proved to be surprisingly skilful in the utilisation of directed human bodily signals. Even very young puppies can follow a variety of human communicative cues (Gácsi et al., 2008). This ability has been extensively investigated using a two-way object choice paradigm in which a human gives a social cue (such as pointing or gazing) to indicate the correct location of the target object. (Soproni et al., 2001; Soproni et al., 2002). Dogs are able to use most of these cues successfully, and their success appears to be based purely on the use of the social cues, as several controls have ruled out alternative explanations, including using odour as a cue (Szetei et al., 2003). They are also able to generalize to a certain degree from familiar pointing gestures to unfamiliar ones, and thereby they perform well on the basis of partially novel or “strange” pointing gestures (Lakatos, Soproni, Dóka, & Miklósi, 2008)

3.1.2 Dog-initiated communication

There are some indications that dogs have a strong propensity to initialise communicative interactions with humans by using visual and sometimes acoustic signals functionally similar to ones used by humans.

3.1.2.1 Visual signals

When facing an insoluble problem dogs often use such attention-getting behaviours. For example, after looking at the owner, dogs display gaze alternation between the location of a target object and the owner (Miklósi et al., 2000). A similar phenomenon was observed in a separate experiment (Miklósi et al., 2003), in which dogs, after having learnt how to solve a task, were prevented to get the target object the same way. Characteristically, after a few attempts most dogs stopped trying and looked at their owner.

3.1.2.2 Vocalisation

In contrast to wolves, dogs seem to bark invariably in a wide range of contexts. Barking has been often observed in dogs living with humans and relatively rarely in stray and feral dogs. Thus some researchers assumed that dogs use barking as a means for communicating with humans. It seems that dogs can vary at least three parameters of their bark (frequency, tonality = noise/harmony, pulsing) (Pongrácz et al., 2005). Humans need relative little experience for decoding the meaning of barking. Children from the age of 6 are able to report correctly the two basic emotions (aggressive versus fearful) involved in some situations (attacking versus left alone) (Molnár et al., 2008). In many respects human non-linguistic signalling is also in accord with the motivation-structural rules, thus we might be able to rely on this ability in decoding vocalisations of non-human communicational partners, like dogs.

3.2 Human-human experiments

3.2.1 Research-Intention

This pilot-study provides a basis for the further work on the concept of long-term-relationships between humans und artificial companions. It focuses on the beginning and maintaining of human-human-friendships in the context of university students in order to find (success-/ risk-) factors which might be applicable to human-companion-relationships.

The small scale study was primary based on the showcase “myFriend” (see showcases workpackage), where the companion myFriend was going to help students to cope with stress during their first weeks and months at the university.

3.2.2 Research Question

Following research-issues have been examined:

1. Motives: Which motives do the subjects report for getting involved in social relationships to other students at the beginning of their studies
2. Success-/ and risk-factors for building-up & maintaining a friendship
3. Identification of personality-factors influencing the friendship
4. Attitudes towards artificial companions: Identification of general attitudes towards human-companion-relationship

3.2.3 Study Design

With a questionnaire developed for the purposes of this study (= FRIENDQ) 15 subjects were explored regarding the above mentioned research questions. The theory-based questionnaire contains 68 items (e.g. “We got to know each other by a mutual friend of us” – Yes or No) and 19 open questions (e.g. “What aroused your interest in getting closer to this person?”).

Furthermore the subjects judged both their own and the student-friend's personality on the dimensions of extraversion, agreeableness, conscientiousness, emotional stability and openness (German version of the TIPI; Muck et al., 2007).

The sample consisted of students of several disciplines. In order to get as precise answers as possible, the students were asked to focus on just one of their close student-friends while responding the FRIENDQ.

For the data-analysis have been used as well quantitative and qualitative methods (Mayring, 2000).

3.2.4 Results

3.2.4.1 Quantitative data-analysis

Descriptive analysis of the subjects

The age of the 15 subjects ranged between 21 and 26 years ($M = 23.5$; $SD = 1.6$), nine of them were female, five male, one person gave no information. They were studying 6.4 semesters on average in different disciplines like e.g. psychology, computer sciences, sinology, and chemistry.

Descriptive analysis of the student-friendships

The subjects are in closer relationships to 4 – 10 fellow students ($M = 6.9$; $SD = 2.2$) of each gender (7 male and 8 female friends, respectively). The friendships are existent for an average of 26.3 month, thus showing a wide range (between 1 and 60 month; $SD = 18.7$). The students meet as well in university context (8.6 meetings a week on average; $SD = 8.0$) as in private context (7.4 meetings a week on average; $SD = 6.1$).

The beginning of the student – friendship The most friendships were initialized by the subject's attendance to the same classes at university (92.9%³), 28.6% got to know each other on a party. 14.3% of the subjects got to know their future friend due to common friends, and also 14.3 % come from the same region. The vast majority reported similar interests (71.4%), similar attitudes (78.6%), related study-interests (92.9%) and comparable attitudes towards their tutors.

According to the student's statements, most of the time both of the friends initiated the friendship (78.6%); furthermore they mainly characterised the friendship as balanced (71.4%).

The maintaining of the student – friendship As important success-factor for the maintaining of the friendship was again the similarity of the student's attitudes and values (92.9%), and analogue interests and hobbies (76.9%). Similar to the friendship's beginning, the student's majority still attends the same lectures (76.9%). 84.6% of the subjects share the same circle of friends. All subjects like working together with their friend in the university context, and they also all like to spend their leisure-time with his/her friend.

92.3% estimate their friend as popular and clever, and for the majority (84.6%) conversations with him/ her are very important. 92.3% are content with the friend's help.

³ multiple answers have been permitted

Personality-factors

Self / FRIEND	Extraversion	Agreeableness	Conscientiousness	Emotional Stability	Openness
Extraversion	.77**				
Agreeableness	–	.77**			
Conscientiousness	–	–	n.s.		
Emotional Stability	–	–	–	.63*	
Openness	–	–	–	–	.81**

Figure 3.5 Correlations among personality traits self-assessment and the assessment of the traits of the friend

Significant correlations between a person's self-assessment of his/her traits and the assessment of the friend's personality were found: Except for conscientiousness, the ratings for the self- and friend are associated in all dimensions (see Figure 3.5).

The relationships are indirectly influenced by a person's personality; the subjective interpretation of being similar, as far as personality traits are concerned, seems to be crucial.

3.2.4.2 Qualitative data-analysis

Motives for getting involved in a college-friendship

Why did the subjects become friends with a fellow-student? 11 of 15 subjects expected to "have a good time" with his/her friend or actually enjoyed the time they spent together. The second-most mentioned motive for getting involved in a college-friendship is the finding of a new friend (e.g. "not being alone"; mentioned 8 times). Further 8 subjects said they liked their potential friend very much and found him/ her likeable (e.g. "charisma"), another 7 persons mentioned the similarity between themselves and the other one ("e.g. "similar experiences").

Shared activities

What are the students doing when they spend time together? Most of the participants of this study mentioned social activities (e.g. "party", "clubbing"; mentioned 16 times) and collective sport activities (e.g. "climbing", "sailing"; "soccer"; mentioned 10 times). Furthermore the mutual cultural interest (e.g. "theatre"; "cinema", "music"; mentioned 7 times) seems to be an important shared activity.

Success-factors for a friendship

According to the subjects statements the friendship's success-factors are changing: While in the starting-time the above mentioned similarity (e.g. "affinity of nature") seems to be most important, mutual interests and investments into the friendship (e.g. "caring for the friendship", "trust") are getting more important the longer the friendship lasts (see Figure 3.6).

Success-factors/ beginning	Numbers mentions	of	Success-factors/ maintaining	Numbers mentions	of
similarity	18		mutual interest & investment	23	
mutual interest & investment	10		similarity	7	
social competences	7		frequent encounter/ proximity		
sympathy	7				

Figure 3.6 Success-factors for the beginning & maintaining of a friendship

Furthermore likeability and social competences like “*reliability*” and “*open-mindedness*” are helpful for the beginning friendship. Later, basic conditions like frequent encounter, and proximity become important factors.

Risk-factors

What are the risk factors for the maintenance of a student- friendship? The reason that was mentioned most frequently for the end of a student-friendship is the growing-apart of the friends (e.g. “*to lose sight of the other*”; mentioned 11 times). Furthermore, basic conditions (e.g. “*too less time*”, “*too great distance*”; mentioned 8 times) and an increasing one-sidedness (e.g. “*missing-engagement*”, mentioned 8 times) causes problems and the fading of the friendship.

Attitudes towards artificial companions

At the end of the questionnaire, subjects were asked whether they could imagine to use an artificial companion or even to get familiar with it.

Only two of the 15 participants could imagine developing a relationship to an artificial companion, the majority refused to interact with it at all (e.g. because „*I have enough human friends*“, and „*the robot is unable to identify the user’s emotions*“).

While this could be the attitude of an average person that never had contact to a robot or artificial companion before, it seems quite harsh. A possible explanation might be that the items on the human-robot-interaction should have been introduced more properly: maybe the subjects were answering the question while still focusing on human-human-relationships which is an inappropriate metaphor for the relationship to an artificial companion: The main argument of our participants against the robot is it’s disability of being human and respectively it’s “*missing of a soul*”.

3.2.5 Summary and future prospects:

The main *motive* for getting involved in a friendship among students seems to be the joy while spending time with the friend. Furthermore, likeability and the “need” for a new friend seem to be important.

In all phases of a friendship similarity seems to be the *main success-factor*. In later phases, mutual interest is getting more important, but still the persons’ affinity – even in their personalities - plays an important role. Well-balanced investments are crucial for the further development of a relation among students and nevertheless simple basic conditions (like the frequent encounter) can determine on success and failure of a friendship.

The mostly-mentioned *risk-factor* is the growing apart of the former friends, therefore permanent input into the friendship seems to be necessary.

The subjects, to a large extent, refused the notion to interact with an artificial companion. For further (questionnaire- or interview-based) studies, questions on human-robot-interaction should be introduced by a background-story, in which the robot does not compete with a human being, but stands for its own. This probably makes it easier for the subjects (or interview-partners) to see the potential benefit of a human-robot-interaction.

4 Experiments with artificial social agents and robots

4.1 Robot-house experiments

Exploratory experiments performed by the University of Hertfordshire team, mainly as work for WPs 6 and 8, have been aimed at informing the Robot House Showcase and associated scenarios. Therefore they have primarily examined users' perceptions of companions relating to migration and non-verbal behaviours. They have not specifically addressed issues of communication with companions, although some aspects have touched on verbal communication. There have been two experimental studies carried out to date aimed at informing the Robot House showcase:

The first exploratory experiment ran from the 5th to the 7th May, 2008, and used 160 participants from a school event "Take Part in the Future" held at UH for visiting schoolchildren (Figure 4.1). The study included the first set of HRI trials investigating perceptions of migration of companions and embodiment (WP4, 6 and 8). Results and data are currently being analysed, and full findings will be presented in the respective deliverable reports, D6.1 and D8.1.



Figure 4.1 Explaining the companion migration experiment to the participant schoolchildren

The second study on 22nd of July, 2008, was an initial exploratory Video-based HRI (VHRI) study which was designed for WP8 with the aim of studying issues related to the migration of companions. The VHRI study videos shown to participants illustrated three different visual indications of the companion migration process by using LED displays mounted on each of two robot embodiments (Pioneer and PeopleBot, see Figure 4.2). The aim was to explore participants' thoughts and feelings on the migration of robotic companions. Specifically, our objective was to understand participants' mental models of the migration process and determine what the key components are that help companions express (communicate) convincingly the migration process through non-verbal cues. The data is currently being analysed, and full details will be presented in WP8 deliverable report (D8.1)

A third study is currently under development. It is planned that this study will be a VHRI user trial, and the scenario is that of a user and guest interacting with a small mechatronic (Pioneer) robot. The robot will exhibit non-verbal behaviour inspired by dogs, and will be informed by a short pilot trial by EOTETO which will use real dogs. It is planned that the actual study will take place early in the 2009, after findings from the EOTETO pilot are assimilated and the trial video finished. More details will be presented in the WP6 deliverable (D6.1)



Figure 4.2 The LED displays indicate to the user that the companion is migrating between different embodiments (Pioneer and PeopleBot)

Although UH does have a customised PeopleBot robot which is capable of gestures (cf. Walter et al., 2008), robot gesture has not yet been specifically addressed in the experiments performed for LIREC to date. However, it will be feasible at some point in the future, when the new robot CHARLY development is sufficiently advanced, and experiments may be performed which include aspects addressing peoples perceptions of robot gesture.

A design of a light robotic hand capable of executing a collection of gestures has been proposed by WRUT, see subsection 2.2.2 of this report.

It is anticipated more complete details will be presented in the forthcoming WP6 deliverable D6.1.

The studies to date have only included a necessary minimum of robot speech (synthesis) to set the scene and progress the experimental scenarios. The speech synthesiser used in these studies is a run-time cut-down version of the Festival open source speech synthesis program, Festival Lite (flite). This is very light on computing resources, but only includes a single, rather course, obviously artificial voice. It is anticipated that the improved speech recognition and synthesis systems developed and implemented under WP3, will also be able to be implemented on the UH robots for use in future experiments.

Regarding the capability of analysing the user's non-verbal behaviour, scenario-dependent experimental tests will be carried out by QMUL to endow robots with an active vision system which can perform reasonably well on mobile platforms under noisy conditions and which is sensitive to different face orientations. Issues related to the use of a pan-tilt-zoom camera will be also considered (see subsection 2.1 of this report).

4.2 Pleo experiments

This study was conducted in order to investigate a situation where subjects meet Pleo (as a representative for an artificial companion) for the first time.

4.2.1 Research Question

- What characteristics of the robot are critical for the first impression?
- How do age, gender and personality of the user affect the way of interaction with the robot?
- Is it possible to generate deep sympathy with an artificial „life form” by short-time interaction?
- Does direct interaction with pleo cause emotions?

4.2.2 Study Design

In the “Pleo – First Contact” study, we used a combination of questionnaires, interviews and behavioural analysis. First of all we gave the subjects some basic information on the study (task: get known to Pleo, duration: 60 min., information on data security, etc.). The number of participants was 20 (10 male, 10 female).

Then subjects completed the first questionnaire, comprising questions on demographical issues, a short “Big Five” personality assessment (TIPI; Muck et al., 2007), questions on their relation towards technical equipment and gadgets, their feelings towards animals and some questions on robots.

After that, the subjects were allocated to one of two experimental groups: Group ‘A’ watched a video clip of 2:30 minutes duration in which a Pleo was destroyed by a “war-bot” in an arena surrounded by cheering people. After the video, the subjects were interviewed about their feeling while watching and whether they had sympathy with the robot dinosaur. After that, the subject “met” Pleo. Group ‘B’ first met Pleo and then watched the “war-bot”-video.

Before the subjects met Pleo, they were told that now they have some time to get known Pleo while the investigator would leave the room, and that they couldn’t do anything wrong and they could call at any time, if they had any questions. The Pleo was switched on in the other room, brought to the table in sleeping position and put in front of the subjects, facing them. After the investigator had left the room, Pleo “woke up”. After 5 minutes of free interaction, the investigator reentered the room and handed the subject 3 tasks on paper cards that they were meant to perform with Pleo during another 5 minutes of interaction (“Allure Pleo”, “Feed Pleo” and “Make Pleo sleep”). After these 5 minutes, Pleo was removed from the subject, brought to the other room and then switched off.

The next part of the study was the video reconstruction: The whole interaction between Pleo and the subject was recorded on video. The subjects were asked to comment on their actions and feelings while watching this video. Topics that were not mentioned by the subjects themselves were mentioned by the investigator in order to organize the reconstruction along a semi-structured interview guide (e.g., what was surprising, what was pleasant / awkward when interacting with Pleo).

Finally, a second questionnaire was handed out containing similar questions as the first one on robots and Pleo.

4.2.3 Results

Analysis of the video material of the interaction, e.g. regarding communication between Pleo and the subjects, is currently ongoing.

4.3 *iCat* experiments

4.3.1 Research-Intention

This study was conducted to evaluate if user's perceived social presence towards a social robot changes over time. We argue that by analyzing social presence over time, some indicators and features about what artificial companions should have to engage users in long-term interactions can be retrieved. We conducted a long-term experiment using "iCat, the Affective Chess Player", a system in which a social robot plays chess against a human opponent on an electronic chessboard. The robot's affective behaviour is influenced by the moves played by the human.

4.3.2 Research Questions

- Evaluate if the user's perceived social presence changes over time.
- Identify, if any, the aspects of social presence that are most affected over time.
- Analyze the user's general interaction with the system over the weeks.

4.3.3 Study Design

Participants

The experiment took place in a local chess club where every Saturday morning children between 5 and 16 years old take chess lessons from an instructor and play with each other. The class is composed of 7 children. Although the evaluations were always performed with every child who attended the sessions, for the analysis of the quantitative results we only considered the ones who did not miss any session and thus played every consecutive week with the iCat. So the total number of subjects is 4, three males and one female.

None of the participants had interacted with the iCat or with any social robot before. Most of them already had some contact with chess software or computerized chess games, where sometimes their opponent is represented as a graphical avatar. Some of the younger participants had limited reading comprehension.

Procedure

The system including the robot, the electronic chessboard and the laptop was placed in a table in the room of the chess club where participants were attending the lessons. The subjects were seated in front of the iCat and the chessboard like in a regular chess game, as Figure 4.3 shows.



Figure 4.3 User playing with the iCat

A set of chess exercises was previously proposed to the chess instructor. He analyzed the exercises and suggested some modifications so that the difficulty of the exercises was adequate for each participant. In each session participants played a different chess exercise against the iCat. The exercises for a particular session were chosen randomly, but none of the participants played the same exercise twice.

While iCat was playing with one subject, the others could be watching the game or playing against each other, continuing their lessons. The idea was to integrate the robot in the group as one of their own. In this way, users were directly interacting with iCat during their game and indirectly during the remaining time, by having the robot in the room as one of their colleagues.

The experiment was performed over five consecutive weeks. All the sessions were video recorded for further analysis. At the end of both the first and the last sessions children were asked to fill a questionnaire that measures social presence. Some of the younger participants needed help in filling the questionnaire.

Measures

We measured social presence in two ways: using a questionnaire and by video observation. Given the reduced number of subjects, the results of the questionnaire alone would not be enough to generate statistical significant results, so the recorded videos of the sessions were also analysed.

The social presence questionnaire was based on the Harms and Biocca (2004) questionnaire, which conceptualizes social presence in six dimensions:

- *Co-presence* refers to “the degree to which the observer believes s/he is not alone”.
- *Attentional allocation* addresses “the amount of attention the user allocates to and receives from an interactant”.
- *Perceived message understanding* is “the ability of the user to understand the message from the interactant”.
- *Perceived affective understanding* refers to “the user’s ability to understand the interactant’s emotional and attitudinal states”.
- *Perceived affective interdependence* refers to “the extent to which the user’s emotional and attitudinal state affects and is affected by the interactant’s emotional and attitudinal states”.
- *Perceived behavioural interdependence* is “the extent to which a user’s behaviour affects and is affected by the interactant’s behaviour”.

We translated the social presence questionnaire to the subjects’ native language, and selected two items for each dimension that would be adequate for children (see Table 4.1). Subjects were asked to express their agreement or disagreement regarding each item in a five-point Likert scale, where zero meant “totally disagree” and five stood for “totally agree”.

The videos from both the first and last sessions of the four users who played all the exercises were analyzed using ANVIL, a free video annotation tool for adding structured human annotations to digital video material (Kipp, 2008). ANVIL allows coders to annotate regions in the video on multiple layers called tracks. Each track can hold a number of attribute-value pairs. For this particular experiment, we defined the following tracks:

- User looking at iCat, in which we also defined the attributes “after user’s own move”, “after playing iCat’s move” and “during the game”. The attributes were chosen to distinguish the motivations of user’s attention towards the robot during the game, considering that: (1) after user’s own move, the iCat performs an emotional reaction; (2) after playing iCat’s move, the user receives feedback from the robot to confirm or disapprove that move, and

(3) during the game, which covers the remaining moments, iCat is blinking and looking to sides and its “face” reflects the mood.

- User looking sideways.
- User talking to iCat, with an attribute to save the utterance.
- Facial expressions of the user, containing two attributes: one to indicate the type of facial expression (e.g., sad, happy...) and the other to explain the reason of such expression.

4.3.4 Results

4.3.4.1 Social Presence Questionnaire

The means of the Likert scale responses for each questionnaire item in both the first and fifth weeks are presented in Table 4.1. In general, perceived social presence decreased after five weeks of interaction.

		1 st week	5 th week
Co-Presence			
Q1	I noticed iCat.	4,00	3,75
Q2	iCat noticed me.	3,75	3,75
Attentional Allocation			
Q3	I remained focused on iCat.	3,50	2,75
Q4	iCat remained focused on me.	3,75	3,25
Perceived Message Understanding			
Q5	My thoughts were clear to iCat.	3,25	2,75
Q6	iCat's thoughts were clear to me.	3,00	3,25
Perceived Affective Understanding			
Q7	I could tell how iCat felt during the game.	3,00	3,00
Q8	iCat could tell how I felt during the game.	2,25	2,50
Perceived Emotional Interdependence			
Q9	I was sometimes influenced by iCat's moods.	3,75	3,00
Q10	iCat was sometimes influenced by my moods.	3,50	2,75
Perceived Behavioural interdependence			
Q11	My behaviour was closely tied to iCat's behaviour.	3,50	2,25
Q12	iCat's behaviour was closely tied to my behaviour.	3,50	2,00

Table 4.1 Social presence questionnaire items and Likert scale means for each item in the first and fifth week

In the co-presence dimension, considering the Q1 means, users seem to notice iCat less on the last week. This could actually be asserted by video observation. The number of times that children looked at iCat on the last interactions is lower than in the first ones, as we will discuss on the next Section. This may happen due to the novelty effect mentioned earlier, as none of the children had interacted with a social robot before. In spite of that, when asking if iCat notice them (Q2), their opinion did not change significantly. The turn-taking nature of the chess game may be the main cause for this result. Since iCat reacts to children's moves and after that asks them to play its move, children probably interpreted that “reactive” behaviour as the robot noticing their presence.

Both the items regarding attentional allocation (Q3 and Q4) decreased after five weeks. From our observations within the chess club, when kids are playing with each other, they refer to previous games they played together, explain to each other some theory behind a certain move and sometimes they even make fun of each other. Some of these behaviours could be implemented in iCat to increase the user's attention to the robot, especially the ones related with memory.

In the perceived message understanding category, it is interesting to see that although user's perception of how clear their thoughts were to iCat (Q5) decreased, the opposite increased, i.e., after several weeks kids claimed to know better what iCat was thinking. This also happens for the perceived affective understanding dimension (Q7 and Q8). For instance, on the last weeks of interaction, when iCat reacted sadly to a good move from the user, some of them talked to the robot: "I know you don't like that".

The last two dimensions (perceived emotional interdependence and perceived behavioural interdependence, from Q9 to Q12) were the ones whose means decreased the most after the long-term experiment. After a few weeks, iCat seems to be perceived much more as an automaton, behaving independently of how the users feel or act, only reacting to their moves.

4.3.4.2 Video Observation

The sample of the data from the questionnaires is too small to apply more sophisticated statistical tests. Still, those results combined with the data collected through video observation already gave us some clues on the most relevant social presence dimensions for long-term interaction.

We exported the data resulting from ANVIL track annotations and attributes for quantitative analyses. Since all the exercises had different durations (from 5 to 15 minutes), we present the values as percentages of the total duration of the exercise, to be able to compare them between sessions.

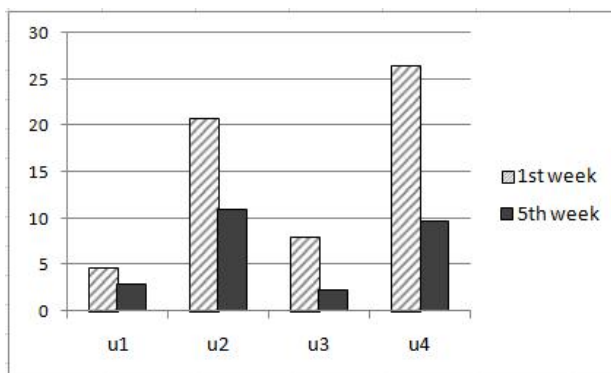


Figure 4.4 Total percentage of time that each user spent looking at iCat in the first and fifth weeks

From the four tracks defined for annotation, "looking at iCat" was the one with the largest number of annotations and also the one with more different results between the two sessions for each user. Figure 4.4 shows the percentages of the "looking at iCat" track for each user in the first and last week of interaction. As one can see in the chart, the total time that children spent looking at iCat on the last session is, on average, half the time that they looked in the first week. These results are aligned with the ones obtained in the social presence questionnaire, especially for the co-presence and the attentional allocation dimensions.

Another relevant issue may be age: the first user (u1) was fourteen years old, u2 was five, u3 was thirteen and u4 was nine. If we consider age when analysing the time users spent

looking at iCat, the data suggest that younger users feel much more engaged to this kind of robots. Even so, the decay of attention after five weeks was also verified in this case.

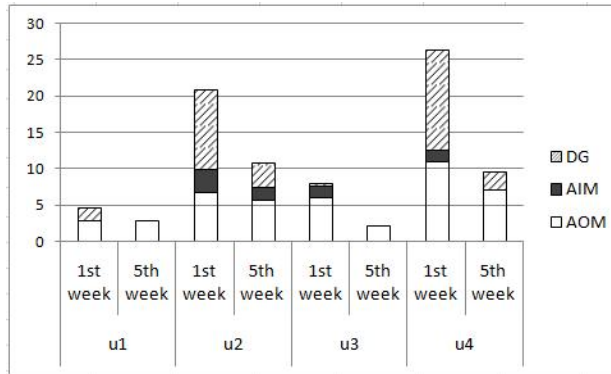


Figure 4.5 Percentages of the “looking at iCat” track for the first and fifth week of interaction broken down by its attributes: during the game (DG), after playing iCat’s move (AIM) and after user’s own move (AOM)

A more detailed analysis of the “looking at iCat” attribute values is shown in Figure 4.5 and Figure 4.6. Here we can see that both the attributes “during the game” (DG) and “after iCat’s move” (AIM) were the ones that decreased the most between the two sessions. The sum for all the users of the DG attribute in the first week is 27%, whereas in the last is 6%, and the sum of the AIM is 6% in the first session and 2% in the last. The accentuated decrease in these two attributes may be explained by the novelty effect of the first sessions.

In the “after own move” (AOM) stage of the game, the fall was not so pronounced: in the first week is 27% and in the last week is 18%. These results were quite expected, as after the user’s own move iCat performs an affective reaction that can help the users in the game. Again, these results strengthen the ones obtained in the perceived message and affective understanding dimensions of the social presence questionnaire, which also remained similar over the weeks.

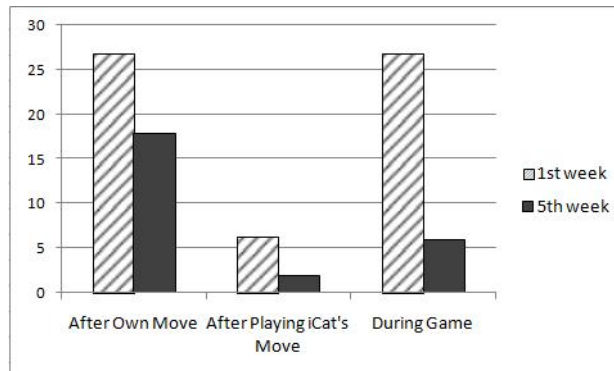


Figure 4.6 Total percentage of each attribute of the “looking at iCat” track for all users

The annotations of the remaining tracks were considerably more sporadic than the “looking at iCat” track, so instead of analysing the quantitative results obtained from ANVIL we will discuss the most relevant findings on each topic.

In the “looking sideways” track we did not find very significant differences between the annotations of the first and the last week of interaction. Most of the time users looked away from iCat or the chessboard was due to some external event at the chess club (e.g. someone arriving at the club). Still, in the last weeks some participants expressed signals of boredom after playing their move, while waiting for iCat’s affective reaction and consequent move. Some of them looked away, but it was only for short periods of time.

Regarding the “user talking to iCat” track, it changed significantly among users and not so much over the weeks. The older subjects barely talked to the robot, but the younger participants did. Within younger participants, this track did not change over the weeks, which may be because in the first weeks users were not so comfortable interacting with the robot. Over the weeks, some participants started talking to iCat even when it was not their turn to play. It remains to be validated if this behaviour would continue over subsequent interactions.

As for the “user’s facial expressions” track, we basically identified two types of facial expressions: the ones users displayed when they did not understand an iCat’s affective reaction and the ones performed in the end of the game. The expressions users displayed to show misunderstanding about iCat’s reactions decreased over the weeks, which again indicates that over time the perceived message and affective understanding dimensions of social presence tend to improve (or at least remain the same). There were no substantial variations on the user’s expressions in the end game though. Usually, when winning the game, users showed happy faces and when losing they made a sad expression or showed no expression at all.

4.3.5 Summary and future prospects

The outcomes of the evaluation indicate that the user’s perception of social presence towards the iCat decreased after five weeks. We are aware that the results were obtained in a specific domain (a social robot in a chess game) and as such more experiments should be performed to see if the results can be generalized to other domains.

Our main contribution was the identification of the dimensions of social presence that decreased the most after five weeks of interaction, which were co-presence, attentional allocation, perceived emotional interdependence and perceived behavioural interdependence. These dimensions are mainly related to agent’s believability and user’s attention to the system. We observed that the attention users dedicated to the robot decreased significantly over the weeks, which suggests that new mechanisms and behaviours must be developed in the agent to maintain the engagement.

In summary, we concluded that the robot’s current behaviour is not enough to maintain the perception of social presence after several interactions. Although it might appear believable and intelligent on the first impressions, *as time goes by*, users need more.

4.4 Mozart experiments

The global idea is to create a friend tutor that can improve the learner process of melodic composition of music. To do so, we will use the work already done with “Pequeno Mozart” (Little Mozart) but we will try to integrate an emotion model that will be influenced by some personality characteristics that we will call for now on “personality”. To implement this idea we will base our research in FLAME model that is based in OCC model. In order to achieve our goals we will extend the FLAME model by introducing some personality characteristics.

In a general way, the event process will occur in the following way:

- Event Evaluation – evaluation based on the goals and the desirability of the event
- Event Appraisal – Initiation of the emotional state, in rough, where we only take in account the desirability and the expectation (originated by the Learning Component).
- Emotional Filtering – Taking in account the personality and the mood we will filter the emotions, increasing and decreasing some of them according to the personality.
- Behaviour Selection – In this, we will select the behaviours that the agent will externalize, taking into account the personality, the state of emotions, the mood and the learning obtained in previous experiences. From this component will result three types of feedbacks: verbal, facial and body motion.

Complementary Ideas:

- The mood will be considered a feeling because is constant along the time and a result of emotional past experiences.
- The Learning Component will evaluate the previous events statistically and will define the expectation of each event
- The “personality” will be limited to aspects that interfere with emotions
- The memory will contain, in this phase, only short-term registry.

4.4.1 Target-Group

We expect to act in two schools of Coimbra, three classes of 7 year old students with proximally 10/15 students each. We predict an hour, twice a week of interaction with the agent.

4.4.2 Main Phases

- First contact with schools
- Formal project presentation
- Involving the class with the agent
- Periodic observations – short, medium and long period
- Final evaluation

4.4.3 Objectives

- Analyze the interaction dynamic of the class with the synthetic agent
- Identify the factors involved in relation and learning process

4.4.4 Hypothesis/Research Questions

- Can students establish long-term relationships with these agents? Why does it happen?
- Which agent established a longer and more useful relation with? Why? What factors were involved?
- In what way the level of control of the student towards de agent influences their relation?
- Does the agent facilitates or potencies the learning process? How so?

4.4.5 Observation/ Evaluation Criteria

- Usability of the interface for interaction with the companion
- User experience - video recording of facial expressions
- Levels of comfort of the interaction
- Levels of learning and how do they relate with emotions and “personality” mood of the agent

4.5 AIBO experiments

Comparison dog-human and robot-human interactions

The data of Study 1 is originated from Kerepesi et al., 2006. Study 2 and 3 presents new data on (partly) the same samples.

4.5.1 Introduction

The companion toy robots are designed specially to interact with people and to provide some kind of "entertainment" for humans. They have the characteristics to induce an emotional relationship ("attachment") (Donath, 2004; Kaplan, 2001). One of the most popular companion toy robot was Sony's AIBO (Pransky, 2001) which is to some extent reminiscent to a dog-puppy.

Two different approaches have been used to investigate humans' interaction with companion toy robots. Some researchers use questionnaires to find out whether humans perceive AIBO similar to a dog and what kind of emotions they attribute to the robot. Such investigations carried out at online AIBO discussion forums focused on describing owners' perceived relationship with their AIBOs (Kahn et al., 2003). About 42% of the participants spoke of AIBO having feelings, while 26% of them spoke of AIBO as a companion. Kahn et al. (2003) suggested that the relationship between people and their AIBO could be similar to the human-dog relationship.

It is also interesting how people speak about the robot. Do they refer to AIBO as a non-living object, or as a living creature? When comparing children's attitudes towards AIBO and other robots Bartlett et al. (2004) found that children referred to AIBO as if it were a living dog, labelled it as "robotic dog" and used rather 'he' or 'she' than 'it' when talked about AIBO. Interviewing children Melson et al. (2004) found that although they distinguished AIBO from a living dog, they attributed psychological, companionship and moral stance to the robot. Interviewing older adults Beck et al. (2004) found that elderly people regarded AIBO much like as a family member and they attributed animal features to the robot.

The second line of studies use ethological methods for describing robot-human interactions by the means of behaviour analysis. In a comparative study Kahn et al. (2004) found that children distinguished between AIBO and a stuffed dog toy in their interactions. Although they engaged in an imaginary play with both of them, they showed more exploratory behaviour and attempts for reciprocity when playing with AIBO. Turner et al. (2004) found that children touched a live dog over a longer period than the robot but ball game was more frequent with AIBO than with the dog puppy.

In our study we investigated children's and adults' behaviour during a play session with AIBO and compared it to playing with a live dog puppy. The aim of this study was (1) to analyse spontaneous play between the human and the dog/robot, including the comparison of the temporal structure of the interactions in both children and adults and (2) to analyse the verbal communication of adult subjects towards the partners. In addition, we measured (3) the attitude towards the partners and animals in general of those adult participants, who were engaged in playing with both AIBO and dog. Some ethical issues were also covered.

4.5.2 Method

Study 1. Analysis of spontaneous interaction with AIBO and dog

Subjects

28 and 28 children participated in the test and were divided into four experimental groups.

1. Adults with AIBO: 7 men and 7 women (Mean age: 21.1 years, SD= 2.0 years)
2. Children with AIBO: 7 boys and 7 girls (Mean age: 8.2 years, SD= 0.7 years)

3. Adults with dog: 7 men and 7 women (Mean age: 21.4 years, SD= 0.8 years)

4. Children with dog: 7 boys and 7 girls (Mean age: 8.8 years, SD= 0.8 years)

Procedure

The test took place in a 3m x 3m separated area of a room. Children were recruited from elementary schools, adults were university students. The robot was Sony's AIBO ERS-210, (dimension: 154 mm x 266 mm x 274 mm; mass: 1.4 kg; colour: silver) that is able to recognise and approach pink objects. To generate a constant behaviour, the robot was used only in its after-booting period for the testing.

The dog puppy was a 5-month-old female Cairn terrier, similar size to the robot. It was friendly and playful, and its behaviour was not controlled in a rigid manner during the playing session. The toy for AIBO was its pink ball, and a ball and a tug for the dog-puppy.

The participants played for 5 minutes either with AIBO or the dog puppy in a spontaneous situation (Figure 4.7). None of the participants met the test partners before. During the interactions participants were not controlled in any way. Those who played with the AIBO knew that it liked being stroked, that there was a camera in its head enabling it to see and that it liked to play with the pink ball.

The video recorded play sessions were coded by ThemeCoder. Transcribed records were analysed using Theme 5.0 (www.patternvision.com). Play behaviour (human moves the toy in front of the partner) and interest in the partner (stroking behaviour and orientation to the dog/AIBO) were analysed. Additionally we investigated the temporal structure of the actions. Temporal patterns (T-patterns) are composed of simpler directly distinguishable event-types, which are coded in terms of their beginning and end points (such as "dog begins walking" or "dog ends orienting to the toy"). In short, within a given observation period, if two actions, A and B, occur repeatedly in that order or concurrently, are said to form a minimal T-pattern (AB) if found more often than expected by chance (Magnusson, 2000).

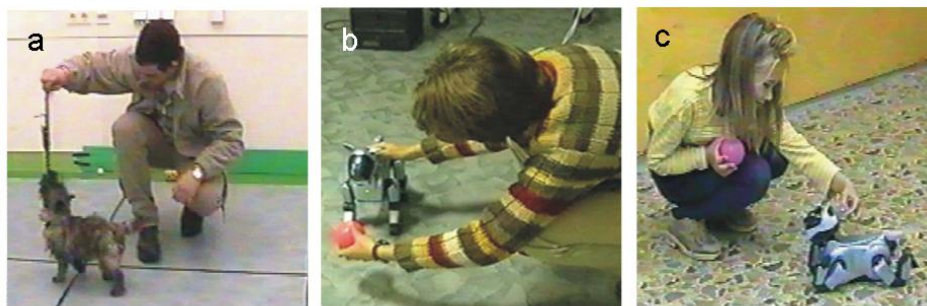


Figure 4.7 Adult interacting with the dog (a) with AIBO (b) and child interacting with AIBO (c)

Study 2. Analysis of verbal communication of adults during interaction with AIBO/dog

Subjects

1. Adults with AIBO: 7 men and 7 women (same as in Study 1; Mean age: 21.1 years, SD= 2.0 years)

2. Adults with dog: 7 men and 7 women (same as in Study 1; Mean age: 21.4 years, SD= 0.8 years)

Procedure

See Study 1. Latencies, frequencies, duration and length of utterances and speech-breaks were measured.

Study 3. Attitudes towards the partners and ethical issues (questionnaire study)**Subjects**

Twenty one university students (11 men, 10 women; Mean age= 21.2, SD=2.0 years) interacted with both partners (dog and AIBO). Out of 21, 14 play sessions with AIBO were included in Study 1. 52% of the students had a dog at home, 71% knew at least 1 programming language, and 57% heard about AIBO before.

Procedure

Subjects interacted with both partners as described in Study 1. Half of the subjects played with the AIBO first, while the other half has begun with the dog puppy (8-weeks-old Hungarian vizsla puppy). Afterwards they were asked to fill in a questionnaire about their attitudes towards robots, and were asked to compare the two play partners.

4.5.3 ResultsStudy 1. Behavioural analysis of the interaction with AIBO and dog

BEHAVIOURAL VARIABLE	SIGNIFICANT DIFFERENCE AMONG GROUPS
Latency of the first touch of the dog/AIBO	No
Duration and frequency of stroking (s)	No
Frequency of stroking (s)	No
Duration and frequency of looking at the dog/AIBO (s)	No
Frequency of looking at the dog/AIBO (s)	No
Duration of moving the toy in front of dog/AIBO	Both children and adults spent more time moving the toy in front of the AIBO.
Frequency of moving the toy	No
Ratio of T-patterns initialized by humans	Adults initialized T-patterns more frequently when playing with dog than participants of the other groups (Figure 4.8)
Ratio of T-patterns terminated by humans	Both children and adults terminated the T-patterns more frequently when they played with AIBO than when they played with the dog puppy (Figure 4.8).

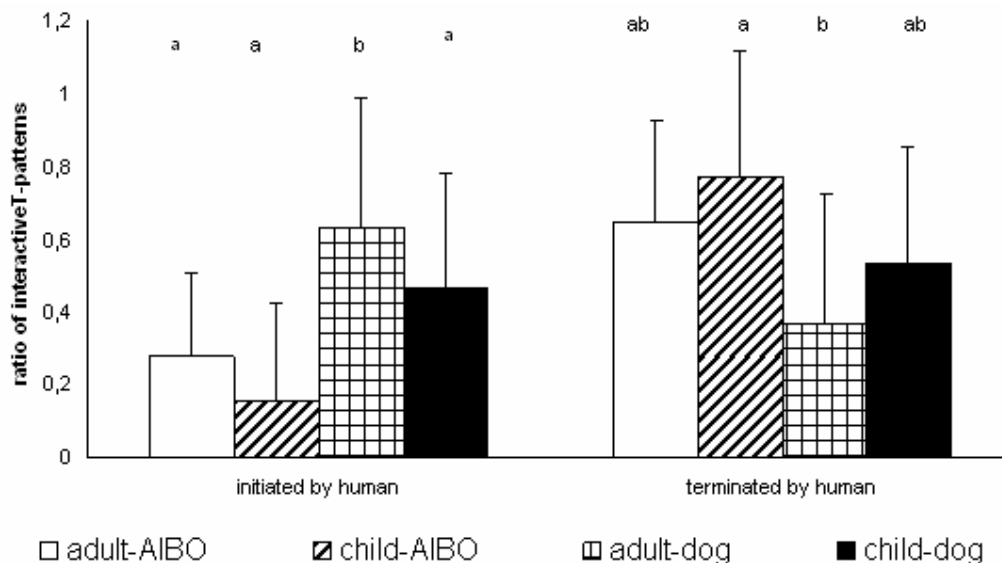


Figure 4.8 Mean ratio of interactive T-patterns initiated and terminated by humans. Different letters mean significant differences among the groups

Study 2. Analysis of verbal communication of adults during interaction with AIBO/dog

We have not found significant differences between neither the latency, nor the duration and frequency of utterances towards the dog and the AIBO. However, people talked 3.5 times more to the dog than to the AIBO (Mean $s \pm SE = 10.4 \pm 3.3$ vs 35.4 ± 9.2 respectively); started to talk to the dog 4.5 times sooner than to the AIBO (Mean $s \pm SE = 42.8 \pm 11.8$ vs 10.04 ± 2.5). The frequency of utterances were almost the same (Mean $s \pm SE = 8.14 \pm 2.18$ vs 10.04 ± 1.4 ; Figure 4.9)

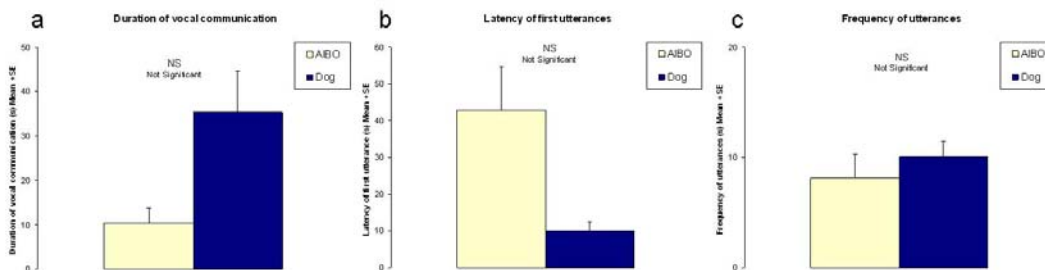


Figure 4.9 Duration, latency and frequency of verbal communication of adults during dog/AIBO interactions

Both length of speech-breaks and utterances differed significantly between the two groups. People’s utterances were longer ($t = 2.294$, $df = 23$, $p = 0.03$; (Mean $s \pm SE = 1.12 \pm 0.39$ vs 4.79 ± 5.26), and the length of speech-breaks were shorter ($t = 2.434$, $df = 24$, $p = 0.02$; (Mean $s \pm SE = 20.58 \pm 17.9$ vs 8.37 ± 5.4) when they spoke to the dog (Figure 4.10).

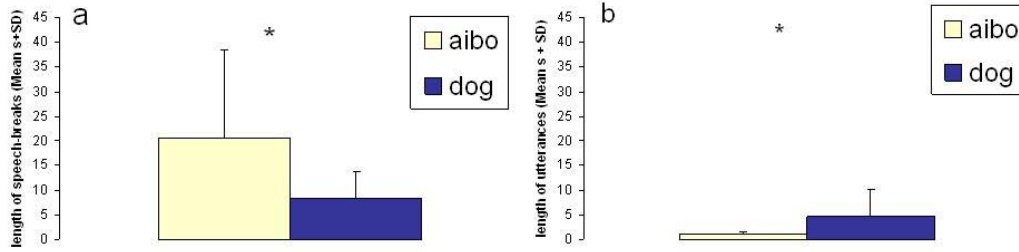


Figure 4.10 Length of speech-breaks (a) and utterances (b)

Study 3. Attitudes towards the partners and ethical issues (questionnaire study)

Attitudes toward AIBO and animals

What do you think is the price of the robot?	68% appraised the price less, 16% more, 10,5% knew the exact price, 5% did not know
Which would be the preferred colour if you could own an AIBO?	29% silver, 0% golden, 52% black, 19% does not care
How would you name your own AIBO?	11% would not give a name, 28% name of a boy, did not specify, 28% dog name, 33% specific robot name (like 'Robo')
Would you like an AIBO without a "switch off" button?	5% yes, 95% no
How much time would you devote for interacting with an AIBO per day?	5% nothing, 24% 10 min, 14% 30 min, 10% 1-2 h, 48% would not use it every day
How much time would you devote/do you devote for a/your dog per day?	0% nothing, 14% 10 min, 29% 30 min, 38% 1-2 h, 14% 2-3 h, 5% more than 4 h per day
Would you buy an AIBO for your child?	19% yes, 81% no
What kind of animal would you buy for your child? (not exclusive)	78% dog, 39% cat, 33% small rodent
Why do you keep animals? (not exclusive)	41% like having animals around, 35%: friend and company, 35%: simply like them, 12%: useful
Which partner did you prefer to interact with?	92% puppy (because it is a living being, warm, hairy, pay attention at me, curious, moves subtly?, has feelings, trainable, communicating, backfeed?, not computable), 8% AIBO (because it moved interestingly)

Communication with dogs

How would you/do you talk to your dog?	10% command, 33% complete sentences, 57% continuous speaking
What is the level of understanding human speech in dogs?	0% nothing, 43% only feelings, 38% understands some words, 14% understands some sentences, 5% understands human speech well

Ethical issues

Do you have bad feelings in connection with the coming of robotic pets?	38% yes, 62% no
Can it be advantageous having a robotic pet?	48% yes, 52% no
Is it possible that children will not be able to distinguish robots from living beings?	19% yes, 81% no
Is it possible that robots will be used for cruel purposes in the future?	81% yes, 19% no
Is it possible that robots will be smarter than their creators?	19% yes, 81% no

4.5.4 Discussion

Previous questionnaire studies on human-robot interaction showed, that people describe their relationship with AIBO similar to a relationship with dog puppy (Kahn et al., 2003), attribute animal characteristics to the robot and view it as a family member (Beck et al., 2004). However, the analysis of their behaviour tended to show that in parallel they behave differently toward the AIBO and a living dog puppy (Turner et al., 2004).

Considering the behavioural pattern of the humans, our results show that neither the latency of the first tactile contact between humans and the dog/AIBO nor the duration of stroking the dog/AIBO nor the verbal communication differed significantly among the groups. This suggests that under the present conditions the robot was as an affective playing partner for both children and adults as the dog puppy.

To investigate whether humans interact with AIBO as a non-living toy rather than a living dog, we have analyzed the temporal patterns of these interactions. We have found that similarly to human interactions (Borrie et al., 2002; Magnusson, 2000; Grammer et al., 1998) and human-animal interactions (Kerepesi et al., 2005), human-robot interactions also consist of complex temporal patterns. In addition the numbers of these temporal patterns are comparable to those T-patterns detected in dog-human interactions in similar contexts.

One important finding of the present study was that the type of the play partner affected the initialization and termination of T-patterns. Adults initialized T-patterns more frequently when playing with dog while T-patterns terminated by a human behaviour unit were more frequent when humans were playing with AIBO than when playing with the dog puppy. In principle this finding has two non-exclusive interpretations. In the case of humans the complexity of T-patterns can be affected by whether the participants liked their partner with whom they were interacting or not (Grammer et al., 1998; Sakaguchi et al., 2005). This line of arguments would suggest that the distinction is based on the differential attitude of humans toward the AIBO and the dog. Although, we cannot exclude this possibility, it seems more likely that the difference has its origin in the play partner. The observation that the AIBO interrupted the interaction more frequently than the dog suggests that the robot's actions were less likely to

become part of the already established interactive temporal pattern. This observation can be explained by the robot's limited ability to recognize objects and humans in its environment.

Although the results of the traditional ethological analysis both in our and other studies (e.g. Kahn et al., 2004; Bartlett et al., 2004) suggest that people interacting with AIBO in same ways as if it were a living dog puppy, and that playing with AIBO can provide a more complex interaction than a simple toy or remote controlled robot, the analysis of temporal patterns revealed some differences. The differences in initialisation and termination of the interactions could have a significant effect on the human's attitude toward their partner, that is, in the long term humans could get "bored" or "frustrated" when interacting with a partner that has a limited capacity to being engaged in temporally structured interactions. This hypothesis is strengthened by the questionnaire study, where 92% of the participants indicated that they preferred playing with the puppy in contrast to the AIBO.

In summary, contrary to the findings of previous studies, it seems that at a more complex level of behavioural organisation human-AIBO interaction is still different from the interactions displayed while playing with a real puppy, and in the future more attention should be paid to the temporal aspects of behavioural pattern when comparing human-animal versus human-robot interaction.

5 Concluding remarks and future work

Although the experiments so far are limited, and many new experiments are planned for the near future, we can draw some preliminary implications and recommendations from them.

- Humans talk to their companions although they somehow know that there is limited capability in the understanding of the companion;
- The interaction with companions is usually based on full sentences rather than just words and orders;
- There is a lot of gestual and visual communication with pet companions;
- Friendship is initiated by both parts and is usually related with common interests;
- Companions (even pet companions) do initiate the communication;
- Humans are willing to interact with robotic pets, use similar types of communication, but prefer real ones;
- The interaction with a robotic playmate becomes less “social” with long term interaction;
- There is a clear lack of responsiveness of the existing systems and robotic companions in terms of behaviour and expression;
- Personality factors affect the communication and the degree of attachment between the companions (similarity seems to be crucial).

Taking into account these findings, in this WP we will investigate further the techniques for the perception of users by creating a framework for the perception of the user actions using several modalities when interacting with companions. In this we will work on:

- Markerless vision systems aware of user body pose and facial expression;
- Limited language and speech recognition and synthesis using off the shelf developed systems for human/companion communication;
- Facial and body expression of the robots and embodied graphic characters;
- Integration of non-verbal communication expression in companions relevant for developing long term relations and tests in the application scenarios.

Aware that such a development is too large, we believe that the findings reported in this deliverable will restrict the problem at hands by allowing us to consider more specific scenarios of companionship.

6 References

- Allen, J. (1994). *Natural Language Understanding*, Benjamin Cummings, Second Edition.
- Ambady, N. & Rosenthal, R. (1992). Thin Slices of Expressive Behaviour as Predictors of Interpersonal Consequences: A Meta-Analysis. *Psychological Bulletin*, 111(2), 256-274.
- Anderson, K. & McOwan, P. (2006). A Real-Time Automated System for Recognition of Human Facial Expressions. *IEEE Trans. Systems, Man, and Cybernetics – Part B*, 36 (1), 96-105.
- Andersson, J., Badino, L., Watts, O., Aylett, M. (2008). The CSTR/Cereproc Blizzard Entry 2008: The Inconvenient Data, University of Edinburgh, UK / CereProc Ltd, UK.
- Aryananda, L. & Weber, J. (2004). MERTZ: A Quest for a Robust and Scalable Active Vision Humanoid Head Robot. *IEEE-RAS/RSJ International Conference on Humanoid Robots (Humanoids 2004)*, November 2004.
- Balestri M., A. Pacchiotti., S. Quazza, P. Salza and S. Sandri (1999). Choose the Best to Modify the Least: a New Generation Concatenative Synthesis System. *Proceedings of EUROSPEECH 1999*, 2291-2294, Budapest.
- Balomenos, T. Raouzaiou, A., Ioannou, S., Drosopoulos, A., Karpouzis, K. & Kollias, S. (2005). Emotion analysis in man-machine interaction systems. S. Bengio and H. Bourlard, eds., *Machine Learning for Multimodal Interaction*, 3361 of LNCS, 318–328, Berlin: Springer Verlag.
- Bartlett, B., Estivill-Castro, V., Seymon, S. (2004). Dogs or robots: why do children see them as robotic pets rather than canine machines? *5th Australasian User Interface Conference. Dunedin*. Conferences in Research and Practice in Information Technology, 28, 7-14
- Beck, A.M., Edwards, N.E., Kahn, P., Friedman, B. (2004). Robotic pets as perceived companions for older adults. *IAHAI0 People and animals: A timeless relationship*, Glasgow, UK, 72.
- Bhagat, R., Leuski, A. & Hovy, E. (2005). Shallow Semantic Parsing despite Little Training Data. *ACL/SIGPARSE 9th International Workshop on Parsing Technologies*, Vancouver, B.C., Canada.
- Bianchi-Berthouze, N. & Kleinsmith, A. (2003) A categorical approach to affective gesture recognition. *Connection Science*, 15(4), 259–269.
- Bohus D., Raux A., Harris T.K., Eskenazi M., Rudnicky A.I. (2007). Olympus: an open-source framework for conversational spoken language interface research, *Proc. of HLT-NAACL 2007 Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technology*, Rochester, NY.
- Boone, R. & Cunningham, J. (1998). Children's Decoding of Emotion in Expressive Body Movement: The Development of Cue Attunement. *Developmental psychology*, 34(5), 1007–1016.
- Boone, R. & Cunningham, J. (2001). Children's Expression of Emotional Meaning in Music Through Expressive Body Movement. *Journal of Nonverbal Behaviour*, 25(1), 21–41.
- Borrie, A., Jonsson, G.K., Magnusson, M.S. (2001). Application of T-pattern detection and analysis in sport research. *Metodologia de las Ciencias del Comportamiento*, 3, 215-226.
- Breazeal, C., Eadsinger, A., Fitzpatrick, P. & Scassellati, B. (2001). Active vision systems for sociable robots, in K. Dautenhahn (ed.), *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, 31(5), 443-453.
- Bruderlin, A. & Williams, L. (1995). Motion signal processing. In *Proceedings of SIGGRAPH '95, Annual Conference Series*, ACM SIGGRAPH, Addison Wesley, 97–104.

- Callaway C., Lester J. (2001). Narrative Prose Generation, *Proc. of the 17th Intl. Joint Conference on Artificial Intelligence*, Seattle, WA, 1241-1248.
- Camurri, A., Coletta, P., Varni, G. & Ghisio, S. (2007). Developing multimodal interactive systems with EyesWeb XML. *Proceedings of the 2007 Conference on New Interfaces for Musical Expression*, 305–308.
- Camurri, A., De Poli, G., Leman, M. & Volpe, G. (2005). Toward communicating expressiveness and affect in multimodal interactive systems for performing art and cultural applications. *IEEE Multimedia Magazine*, 12(1), 43–53.
- Camurri, A., Lagerlöf, I., & Volpe, G. (2003). Recognizing emotion from dance movement: Comparison of spectator recognition and automated techniques. *International Journal of Human-Computer Studies*, 59(1- 2), 213–225.
- Camurri, A., Mazzarino, B. & Volpe, G. (2004). Analysis of expressive gesture: The Eyesweb Expressive Gesture Processing Library. A. Camurri and G. Volpe, eds., *Gesture-based communication in human computer interaction*, 2915 of LNAI, 460–467. Berlin: Springer Verlag.
- Caridakis, G., Raouzaoui, A., Karpouzis, K. & Kollias, S. (2006). Synthesizing Gesture Expressivity Based on Real Sequences. *Proceedings of Workshop on Multimodal Corpora: from multimodal behaviour theories to usable models, LREC Conference 2006*, Genoa, Italy, May 2006.
- Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S. & Stone, M. (1994). Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In A. Glassner, editor, *Proceedings of SIGGRAPH '94, Computer Graphics Proceedings, Annual Conference Series*. ACM SIGGRAPH, ACM Press, 413–420.
- Cassell, J., Vilhjálmsdóttir, H. & Bickmore, T. (2001). BEAT: The Behaviour Expression Animation Toolkit. *In Proceedings of SIGGRAPH '01*, 477–486.
- Castellano, G. (2008). Movement expressivity analysis in affective computers: from recognition to expression of emotion, Ph.D. thesis, InfoMus Lab, DIST (Department of Communication, Computer and System Sciences), University of Genova, February 2008.
- Castellano, G., Kessous, L., Caridakis, G. (2008). Emotion recognition through multiple modalities: face, body gesture, speech. In C. Peter, R. Beale, editors, *Affect and Emotion in Human-Computer Interaction*. LNCS, 4868, Springer, Heidelberg.
- Castellano, G., Mancini, M. Analysis of emotional gestures for the generation of expressive copying behaviour in an embodied agent. In: M. Sales Dias, S. Gibet, M. Wanderley (Eds.), *Advances in gesture-based human-computer interaction and simulation*, LNCS, 5085, Springer Verlag. In press.
- Castellano, G., Villalba, S. & Camurri, A. (2007). Recognising human emotions from body Movement and gesture dynamics. In A. Paiva, R. Prada, and R. W. Picard, eds., *Affective Computing and Intelligent Interaction, Second International Conference, ACII 2007*, Lisbon, Portugal, September 12- 14, 2007, Proceedings, volume 4738 of LNCS, 71-82. Berlin: Springer- Verlag.
- Chen, L. et al. (2005). A robust algorithm for eye detection on gray intensity face without spectacles. *Journal of Computer Science and Technology*.
- Chi, D., Costa, M., Zhao, L. & Badler, N. (2000). The EMOTE model for Effort and Shape. In *Proceedings of SIGGRAPH '00*, Computer Graphics Proceedings, Annual Conference Series. ACM SIGGRAPH, ACM Press, 173–182.
- Cowie, R. (2007). *Emotional life: Terminological and conceptual clarifications*. Deliverable D3i HUMAINE EU-IST Project.

- Givens, D. B. Facial expression, <http://members.aol.com/nonverbal3/facialx.htm>
- De Meijer, M. (1989). The contribution of general features of body movement to the attribution of emotions. *Journal of Nonverbal Behaviour*, 13(4), 247-268.
- De Silva, P., Kleinsmith, A. & Bianchi-Berthouze, N. (2005). Towards Unsupervised Detection of Affective Body Posture Nuances. In J. Tao, T. Tan, and R. W. Picard, eds., *Affective Computing and Intelligent Interaction, First International Conference, ACII 2005*, Beijing, China, October 22-24, 2005, Proceedings, 3784 of LNCS, 32–39. Berlin: Springer-Verlag.
- Devon, M. (2006). *The origin of emotions*. Charleston, South Carolina.
- Donath, J. (2004). Artificial pets: Simple behaviors elicit complex attachments. M. Bekoff ed., *The Encyclopedia of Animal Behaviour*, Greenwood Press.
- Ekman, P & Friesen, W. (1969). The repertoire of nonverbal behaviour. *Semiotica*, 1, 49-98.
- Ekman, P. & Friesen, W. (1974). Detecting deception from the body or face. *Journal of Personality and Social Psychology*, 29, 288–298.
- Ekman, P. & Friesen, W. (1975). *Unmasking the face*. Englewood Cliffs, NJ, Prentice Hall, Inc.
- Ekman, P., Friesen, W. & Hager, J. (2002). *Facial Action Coding System: The Manual*. Salt Lake City.
- Filmakademie Baden-Wurtemberg. Research and Development at the Institute of Animation, <http://aistud.filmakademie.de/actor/9.0.html>
- Foster M.E., Bard E.G., Guhe M., Hill R.L., Oberlander J., Knoll A. (2008). The roles of haptic-ostensive referring expressions in cooperative, task-based human-robot dialogue, *HRI '08: Proc. of the 3rd ACM/IEEE Intl. Conference on Human Robot Interaction*, 295-302.
- Gácsi, M., Kara E, Belényi, B, Topál, J, Miklósi, Á. (2008). Critical analyses of dogs' comprehension of human pointing: Developmental and methodological concerns. *Animal Cognition* (in press)
- Gácsi, M., Miklósi, A., Varga, O., Topál, J., Csányi, V. (2004). Are readers of our face readers of our minds? Dogs (*Canis familiaris*) show situation-dependent recognition of human's attention. *Animal Cognition*, 7, 144-153.
- Gilbert, M. & Feng, J. (2008). Speech and Language Processing over the Web, *IEEE Signal Processing Magazine*, 18-28.
- Grammer, K., Kruck, K.B., Magnusson, M.S. (1998). The courtship dance: patterns of nonverbal synchronization in opposite-sex encounters. *Journal of Nonverbal Behavior*, 22, 3-29
- Gunes, H., & Piccardi, M. (2007). Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications*, 30(4), 1334–1345.
- Hakulinen J., Turunen M., Smith C., Charlton D., Zhang L., Cavazza M. (2008). A Model for Flexible Interoperability between Dialogue Management and Domain Reasoning for Conversational Spoken Dialogue Systems, *Proc. of the 4th Intl. Workshop on Human-Computer Conversation*.
- Hartmann, B., Mancini, M. & Pelachaud, C. (2005). Implementing expressive gesture synthesis for embodied conversational agents. In S. Gibet, N. Courty, and J.F. Kamp, eds., *Gesture in Human-Computer Interaction and Simulation: 6th International Gesture Workshop, GW2005*, Berder Island, France, May 18-20, 2005, Revised Selected Papers, 3881 of LNCS, 188–199. Berlin: Springer-Verlag.

Head Robot Team Takanishi Laboratory. Emotion Expression Humanoid Robot WE-4RII, <http://www.takanishi.mech.waseda.ac.jp/research/>

Hirsch-Pasek, K, Treiman R (1981). Doggerel: motherese in a new context. *Journal of Child Language*, 9, 229-237.

Huang, Y., Chiang, C. (2006). A rule-based real-time face detector. Department of Computer Science and Information Engineering, National Dong-Hwa University, Shoufeng, Hualien.

Intel. OpenCV Library, <http://www.intel.com/technology/computing/opencv/index.htm>

J. Fredslund. Feelix, http://www.daimi.au.dk/hili/feelix/feelix_home.htm

Jurafsky, D. & Martin. J. (2008). *Speech and Language Processing (2nd Edition)*. Prentice-Hall, Inc.

Kahn, P. H., Jr., Friedman, B., & Hagman, J. (2003). Hardware Companions? – What Online AIBO Discussion Forums Reveal about the Human-Robotic Relationship. *Conference Proceedings of CHI 2003*, New York, NY: ACM, 273-280.

Kahn, P.H., Friedmann, B., Perez-Granados, D.R., Freier, N.G. (2004). Robotic pets in the lives of preschool children. *CHI 2004*. Vienna, Austria, 1449-1452.

el Kaliouby, R. & Robinson, P. (2005). Generalization of a Vision-Based Computational Model of Mind-Reading. In J. Tao, T. Tan, and R. W. Picard, eds., *Affective Computing and Intelligent Interaction, First International Conference, ACII 2005*, Beijing, China, October 22-24, Proceedings, 3784 of LNCS, 582–589. Berlin:Springer-Verlag.

Kaplan, F. (2001). Artificial Attachment: Will a robot ever pass Ainsworth's Strange Situation Test? Hashimoto, S., ed., *Proceedings of Second IEEE-RAS International Conference on Humanoid Robots, Humanoids*. Institute of Electrical and Electronics Engineers, Inc., Waseda University, Tokyo, Japan, 99–106.

Kapoor, A., Qi, Y. & Picard, R. (2003). *Fully automatic upper facial action recognition*. MIT Media Laboratory, Cambridge.

Kapur, A., Virji-Babul, N., Tzanetakis, G. & Driessen, P. (2005). Gesture-Based Affective Computing on Motion Capture Data. In J. Tao, T. Tan, and R. W. Picard, eds., *Affective Computing and Intelligent Interaction, First International Conference, ACII 2005*, Beijing, China, October 22-24, Proceedings, volume 3784 of LNCS, pages 1–7. Berlin: Springer-Verlag.

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T. & Banno, H. (2008). TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0 and aperiodicity estimation, *Proc. ICASSP'2008*.

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press.

Kerepesi, A., Jonsson, G.K, Miklósi Á., Topál, J., Csányi, V., Magnusson, M.S. (2005). Detection of temporal patterns in dog-human interaction. *Behavioural Processes*, 70(1), 69-79.

Kipp, M., Neff, M., Kipp, K. & Albrecht, I. (2007). Towards Natural Gesture Synthesis: Evaluating Gesture Units in a Data-Driven Approach to Gesture Synthesis. In C. Pelachaud, J.-C. Martin, E. Andr, G. Collet, K. Karpouzis, and D. Pel, eds., *Intelligent Virtual Agents, 7th International Conference, IVA 2007*, Paris, France, September 2007, Proceedings, 4722 of LNAI, 15–28. Berlin: Springer-Verlag.

Kleinsmith, A. & Bianchi-Berthouze, N. (2007). Recognizing Affective Dimensions from Body Posture. In A. Paiva, R. Prada, and R. W. Picard, eds., *Affective Computing and Intelligent*

- Interaction, Second International Conference, AClI 2007*, Lisbon, Portugal, September 12-14, Proceedings, 4738 of LNCS, 48–58. Berlin:Springer-Verlag, 2007.
- Krstulovic, S., Hunecke, A. & Schroeder, M. (2007). An HMM-Based Speech Synthesis System applied to German and its Adaptation to a Limited Set of Expressive Football Announcements, *Proc. of Interspeech*.
- Laban, R. & Lawrence, F. (1947). *Effort*. London: Macdonald and Evans Ltd.
- Lakatos, G., Soproni, K., Dóka, A., Miklósi, Á. (2008). A comparative approach to dogs' (*Canis familiaris*) and human infants' comprehension of various forms of pointing gestures. *Animal Cognition* (in press).
- Leuski, A., Patel, R., Traum, D. & Kennedy, B. (2006). Building effective question answering characters. *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, Sydney, Australia.
- Littlewort, G., Bartlett, M., Fasel, I., Susskind, J. & Movellan, J. (2006). Dynamics of facial expression extracted automatically from video. *J. Image & Vision Computing*, 24(6), 615-625.
- Maeda, Y. & Tanabe, N. (2006). Basic Study on Interactive Emotional Communication by Pet-type Robot. *Transactions of the Society of Instrument and Control Engineers*, 42(4), 359-366.
- Magnusson, M.S. (2000). Discovering hidden time Patterns in Behaviour: T-patterns and their detection. *Behaviour Research Methods, Instruments & Computers*, 32, 93-110.
- Marcel, S. (2002). *Gestures for Multi-modal Interfaces: A review*. IDIAP Research Report. Dalle Molle Inst. for Perceptual Artificial Intelligence, Switzerland.
- Martin, J.-C., Abrilian, S., Devillers, L., Lamolle, M., Mancini, M. & Pelachaud, C. (2005). Levels of Representation in the Annotation of Emotion for the Specification of Expressivity in ECAs. In T. Panayiotopoulos, editor, *Proceedings of the 5th International Working Conference on Intelligent Virtual Agents (IVA'2005)*, Kos, Greece, September 12-14, 3661 of LNAI, 405–417. Berlin: Springer-Verlag.
- Mellish C., Reape M., Scott D., Cahill L., Evans R. & Paiva D. (2004). A Reference Architecture for Generation Systems, *Natural Language Engineering*, 10, 227-260.
- Melson, G.F., Kahn, P., Beck, A., Friedman, B., Roberts, T. (2004). Children's understanding of robotic and living dog. *IAHAIO People and animals: A timeless relationship*, Glasgow, UK, 71.
- Miklósi, A, Kubinyi E, Topál, J, Gácsi, M., Virányi, Zs., Csányi, V. (2003). A simple reason for a big difference: wolves do not look back at humans but dogs do. *Current Biology*, 13, 763-766.
- Miklósi, Á., Polgárdi, R., Topál, J., Csányi, V. (2000). Intentional behaviour in dog-human communication: An experimental analysis of 'showing' behaviour in the dog. *Anim.Cogn.*, 3, 159-166.
- Mitchell, R W (2001). Americans'talk to dogs: similarities and differences with talk to infants. *Research on Language and Social Interactions*, 34, 183-210.
- Molnár Cs., Pongrácz, P., Dóka, A., Miklósi Á. (2008). Children can understand man's best friend: Emotional and contextual classification of dog barks by pre-adolescents. *Dev. Psychobiol* (submitted)
- Mota, S. & Picard, R. (2003). Automated posture analysis for detecting learner's interest level. *Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction*, CVPR HCI.

- Nakata, T., Mori, T. & Sato, T. (2001). Quantitative Analysis of Impression of Robot Bodily Expression based on Laban Movement Theory. *Journal of Robotics Society of Japan*, 19(2), 104-111.
- Namysl, M. (2008a). *Vision system in human emotions recognition*. Master's Thesis, Institute of CECR WRUT, Wroclaw, (in Polish).
- Namysl, M. (2008b) A concept of vision system for emotion recognition from facial images. In: *Problemy Robotyki*, ed. K. Tchon, Warszawa (in Polish).
- Orlov. Saya, <http://www.cs.bgu.ac.il/örlovm/teaching/saya/>
- Pantic, M. (2006). Face for Ambient Interface. *Lecture Notes in Artificial Intelligence*, 3864, 35-66.
- Pantic, M. & Bartlett, M. (2007). Machine Analysis of Facial Expressions, in *Face Recognition*, K. Delac & M. Grgic, eds., Vienna, Austria: I-Tech Education and Publishing, 377-416.
- Pantic, M. & Rothkrantz, L. (2000). Automatic Analysis of Facial Expressions – The State of the Art. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(12), 1424-1445.
- Pelachaud, C. (2005). Multimodal expressive embodied conversational agents. *MULTIMEDIA '05: Proceedings of the 13th annual ACM International Conference on Multimedia*. ACM Press, 683–689.
- Pollick, F., Paterson, H., Bruderlin, A. & Sanford, A. (2001). Perceiving affect from arm movement. *Cognition*, 82(2).
- Pongrácz, P., Miklósi, A., Csányi, V. 2001. Owners' beliefs on the ability of their pet dogs to understand human verbal communication. A case of social understanding. *Current Cognitive Psychology*, 20, 87-107.
- Pongrácz, P., Miklósi, Á., Molnár, Cs., Csányi, V. (2005). Human listeners are able to classify dog barks recorded in different situations. *Journal of Comparative Psychology*, 119, 136-144.
- Pongrácz, P., Miklósi, Á., Timár-Geng K., Csányi V. (2004). Verbal attention getting as a key factor in social learning between dog (*Canis familiaris*) and human. *J of Comp. Psychology*, 118, 375-383.
- Pransky, J. (2001). AIBO - the No. 1 selling service robot. *Industrial Robot: An International Journal*, 28, 24-26.
- Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition, *Proc. IEEE*, 77(2), 257-286.
- Reimondo, A. Haar cascades, <http://alereimondo.no-ip.org/OpenCV/34>
- Reis, P., Matias, J. & Mamede, N. (1997). Edite – A Natural Language Interface to Databases: a New Dimension for an Old Approach. *Proceeding of the Fourth International Conference on Information and Communication Technology in Tourism (ENTER' 97)*.
- Reiter E., Dale R. (2000). *Building Natural-Language Generation Systems*, Cambridge University Press, ISBN 0521620368.
- Rich C., Sidner C.L., Lesh N. (2001). Collagen: Applying collaborative discourse theory to human-computer interaction, *AI Magazine*, 22(4), Special issue on Intelligent User Interfaces, 15- 25.
- Russell, J. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39, 1161–1178.

- Sakaguchi, K., Jonsson, G.K. & Hasegawa, T. (2005). Initial interpersonal attraction between mixed-sex dyad and movement synchrony. L. Anolli, S. Duncan, M. Magnusson and G. Riva, eds., *The hidden structure of social interaction: From neurons to culture patterns*. IOS Press.
- Shibata, S., Yamamoto, T., Jindai, M. & Shimizu, A. (2003). A Fundamental Approach of Avoidance Planning of Robots Considering Human Emotions. *JSME International Journal Series C*, 46(1), 270-277.
- Shinozaki, K., Iwatani, A. & Nakatsu, R. (2007). Concept and construction of a robot dance system. *The International Journal of Virtual Reality*, 6(3), 29-34.
- Soproni, K., Miklósi, Á., Topál, J., Csányi, V. (2001). Comprehension of human communicative signs in pet dogs. *Journal of Comparative Psychology*, 115, 122-126.
- Soproni, K., Miklósi, Á., Topál, J., Csányi, V. (2002). Dogs' responsiveness to human pointing gestures. *Journal of Comparative Psychology*, 116, 27-34.
- Strom, V., Clark, R. & King, S. (2006). Expressive prosody for unit-selection speech synthesis. *Interspeech*, Pittsburgh, U.S.A.
- Syrdal, A.K., Kim Y.-J. (2008) Dialog speech acts and prosody: Considerations for TTS. *Proceedings of Speech Prosody 2008*, 661-665, Campinas, Brazil.
- Szetei, V., Miklósi, Á., Topál, J., Csányi V. (2003). When dogs seem to lose their nose: an investigation on the use of visual and olfactory cues in communicative context between dog and owner. *Applied Animal Behavior Science*, 83, 141-152.
- Szöke, I. et al. (2005). Comparison of Keyword Spotting Approaches for Informal Continuous Speech, *Proceedings of Interspeech'05*, Lisbon, Portugal.
- T. Kanade, J. F. Cohn. Automated Face Analysis, <http://www.cs.cmu.edu/afs/cs/project/face/www/facs.htm>
- Tchou, K., Arent, K., Janiak, M., Kedzierski, J., Kreczmer, B., Malek, L., Muszynski, R., Oleksy, A. & Wnuk, M. (2008). *Toward a robotic companion design*. WRUT LIREC Report, Inst. Comp. Eng., Contr. and Robotics, Wroclaw University of Technology, Poland.
- Turner, D.C., Ribi, F.N., Yokoyama, A. (2004). A comparison of children's behaviour toward a robotic pet and a similar sized, live dog over time. *IAHAIO People and animals: A timeless relationship*, Glasgow, UK, 68.
- Valstar, M., Gunes, H. & Pantic, M. (2007). How to distinguish posed from spontaneous smiles using geometric features. *ACM International Conference on Multimodal Interfaces (ICMI'07)*, 38-45, Nagoya, Japan.
- Valve Developer Community. Facial Expressions Primer, http://developer.valvesoftware.com/wiki/Facial_Expressions_Primer
- Vanger, P., Hoenlinger, R. & Haken, H. Computer aided generation of prototypical facial expressions of emotion, <http://www.dgps.de/fachgruppen/methoden/mpr-online/issue4/art3/article.html>
- Virányi, Zs., Topál, J. Gácsi, M., Miklósi, Á. & Csányi, V. (2004). Dogs can recognize the focus of attention in humans. *Behavioural Processes*, 66, 161-172.
- Volpe, G. (2003). *Computational models of expressive gesture in multimedia systems*. Ph.D. Dissertation, Faculty of Engineering, University of Genova.
- Vukadinovic, D. & Pantic, M. (2005). Fully automatic facial feature point detection using Gabor feature based boosted classifiers, *Proc. IEEE Int'l Conf. Systems, Man and Cybernetics*, 1692-1698.
- Wallbott, H. (1998). Bodily expression of emotion. *European Journal of Social Psychology*, 28(6), 879-896.

Walters, M., Syrdal, D., Dautenhahn, K., te Boekhorst, R. & Koay K. L. (2008). Avoiding the Uncanny Valley – Robot Appearance, Personality and Consistency of Behaviour in an Attention-Seeking Home Scenario for a Robot Companion. *Journal of Autonomous Robots*, 24(2). 159-178.

Weiss, C. Oliveira, L. C. Paulo, S. Duarte Mendes, C. M., Dias Mestre Figueira, L.A., Vala, M., Sequeira, P., Paiva, A., Vogt, T., Andre, E. (2007). ECIRCUS: Building Voices for Autonomous Speaking Agents, *6th Speech Synthesis Workshop*, ISCA, Bonn.

Weizenbaum, J. (1966). ELIZA - A Computer Program for the Study of Natural Language Communication between Man and Machine. *Communications of the ACM*, 9(1), 36-45.

Young, S. (2002). Talking to machines (statistically speaking), *Proc. Int. Conf. Spoken Language Processing*, Denver, CO.

Zeng, Z., Pantic, M., Roisman, G. & Huang, T. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, TPAMI-2007-09-0496, To Appear.

Zhao, L. (2001). *Synthesis and Acquisition of Laban Movement Analysis Qualitative Parameters for Communicative Gestures*. Ph.D. Dissertation, University of Pennsylvania.

7 Appendices

This Section contains additional information on some of the experiments presented in this deliverable.

7.1 “Pleo – first contact” experiment

7.1.1 Manual

Set-up

Group 1	Dauer	Group 2	Dauer
Introduction			
Q 1	10 Min	Q 1	10
Watching Videos (Pleo Fight) Short Interview: What did you feel while watching the video?	10	Interaction with Pleo <ul style="list-style-type: none"> ○ 5 minutes: free interaction (experimenter leaves room) ○ 5 minutes: s structured interaction - subject tries to fulfil tasks with Pleo (tasks are specified on paper in front of them on the table) 	10
Interaction with Pleo <ul style="list-style-type: none"> ○ 5 minutes: free interaction (experimenter leaves room) ○ 5 minutes: structured interaction - subject tries to fulfil tasks with Pleo (tasks are specified on paper in front of them on the table) 	10	Watching Videos (Pleo Fight) Short Interview: What did you feel while watching the video?	10
Joint video analysis of interaction, semi-structured interview on the interaction with PLEO	20	Joint video analysis of interaction, semi-structured interview on the interaction with PLEO	20
Q 2	10	Q 2	10
Time:	60Min	Time:	60 Min

Introduction Experimenter

- We want to know how you like Pleo. It is you personal opinion we are interested in, not any wrong or right answers or behaviours...
- Duration: 60 mins.
- Videotaping for scientific purpose...

- Data security issues
- Questions?

Q-I

- Experimenter goes to the next room but offers to answer questions if there are any.

Video „Fight“

- Questions: What did you think during the video? What did you feel?

Interaction 1: free

- Experimenter explains that subjects now have five minutes to get to know PLEO while she is in the next room.
- In the next room, experimenter takes the time and goes back in the room after 5 mins interaction

Interaction 2: structured

- Experimenter hands subject paper with 3 tasks:
 1. Try to get PLEO to move towards you.
 2. Try to feed PLEO.
 3. Try to make PLEO fall asleep.
- Experimenter explains that subjects now have five minutes for the tasks while she is in the next room.
- In the next room, experimenter takes the time and goes back in the room after 5 mins.

After time is over, experimenter takes PLEO out of the room and turns it off *outside* the room.

Joint video analysis – semi-structured interview

- Experimenter asks subject to remember what he/she felt when interacting with PLEO and explain it, while the two of them watch the video of the interaction.
- Questions:
 - What was your first impression of PLEO?
 - How did you feel when PLEO first started to move and make noises? Were you surprised? What was surprising to you? Did you feel uncomfortable?
 - How did you feel during feeding it, luring it, putting it down (tasks)? Did it go well? How?
 - Did PLEO meet you expectations? What do you like / dislike with PLEO? What should it be able to do (in order to be more interesting)? Do you think / feel differently about PLEO after the interaction? Why?

7.1.2 Questionnaire No. 1

Time / Date	Participant
-------------	-------------

Demographic Data

Age: _____ yrs.	<input type="checkbox"/> female <input type="checkbox"/> male
-----------------	---

Questionnaire

In the following we will ask you some questions. Please try to answer openly and truthfully, there are no wrong or right answers, only your personal opinion. Thank you very much!

1. What do you use your computer for most of the time? (e.g. e-mail, internet, games, music)	<hr/> <hr/> <hr/> <hr/> <div style="text-align: right;"><input type="checkbox"/> I don't have a computer</div>
--	--

2. Do you enjoy exploring technical devices?	<div style="text-align: right;"><input type="checkbox"/> yes <input type="checkbox"/> no</div>
---	--

3. I see myself as extraverted, enthusiastic	<table style="width: 100%; border: none;"> <tr> <td style="text-align: left;">not true at all</td> <td style="text-align: right;">absolutely true</td> </tr> <tr> <td colspan="2" style="text-align: center;">1 --- 2 --- 3 --- 4 --- 5 --- 6 --- 7 --- 8 --- 9 --- 10</td> </tr> </table>	not true at all	absolutely true	1 --- 2 --- 3 --- 4 --- 5 --- 6 --- 7 --- 8 --- 9 --- 10	
not true at all	absolutely true				
1 --- 2 --- 3 --- 4 --- 5 --- 6 --- 7 --- 8 --- 9 --- 10					

4. I see myself as critical, quarrelsome	<table style="width: 100%; border: none;"> <tr> <td style="text-align: left;">not true at all</td> <td style="text-align: right;">absolutely true</td> </tr> <tr> <td colspan="2" style="text-align: center;">1 --- 2 --- 3 --- 4 --- 5 --- 6 --- 7 --- 8 --- 9 --- 10</td> </tr> </table>	not true at all	absolutely true	1 --- 2 --- 3 --- 4 --- 5 --- 6 --- 7 --- 8 --- 9 --- 10	
not true at all	absolutely true				
1 --- 2 --- 3 --- 4 --- 5 --- 6 --- 7 --- 8 --- 9 --- 10					

5. I see myself as dependable, self-disciplined	<table style="width: 100%; border: none;"> <tr> <td style="text-align: left;">not true at all</td> <td style="text-align: right;">absolutely true</td> </tr> <tr> <td colspan="2" style="text-align: center;">1 --- 2 --- 3 --- 4 --- 5 --- 6 --- 7 --- 8 --- 9 --- 10</td> </tr> </table>	not true at all	absolutely true	1 --- 2 --- 3 --- 4 --- 5 --- 6 --- 7 --- 8 --- 9 --- 10	
not true at all	absolutely true				
1 --- 2 --- 3 --- 4 --- 5 --- 6 --- 7 --- 8 --- 9 --- 10					

6. I see myself as anxious, easily upset	<p>not true at all absolutely true</p> <p>1 — 2 — 3 — 4 — 5 — 6 — 7 — 8 — 9 — 10</p>
7. I see myself as open to new experiences, complex	<p>not true at all absolutely true</p> <p>1 — 2 — 3 — 4 — 5 — 6 — 7 — 8 — 9 — 10</p>
8. I see myself as reserved, quiet	<p>not true at all absolutely true</p> <p>1 — 2 — 3 — 4 — 5 — 6 — 7 — 8 — 9 — 10</p>
9. I see myself as sympathetic, warm	<p>not true at all absolutely true</p> <p>1 — 2 — 3 — 4 — 5 — 6 — 7 — 8 — 9 — 10</p>
10. I see myself as disorganized, careless	<p>not true at all absolutely true</p> <p>1 — 2 — 3 — 4 — 5 — 6 — 7 — 8 — 9 — 10</p>
11. I see myself as calm, emotionally stable	<p>not true at all absolutely true</p> <p>1 — 2 — 3 — 4 — 5 — 6 — 7 — 8 — 9 — 10</p>
12. I see myself as conventional, uncreative	<p>not true at all absolutely true</p>

	1 -- 2 -- 3 -- 4 -- 5 -- 6 -- 7 -- 8 -- 9 -- 10
--	---

13. Do you like animals?	not true at all absolutely true
	1 -- 2 -- 3 -- 4 -- 5 -- 6 -- 7 -- 8 -- 9 -- 10

14. Do you own pets?	<input type="checkbox"/> yes <input type="checkbox"/> no
	If „yes“, which pets? _____ _____

15. Did you have contact to robots before?	<input type="checkbox"/> yes <input type="checkbox"/> no
	If „yes“, please describe the kind of contact: _____ _____ _____

16. Are you able to program your car stereo?	<input type="checkbox"/> yes <input type="checkbox"/> no
---	--

17. What is you first thought when you hear the word “robots”?	_____

<p>18. Please rate the following statement:</p> <p>„I would use a robot“</p>	<p>not true at all absolutely true</p> <p>1 --- 2 --- 3 --- 4 --- 5 --- 6 --- 7 --- 8 --- 9 --- 10</p>
--	---

<p>19. What do you know about PLEO?</p>	<hr/> <hr/> <hr/>
---	-------------------

<p>20. What are your expectations regarding PLEO?</p>	<hr/> <hr/> <hr/>
---	-------------------

<p>21. Look at the picture of PLEO: Do you think PLEO has personality?</p>	<p style="text-align: right;"><input type="checkbox"/> yes <input type="checkbox"/> no</p> <p>If „yes“, please tick the traits that characterise PLEO. This PLEO is ...</p> <p><input type="checkbox"/> ... trusting</p> <p><input type="checkbox"/> ... demanding</p> <p><input type="checkbox"/> ... playful</p> <p><input type="checkbox"/> ... stubborn</p> <p><input type="checkbox"/> ... curious</p> <p><input type="checkbox"/> ... cheerful</p> <p><input type="checkbox"/> ... intelligent</p> <p><input type="checkbox"/> ... self-contained</p>
--	---

	<input type="checkbox"/> ... likeable
--	---------------------------------------

22. Can you imagine having PLEO at your home?	<input type="checkbox"/> yes <input type="checkbox"/> no
--	--

23. What would you pay for PLEO?	_____ EUR
---	-----------

7.1.3 Questionnaire No. 2

Date / Time	Participant
-------------	-------------

Questionnaire

Please try to answer the question as open and truthful as possible. Thank you very much!

1. In your opinion, what is PLEO? (e.g. machine? toy? pet? artificial companion?...)	<hr/> <hr/> <hr/>
--	-------------------

2. did you have fun interacting with PLEO?	<input type="checkbox"/> yes <input type="checkbox"/> no
---	--

3. Do you think PLEO has personality?	<input type="checkbox"/> yes <input type="checkbox"/> no
	<p>If „yes“, please tick the traits that characterise PLEO. This PLEO is ...</p> <p><input type="checkbox"/> ... trusting</p> <p><input type="checkbox"/> ... demanding</p>

	<input type="checkbox"/> ... playful <input type="checkbox"/> ... stubborn <input type="checkbox"/> ... curious <input type="checkbox"/> ... cheerful <input type="checkbox"/> ... intelligent <input type="checkbox"/> ... self-contained <input type="checkbox"/> ... likeable
--	--

4. Would you like to take PLEO home and keep it?	<input type="checkbox"/> yes <input type="checkbox"/> no
--	--

5. How long do you think your interest in PLEO would persist if you had it all the time?	_____
--	-------

6. How much would you pay for PLEO?	_____ EUR
-------------------------------------	-----------

7. In general, do you think robots are useful?	not true at all absolutely true 1 --- 2 --- 3 --- 4 --- 5 --- 6 --- 7 --- 8 --- 9 --- 10
--	--

8. Do you think your general opinion about robots has changed due to your interaction with PLEO?	not true at all absolutely true 1 --- 2 --- 3 --- 4 --- 5 --- 6 --- 7 --- 8 --- 9 --- 10
--	--