

humaine

D9h

Mid-term report on usability exemplar progress

Workpackage 9 Deliverable



Date: 30th November 2006

IST project contract no.	507422
Project title	HUMAINE Human-Machine Interaction Network on Emotions
Contractual date of delivery	<i>November 30, 2006</i>
Actual date of delivery	<i>November 30, 2006</i>
Deliverable number	D9h
Deliverable title	Mid-term report on usability exemplar progress
Type	Report
Number of pages	26
WP contributing to the deliverable	WP 9
Task leader	KTH
Author(s)	See page 5
EC Project Officer	Philippe Gelin

Address of lead author:

Kristina Höök
 Department of Computer and Systems Sciences
 Stockholm University/Royal Institute of Technology (KTH)
 Forum 100
 164 40 Kista
 Sweden

Table of Contents

1	THE PLACE OF THIS REPORT WITHIN HUMAINE.....	5
2	BRIEF OVERVIEW OF WORKPACKAGE 9 AND THE EXEMPLAR.....	6
2.1	The field covered by Workpackage 9	6
2.2	The research objectives.....	7
2.2.1	Main elements of the exemplar	8
2.2.2	How the elements link to each other	8
3	PROGRESS TOWARD MAIN ELEMENTS OF EXEMPLAR.....	10
3.1	Element 1: Criteria for usable affective interaction systems	10
3.2	Element 2: Evaluation metrics	11
3.3	Element 3: Existing user-centred methods for design and evaluation	12
3.4	Element 4: New methods for design and evaluation	14
3.5	Affective Interface Development Use Case	17
4	PLANNED PROGRAM OF RESEARCH.....	22
4.1	Element 1: Criteria for usable affective interaction systems	22
4.2	Element 2: Evaluation metrics for criteria	22
4.3	Element 3: Existing user - centred methods for design and evaluation	22
4.4	Element 4: New methods for design and evaluation	23
5	REFERENCES.....	25

1 The place of this report within HUMAINE

The HUMAINE Technical Annex identifies a common pattern that is followed by most of the project's workpackages

The measure of success will be the ability to generate a piece of work in each of the areas which exemplifies how a key problem in the area can be solved in a principled way; and which also demonstrates how work focused on that area can integrate with work focused on the other areas. We call these pieces of work *exemplars*. The exact form of an exemplar is not prespecified: it may be a working system, but it might also be a well-developed design, or a representational system, or a method for user-centred design. (p 4)

To that end, each thematic group will work out a proposal for common action, embodied in one or more exemplars to be built during the second half of the funding period (p.16)

The process will begin with production by each thematic group of a review of key concepts achievements and problems in its thematic area; and drawn from the review, an assessment of the key development goals in the area. This review and assessment will be circulated to the whole network for discussion and comment, aimed both at building understanding of basic issues across areas, and at identifying the choices of goal that would be most likely let the different groups achieve complementary developments. That consultation phase will provide the basis for deliverables in month 11, which describe in some detail a few alternatives that might realistically be chosen as exemplars in each area, and their linkages to issues in other thematic areas. A decision and planning period will follow, involving consultation within and between thematic areas, leading to presentations at the second plenary conference, which will describe a single exemplar that has been chosen for development in each area, and the way work on the exemplar will be divided across institutions. The remainder of the project will be absorbed in developing the chosen exemplar. (p. 21)

The consultation phase has now ended. Near-final plans were presented to the whole network at the Plenary in May 2005, and adjustments have been made accordingly.

This deliverable reports progress made on the exemplar for WP9.

Ethical issues affect the whole of HUMAINE, but rather than repeating essentially similar points in multiple deliverables, they will be handled coherently in a single document, D0o (Science and Society).

The following persons have contributed to the work reported in the deliverable:

Jarmo Laaksolahti, Kristina Höök, Fiorella de Rosis, Sabine Payr, Benoit Morel, Catherine Pelachaud, Asimina Vasalou, Tanja Bänziger, Stephen Westerman, Ana Paiva

The institutions that have contributed are:

KTH, GERG, BARI, PARIS8, CANTOCHE, OFAI, IMPERIAL, LEEDS, IST

2 Brief overview of Workpackage 9 and the exemplar

While the rest of HUMAINE will produce knowledge and theories of many different aspects of emotional systems (ECA behaviours, emotion recognition, interaction principles, exemplar databases, and emotion theory) the overall goal of WP9 is to use that knowledge to produce design and evaluation methods able to deliver functioning affective end-user applications. Such applications allow users to become affectively involved in the course of interaction with the system.

Our strategy for working within an end-user application framework to evolve the usability of affective systems can be divided into two focus areas:

1. Ends: Determining qualities and criteria that demark emotional systems as usable and evoking the desired experiences for users.
2. Means: Forging process—design methods, project goals and evaluation strategies — that will steer a project towards producing a ‘successful’ affective application.

The WP9 exemplar is designed to develop, explore and determine key criteria and methods for designing and evaluating affective systems

2.1 The field covered by Workpackage 9

The area covered by this workpackage is described in the Technical Annex, particularly in Section 6.2, and in more depth in the review and assessment document for the workpackage. We summarise the area here so that the deliverable can be read as a stand-alone document.

Regarding Ends, we do not believe there are simple, replicable correlations between system properties and all end users’ experiences. The experience of using an emotional system is not a property of the system itself, but rather is something that arises in the interaction between user and system. To quote Sengers and colleagues (2004):

“Rather than experience as something to be poured into passive users, we argue that users actively and individually construct meaningful human experiences around technology. They do so through a complex process of interpretation, in which users make sense of the system in the full context of their everyday experience.”

Along the same lines, Suchman has criticized the cognitivistic reductionist basis for the field of affective computing (Suchman, 2002), reducing human emotional responses to discrete, universal component parts, arising from some underlying state, and offers examples of alternative designs that put emotional interaction at the core but where: “we might differently conceptualise affective encounters at the interface: as irreducibly contingent meetings of particularly situated persons with equally particular, dynamic, and culturally inflected things.” As we discuss below, usability of emotional systems and the process of arriving at functioning emotional systems need to be sensitive to how end-users are part in co-constructing the experience and be part of this dynamic and cultural process.

Usability traditionally focuses on goals such as effectiveness, efficiency, safety, utility, learnability, and memorability. These objective usability goals contrast with user experience goals, which cover subjective qualities such as being fun, rewarding, motivating, satisfying,

enjoyable, and helpful. Usability goals and user experience goals often stand in complex relationships, involving tradeoffs such as safety vs. fun or efficiency vs. enjoyability (Preece et al., 2002). Introducing emotion thus raises many new dimensions for research on usability to address. It becomes, for instance, a serious issue whether users feel a system is 'sympathetic' or morally acceptable, whether it engages them emotionally. Furthermore, emotions address directly inherently adaptive faculties of humans, posing challenges for methods of user studies and artefact design.

Work Package 9 will take a primarily qualitative, situated/contextual approach to measuring Ends. A sub-group of Work Package 9 may also focus on evaluating and developing measures aimed at isolating meaningful evaluative variables and testing various components of systems against these variables.

Regarding Means, we believe that evaluation of affective systems is vital not just at the end of the design process, but as an integral part of the design process from the beginning. Having the ability to bounce early intuitions and design sketches off of real users can make key contributions to the evolution of a truly engaging end application, and may even inform the affective theory that led to the application itself.

Work in the field of HCI has made great strides toward merging design and evaluation of productivity-oriented systems (see the goals section of this document for some relevant citations); we seek to extend and adapt participatory design methods and approaches to the needs and desired outcomes of affective system projects. For example, encouraging nonverbal participation and evaluation, and using physical prototypes and fake systems operated by humans. We believe this holistic and integrated approach will not only lead to better system designs, but may also provide important 'in situ' insights back to affective theory makers.

Thus, it is not the primary intent of WP9 to test computational emotion theories or evaluate how particular parameters for e.g. raising an eyebrow in an ECA should work. Instead, the task of WP9 will be to produce design process and evaluation theory, methods, and measures needed to take all that knowledge into the design process and through it, to produce emotional systems that will engage with end-users to create compelling experiences.

2.2 The research objectives

The exemplar proposed for WP9 is a *Framework for Design and Evaluation of Usable Affective Interaction Applications*. Similar to how HCI has benefited from user-centred design processes, our aim is to test and further develop such methods for the area of affective interaction. This in turn will hopefully push the research frontier forwards and serve as an important exemplar to others in the field. The framework contains an assortment of techniques and methods for various aspects of the design process including generating initial ideas for affective applications, refining ideas into systems/products or validating the affective interaction loop in finished systems/products. As a whole the framework is intended to be a broad resource for designing and evaluating affective systems – representing first and foremost a user-centred perspective on affective interaction applications. The tools and methods of the framework will be applied and evaluated in a variety of situations/applications/domains. Records from the sessions will provide valuable guidance for future users of the tools and methods regarding their proper usage and expected results.

2.2.1 Main elements of the exemplar

The exemplar consists of four main elements.

Criteria for usable affective interaction systems

Develop an understanding for what makes affective interaction systems successful and formulate some criteria. These criteria will not be objective, independently measurable entities, but will make sense relative to the specific application domain, aim to capture subjective experiences of the application, and foremost, be related to the designer's intention for the application.

Evaluation metrics for criteria

Translate criteria for affective interaction systems into usable metrics and methods for eliciting them.

Existing user-centred methods for design and evaluation

Condense experiences from applying existing user-centred design methods to the design and evaluation of affective interaction systems – which methods work and which do not? Examples of such methods include body-storming (Oulasvirta et al, 2003), interaction re-labelling (Djajadiningrat, et al., 2000) and personas (Cooper, 1999).

New methods for design and evaluation

Develop new methods for capturing unique aspects of affective interaction that can be used during an iterative design-evaluate-redesign process. Methods that will be investigated include: a sensual method for non-verbal mediation of affective state, a Wizard-of-Oz environment for multimodal emotional interaction, and an extended think-aloud protocol designed to capture emotional interactions.

2.2.2 How the elements link to each other

Design and evaluation is a broad area, participants have differing interests and backgrounds, different use-cases require different methods, resources, and study angles. We are interested in the whole design-evaluation process (see Figure 1) and the subtasks of WP9 address different aspects of the process. The exemplar can still be viewed as a single piece though, since it explores one and the same idea: that a user-centred perspective will help focus the design process and create final application systems that involve users affectively. This exemplar will be a touchstone for integrating this perspective throughout the whole life-cycle of a system development process. While we strongly believe that this is the best kind of exemplar for WP9, there are still many white spots on the map. We do not know how to capture subjective experiences and affective involvement in such a way that it can provide feedback into the design process. A user-centred design perspective and development philosophy is most probably a good path to explore, but little is, as of yet, known in what to design for (flow, pleasure, fun, experiences, excitement, fear), and how to best make users involved.

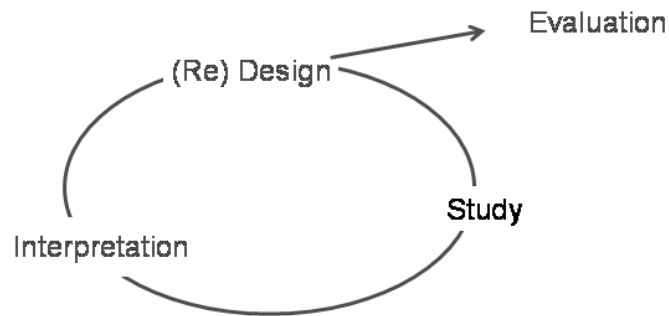


Figure 1 User-centred design framework

Positioning the subtasks within the user-centred design framework above ensures that subtasks are working towards a common goal. The nature of the framework with its different methods and tools all sharing the same conceptual framework also helps in this respect. The four subtasks follow on one-another. Once criteria for the success of a specific application project are established, there has to be corresponding measurements to know whether the design goals are met according to the same criteria. The process to ensure such a design and evaluation cycle in turn needs methods. All these tasks together will help shape the framework produced by this WP.

3 Progress toward main elements of exemplar

3.1 Element 1: Criteria for usable affective interaction systems

In this element we develop an understanding for what makes affective interaction systems successful and formulate some criteria. These criteria will not be objective, independently measurable entities, but will make sense relative to the specific application domain, aim to capture subjective experiences of the application, and foremost, be related to the designer's intention for the application.

- a. Compilation of a commented literature list on ECA evaluation studies and of initial definition criteria for empathic ECAs.

In joint work of WP9 with WP7 Element 3, systems that employ affective agent architectures were analysed by OFAI with respect to their explicit and implicit design aims. By describing the target scenarios of current systems, we tried to capture motivation for and purpose of building affective systems, as well as the specifics of the envisioned deployment¹. Such scenario descriptions form a sound basis to pinpoint the emotional potential of a specific application setting: which kind of emotional phenomena are/can be expected to occur in the interaction between human and system, and which cannot. Regarding the affective agents in the system, the scenario description can help answer the question: which emotional phenomena are crucial to be covered and which may only be simulated or portrayed (Rank & Petta, 2006a). Thus, it relates the different perspectives of WP7 and WP9.

Further, an understanding of motivation and purpose of a system is the necessary foundation of devising evaluation criteria for these systems (Rank & Petta, 2006b).

- b. Explore success criteria for affective interaction wrt the specific design aims.

eMoto is a mobile emotional messaging system that we have previously described (see e.g deliverable d9d). With it, users can compose messages through using emotion-related gestures as input, rendering a message background of colours, shapes and animations expressing the emotional content. The design intent behind eMoto is that it should be engaging physically, intellectually and socially, and allow users to express themselves emotionally in all those dimensions, involving them in an affective loop experience. Note that these criteria are very specific for the application at hand and may not always transfer to other applications, use contexts or user groups. In a study of the system five friends used eMoto for two weeks. The study method, which we name *in situ informants*, helped us enter and explore the *subjective and distributed experiences* of use, as well as how emotional communication unfolds in everyday practice when channelled through a system like eMoto. The *in situ informants* are on the one hand users of eMoto, but also spectators, that is close friends who observe and document user behaviour. Design conclusions include the need to support the sometimes fragile communication rhythm that friendships require – expressing memories of the past, sharing the present and planning for the future. We saw that emotions are not singular state that exist within one person alone, but permeates the total situation, changing and drifting as a process between the two friends

¹ <http://emotion-research.net/restricted/wp7/WP7Element3/wp7e3wiki/>

communicating. We also gained insights into the under-estimated but still important physical, sensual aspects of emotional communication. Experiences of the in situ informants method include the need to involve participants in the interpretation of the data obtained, as well as establishing a closer connection with the spectators.

An article describing the design and evaluation of eMoto (*In situ informants exploring an emotional mobile messaging system in their everyday practice*, (Sundström et al, 2007)), was accepted to a forthcoming special issue of International Journal of Human-Computer Studies (IJHCS) named “Evaluating Affective Interactions: Innovative Approaches and Future Directions”.

3.2 Element 2: Evaluation metrics

In this element we translate criteria for affective interaction systems into usable metrics and methods for eliciting them.

- a. Construction of a collection of design, test and evaluation strategies and methods for empathic ECAs.

Work undertaken by OFAI will contribute to the collection of design and evaluation methods for (E)CAs (embodied or non-embodied conversational agents). As in the related work in WP7 element 4.2, research focuses on pervasive emotional phenomena in communication and social relationship and their controlling function in dialog. The ongoing critical review takes into account both quantitative (e.g. measurement of dialog length, number of turns, pauses, success rate, distribution of turn types, cf. Rienks & Heylen 2006, Kopp 2006) and qualitative methods (e.g. conversation and semantic-functional analysis, cf. Fischer 2006) and appraises their respective contributions to gaining insights into the kinds of (social) relationships between human users and their dialog partners, be they actually ECAs or non-embodied dialog systems, or “wizards” in various experimental settings (from strictly script-following wizards to those with various degrees of freedom, where the study of the wizard’s behaviour in turn becomes a design strategy (Wallis 2005, Porzel 2006). Comparative (re)analysis of human-human dialogs is necessary for reference, both for similarities and dissimilarities: the phenomena that do not or rarely occur in human-machine verbal communication (e.g. interruptions, speech particles) are as revealing as those that do occur. Furthermore, given that neither humans nor machines enter into the dialog without preconceptions (Fischer 2006), to which the actual context of the communication has to be added as a factor, extra-conversational analysis (collection and analyses of data on users’ attitudes of the system in question, e.g. through polarity profiles, semantic differential) is also dealt with. (Payr 2006).

- b. Assessment of subjective evaluation metrics

See 3.1.b.

3.3 Element 3: Existing user-centred methods for design and evaluation

In this element we condense experiences from applying existing user-centred design methods to the design and evaluation of affective interaction systems – which methods work and which do not? Examples of such methods include body-storming (Oulasvirta et al, 2003), interaction re-labelling (Djajadiningrat, et al., 2000) and personas (Cooper, 1999).

- a. WP9 Brainstorming session on generating application ideas using existing design and evaluation methods

Prior to the WP9 workshop in January 2006 KTH organised a design workshop with participants from several HUMAINE members as well as some external researchers. The purpose of the design workshop was to demonstrate and teach user centred design methods by using them in practice. During two intense days participants experienced a full cycle in the user centred design process and went from getting to know their user, to generating initial ideas to creating testable low fidelity prototypes and evaluating them. The design exercise was framed by our choice of scenario, to design applications that involved confiding thoughts and emotions in some way, and the fact that we wanted the applications to involve an ECA. The exercise actually started a month prior to the workshop when participants, as the intended target user group for the design exercise, received *cultural probes* (Gaver et al, 1999): a package consisting of postcards with tasks written on them, pens, pieces of cloth, diaries and in some cases a disposable camera. The purpose of the probe was to collect data about the user group, in their daily life, that could inform the designs that were later created. At the design workshop the first step of the process was to familiarise participants with the probe data and from that generate keywords belonging to three categories: situation of use, character of ECA and technology used. The group was then split into two and each group proceeded to create an initial set of ideas by randomly combining keywords from each of the categories. After that the groups had an opportunity to develop/critique each others ideas using the *six thinking hats* method (DeBono, 1985) which provides a structured way of examining an idea from different viewpoints. Using bodystorming (Oulasvirta et al, 2003) participants tried out 1-3 ideas to get a better feel for how the ideas would work out in a real situation. At the end of this process, and the first day, the initial set of ideas had been condensed to one idea per group that they continued to work on. During the second day each group proceeded to create testable low-fidelity prototypes of their applications, using the *tiny fingers* method (Rettig, 1994), as well as setting up small scale user studies with real users (users that were not participating in the design exercise) to test their applications. The result of this exercise was reported during the WP9 workshop that started a day later.

The design workshop included participants from both academia as well as industry, and was greatly appreciated by all. Participants reported having gained a greater understanding of the user centred design process as well as how they would be able to apply it to their own work to create better affective interaction applications.

For PARIS8 being able to get acquainted with several methods of user centred design was extremely enlightening. On the one hand it provides clear evidence of the importance of going through a design process when developing an ECA application. Trying out several methods help us finding out the strengths and weaknesses of ideas. Once having gone through the cultural probe, the six thinking hats, bodystorming and the tiny fingers we could elaborate a first description of an application. We had then the possibility to enact the idea of our application and call for a user to try it and report on his experience. Even

though, the user study was on a very small scale, it did point out what went wrong in the design process, what should be left out but it also shows the user was able to get affectively involved.

b. Application of existing user-centred methods to use-cases

The process used by CANTOCHE to choose and determine an ECA in our different projects is based on a questionnaire we provide our clients. Four chapters composed the questionnaire in order to get information about (1) the company (mission, value, culture, ..) (2) the project (goal, users,..), (3) general questions about the mission of the character and (4) description of the character they want. From this questionnaire we try to help our client by giving advice and by providing several options corresponding to the physical aspect, the voice, the way the ECA can perform and behave. We use to compare this process to the recruitment of a real person. (Morel, 2004).

The WP9 workshop and the exercises we have done help us to learn new methods we decided to apply to our customers. We have been interested in particular by the *six thinking hats* method as we think it was adapted to our process and clients.

Every questionnaire is customized depending of the client but we try to have a general base. Even if we didn't change the original questionnaire, we adapted it by creating imaginary situations with the ECA. For example, we ask our client to answer questions the ECA ask to the user in the real application. Like that, our client has a real opportunity to understand better how the user can interact with the ECA.

Once we receive the answers of the questionnaire, we organize a brainstorming with different persons of the client company and we encourage them to review their answers even if they are unable to explain why they gave this answer. If we can, we show some pictures or we try to create some specific situations depending of their answers.

This new process helps us to collect new kind of information we start to categorize as (1) general information, (2) intuitions, ideas, (3) elements to avoid or that can be negative for the company, (4) elements that are very positive, (5) innovation and changes the client needs or wants, (6) other parameters that can help us to describe the character.

The result of this new process is very positive. As the client acts a role by being in a realistic situation, he can understand better the mission of the ECA. We notice also that we provide ECA with more details that we could do before as we have more information.

c. Develop a taxonomy of affective systems usability testing

The University of Leeds has been reviewing literature on affective computing systems with a view to proposing a simple taxonomy that can be used as the basis for considering usability evaluations. This will be complementary to other activities in this workpackage insofar as it focuses on studies that provide qualitative/summative assessments. Related to this, a review of primary research on affective computing is in progress that will serve to 'populate' the taxonomy. This work will be made available shortly in the form of HUMAINE project deliverable d9g.

3.4 Element 4: New methods for design and evaluation

In this element we develop new methods for capturing unique aspects of affective interaction that can be used during an iterative design-evaluate-redesign process. Methods that will be investigated include: a sensual method for non-verbal mediation of affective state, a Wizard-of-Oz environment for multimodal emotional interaction, and an extended think-aloud protocol designed to capture emotional interactions.

a. Sensual evaluation instrument

Activities during this period have mainly consisted of demonstrating the Sensual Evaluation Instrument on various occasions to get feedback from researchers and practitioners, and writing up and publishing research results. A demonstration of the SEI was e.g. done at the WP9 workshop held in January 2006, where participants could see and try out the method for themselves. We have also made several copies of the SEI object set that we have shared with researchers interested in trying the method.

Two papers about the SEI have been accepted for publication this year. The first, *The Sensual Evaluation Instrument: Developing an Affective Evaluation Tool* (Isbister et al, 2006) was accepted to CHI06 where it also received a nomination for best paper. The second, *The Sensual Evaluation Instrument: Developing a trans-cultural self-report measure of affect* (Isbister et al, 2007), was accepted to a special issue of International Journal of Human Computer Studies (IJHCS).

KTH has also continued to explore the possible uses of the Sensual Evaluation Instrument in a small study of computer games. Three games were evaluated wrt to the experiences they created in the user with special attention payed to *dramatic* experience of the games. The study showed that the SEI was useful for pinpointing troublesome episodes in the game experience but also for gaining an overall impression of the emotional tone of the gaming experience. Results from the study will be published during 2007.

b. WoZ – test bed

BARI continued their work on evaluation methods applied to human-ECA interaction. Various aspects were investigated:

- With the intent of going deeper into the Stanford's 'CASA' hypothesis, we reviewed the psychological literature about 'social emotions' and related concepts: interpersonal stance, social attitude, engagement etc. We extracted, from this review, a list of 'signs' of social attitude that might be employed in observing and interpreting users' behavior when interacting with the ECA, and recognizing their attitude towards the agent.
- With the intent of evaluating whether and how the interaction mode (text-based vs speech-based) influences the user attitude towards ECA, we implemented a new version of our Wizard of Oz tool, in which subjects interact with speech and a touch-screen.
- We performed a new set of evaluation studies with this new version of the tool. These studies had various intents:
 - to extend our previous comparison of 'subjective' evaluation with 'behavior observation' , in order to assess limits and advantages of the two methods;

subjective evaluation was performed with a 'sensual' evaluation of the individual ECA's moves and a final questionnaire based evaluation of the whole dialog, while 'behavior observation' included a refined analysis of the linguistic characteristics of every individual dialog move

- to assess how some individual features like 'level of involvement' and 'degree of initiative' in the dialog may be estimated from analysis of dialog features;
- to assess the relationship between 'stable' user characteristics (gender, background, personality traits) and their social attitude towards the ECA.

While the data collection step has been concluded (60 subjects and about 1500 dialog moves overall), data analysis is still ongoing and will be concluded in the next few months.

c. Extended think aloud protocol

IMPERIAL and GERG have continued to work on a variation of the think aloud method. Thinking aloud is a method traditionally used in usability testing as a way of finding out the 'why' behind a particular problem. While using an interface, a user verbalises his/her cognitive activity by saying what he or she is doing. Boren and Ramey (2000) have described the importance of clearly defining the roles between the facilitator who is also in the room during usability testing and the user who is interacting with the technology. The user is the speaker, giving information about the interface and the facilitator is there to listen and learn while partaking little in the verbal protocol.

For affective evaluations, we wanted to know if this kind of setting elicits more information than just a cognitive reflection of a user's interface activities. Users are given the role of an expert who speaks out while using an application that can provoke an emotional experience. Does this setting provide a platform for expression? And if so, how reliable are these expressions? People often regulate their expressions to others. One can imagine that a user encountering a severe usability problem may suppress his or her expression in consideration of the facilitator who is present. Additionally, a positive expression may be ironic signifying a rather negative experience. And finally, using the think aloud in the lab, can we, the usability experts, get data that goes beyond performance problems with the interface and taps into the user's experience? Recent work concerned with user experience considers the lab to be an artificial setting which by virtue of its format should be considered more appropriate for performance usability. However, the lab is sometimes our only solution. Thus can it be used for evaluations that go beyond its designated performance role?

We set out to answer all of the questions above in a study conducted at the Geneva Emotion Research Group in July of 2006. 20 participants between the ages of 20 and 40 took part in our study. The application tested was Yahoo! Avatars. The site provides users with tools to customise their personalised avatar and also with tools that broadcast their persona to their social networks. The customisation of Yahoo! Avatars was chosen for this study as it is web based and thus lends itself to being studied in the lab. Upon arriving users were trained on how to do the think aloud. They were then given two tasks to complete. In the first task, they rehearsed an upcoming date they were looking forward to. In the second task, they used the application to send their mother a hypothetical gift i.e. a week long holiday in the Greek islands to celebrate her birthday. Although we are currently analysing the data, we mention some of the preliminary conclusions we have drawn that address our three main questions.

Are users expressive while thinking aloud? Users appear to be expressive while doing the think aloud. At the moment we are undergoing a rating study in an attempt to ‘objectively’ verify whether these non-verbal expressions are perceivable by raters who do not have access to the context.

How reliable are these expressions? When should we look beyond the surface and investigate how the context modulates users’ expressions? Although in many instances, the user’s action corresponded to his or her expression, we found that expression as such has to be approached with some caution. Users were ironic when expressing dislike. They joked about options they found ridiculous and usability problems were at times fun.

Can we retrieve user experience data with this method given its artificial setting? Our experience with this method leads us to believe that usability experts can expect to obtain more than performance information. A few examples of such findings are listed here:

- Users used their own physical appearance as a guide for choosing their face. Others chose faces on the basis of their mood at the time.
- Users customised their avatars to send intimate messages; they had fun imagining others reactions when seeing their avatar.
- Users drew on personal values to make their customisation choices e.g. a place I would visit in real life; rejecting places they dislike in their offline life.

d. Psychometric assessment model/approach

The University of Leeds has continued working on the psychometric assessment of user responses to interface designs. A paper is currently under review in which a study is described that assessed different web page designs using a 53 item questionnaire designed to assess multiple components of user response. Results indicated two broad patterns of user response. One related to evaluations of design quality and usability that seemed to arise from elements such as design regularity. The other related to user experiences of playfulness/tenderness and seemed to be generated by flamboyant and novel designs. However, some of the components (factors) that were proposed were poorly represented by items from the questionnaire (e.g., too few item loadings). Therefore, a revised and extended set of 95 items was constructed and a further experiment conducted. In this study psychophysiological (facial EMG and GSR) and eye movement data were also gathered in the hope that these would help to validate user response factors. From preliminary analyses, the factor structure in this second study seems less defined. However, effects of an interface manipulation (colour versus black and white presentation) suggest some similarities between ‘decorative’, ‘individual’, and ‘entertaining’ factors, and this may point to a broad construct similar to one identified in the first experiment. GSR was elevated when websites were presented in colour. Further analyses of this data are planned followed by dissemination of results. In addition to this, the Leeds group have undertaken a review of the use of psychometric assessments of emotion-related computing systems and, particularly, usability assessments. This will be completed shortly and made available in the form of HUMAINE project deliverable d9f.

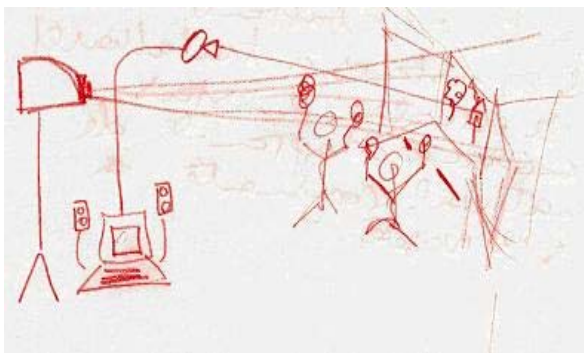
3.5 Affective Interface Development Use Case

Work at IST has focused on exploring affective interaction design in the context of the IShadows project. As the work spans across several of the elements of the exemplar we present it as a whole.

3.5.1 Introduction

Around 200 B.C. in late China, Wu Emperor of Han was taken upon a great sadness after his wife Lady Li died. His ministers were very concerned, because the emperor was not filling his duties. A known magician of that time offered himself to make Lady Li reappear from the death using a Chinese Shadow trick every night. The trick worked for a while. The emperor engaged emotionally with the shadow and reengaged to his responsibilities. But when he found out the truth he went to the other side of the screen...

The I-Shadows system is inspired by this story where we somehow can find the first historical report of an ECA(Embodied Conversational Agent). Using this as an inspiration in Humaine we have created a new Chinese Shadow kind of play, where interactivity emerges and emotional characters (shadows) create stories following an Interactive Drama paradigm. In I-Shadows users are no longer only spectators, but also participate actively in the story development using their own puppets. The system participates also actively, by monitoring the scene using a video camera, and participating by projecting the image of autonomous emotional characters (our shadows). Users and system collaborate in the improvisation of a new story.



This new concept raises the challenge of creating a common expressive language between the computer system and the users in the context of a Chinese Shadow Theatre. This common language should allow users to understand the emotions expressed by the virtual characters, as well for the virtual characters to understand the emotions expressed by the users.

This case study presents the undergoing work in IShadows and its development in the building of this common language. It starts by reporting the User-centred methods used in the design of I-shadows, and the innovative methods tested. Then it explains the criteria for usable affective interactive systems that are being discussed, finishing with some conclusions about the work done

3.5.2 User-centred methods for design and evaluation

The involvement of users in the design process is a time consuming and difficult task. Adding to the fact that our users are children, resulted into special attention needed to keep them focused on the objectives of each experiment.

Our main concern when involving children was to assure that they wouldn't feel being tested and observed but rather authors and participants in the design of a new and different kind of

play. We wanted children to be as natural as possible to get more accurate results. To assure these concerns we defined four rules:

- Children are members of the development team as users.
- All team members’ opinion is important.
- Children collaborate, and are not developers.
- Children and adults have fun, but only adults take notes

To implement these rules, and concerns, we adopted a fast prototyping method which allowed children to participate in several experiments from the beginning of the project until the present day.

Five different kinds of tests were already made until the present day: Acceptance Tests, Observation Tests, Design Workshop, Expression Tests and Expression Recognition. All tests were held at a local school, and are included in the free time activities as a free activity in which they can choose to participate.

3.5.3 Acceptance and Observation Tests

It wouldn’t make any sense to build such a complex installation as I-Shadows if children didn’t like it. Consequently we had to test children’s acceptance of the idea. For that, we built a non-functional prototype of I-Shadows that embedded a simple Chinese Shadows Theatre.



This Prototype was also used with success to see how children expressed, and how they created narratives using it. The results of these tests included a non-quantitative definition of four different patterns of expressions, which corresponded to four emotions, Happiness, Sadness, Anger and Fear. These results can be seen in the following table.

	<i>Direction</i>	<i>Speed</i>	<i>Amplitude</i>	<i>Frequency</i>
Happy	Horizontal	Slow	Short	Low
	Vertical	High	Short	High
	Horizontal	Slow	Very long	High
	Vertical	Slow	Very long	High
Sad	Horizontal	Very slow	??	??
	Vertical	Very slow	??	??
Angry	Horizontal	Very high	Wide	Low

	<i>Direction</i>	<i>Speed</i>	<i>Amplitude</i>	<i>Frequency</i>
	Vertical	Very high	Wide	High
Scared	Horizontal	Very high	Short	High
	Vertical	Very high	Short	High

Table 1: Qualitative Patterns

Other conclusions reached with these tests were that children like the I-Shadows concept, accept suggestions for starting a narrative, as well as they need it to keep their narrative logical enough to be seen by an audience.

3.5.4 Design Workshop

While implementing a first functional prototype with the requisites and patterns taken from the first experiments, we had to support the users’ continuous integration in the development process and make them feel like developers. Due to this, we prepared a Design Workshop, where children were invited to create characters, and sets for their stories.

The goal of this workshop was simple; we wanted IShadows to be as similar as possible to the users’ expectations. We wanted IShadows to present characters and sets according to children’s perspectives.

3.5.5 Expression Tests

This test was taken with eight children (4 boys and 4 girls) aged 8. The goal of this test was to see the accuracy of the proposed expression patterns this time in a quantitative way, with a functional prototype simulated in a computer, controlled by a mouse.

The experiments were preceded by a small introduction where children were invited to play a mime game within the group. “*This mime game has some special words. These words are emotions like Happiness, Sadness, Fury and Fear*”. Once the game started to slow down, we started the experiment that consisted of two tests.

1st test – After choosing two virtual puppets, from the functional prototype, a character to manipulate and a friend to play with, the test worked like the previous mime game, but it was up to their “friend” to recognize their emotion and repeat it. This test was taken in two rounds for each set of emotions (Happy, Sad, Angry and Sad). The results are shown in table 1, the ‘Success’ percentage means the percentage of expressions correctly detected by the prototype.

	Success (1st Round)	Success (2nd Round)
<i>Happy</i>	63%	100%
<i>Sad</i>	88%	100%

<i>Scared</i>	75%	88%
<i>Angry</i>	13%	13%

Table 2 - Tests Results

The children responded enthusiastically to the 1st experiment. Sometimes less expressive kids showed some difficulty at the beginning. However, they adapted to the game very fast by watching how the others did it.

The results were in line with the expected. All patterns except for Angry showed a significant value above the random value. Most of the children showed great difficulty when trying to produce the Anger pattern, because of its high speed, which made them loose control over the mouse.

2nd test – Everyone should try to show their “friend” how he should move when he is angry. The movements’ quantified data was collected and a new Anger pattern was implemented.

3.5.6 Fifth Experiment – Expression Recognition

We were then ready to test the generated expressions of the animated shadows. The experiment involved 10 children (7 boys and 3 girls).

In this test children were invited to guess which spell (emotion) a computer generated shadow was experiencing. The Anger pattern used was the improved version due to the previous test. The results are shown in table 3.

	Success
Happy	70%
Sad	30%
Scared	40%
Angry	50%

Table 3 - Expressions' recognition

Three expressions showed a significant value above a random distribution, which makes us believe in the success of this test. The Sadness expression presented a low success percentage. We interpret this result as an alert that points out the incoherence between the sad expression animation patterns defined by the children.

Afterwards in an informal chat, children mentioned that the lack of facial expressions and sounds of the character made their task more difficult. We also considered by observation that the lack of context in which the expressions occurred might have led to the low results of some expressions.

3.5.7 New methods for design and evaluation

The urge to find new evaluation methods took us to experiment two new methods in the design of I-shadows: Cultural Probes, and Bodystorming.

Cultural Probes

In an attempt to close the gap between the users and the development created between experiments, we proposed to implement a Cultural Probe test. Unfortunately we found an incompatibility between this test and the school rules. The problem was that offering a Probe kit only to those children that participate in the experiment would leave the other children feeling left out, and influence the freedom of choice between the free-time activities. In particular the teachers were quite against this method. Consequently this idea had to be abandoned.

Bodystorming

After defining the emotional patterns, we were ready to start designing the minds of the characters, that will improvise with the users. The need to understand the dimension of this problem, took us to prepare a body-storming session with members of the research group. From this session several important ideas were pointed for this part of the project, such as the need to implement a turn taking mechanism in the cooperative development of a story.

3.5.8 Criteria for usable affective interactive systems

According to Spolin (Spolin, 1963) “Improvisational theatre requires very close group relationships because it is from group agreement and group playing that material evolves from scenes to plays”, in IShadows the success of the Narrative Cooperation will depend entirely on the success of the affective communication between the user and the system. Based on this assumption we proposed to evaluate the success of the of the affective interaction, using the evaluation criteria of the Interactive Drama.

Following Murray (Murray, 1998) there can be defined three aesthetic categories to analyses a players experience in an interactive narrative:

- Immersion
 - When the user totally accepts the logic of the environment.
- Agency
 - Is all about the environment’s ability of letting the user take any action that he wants, and react to it in coherent way.
- Transformation
 - Masquerade – the experience allows the player to transform into another character while he is on the play.
 - Variety – several experiences have several different stories to tell

3.5.9 Conclusions

At this point of the project we feel that the obtained results are very satisfying as the guidelines for the design of an affective system were somehow followed and influenced the resulting installation. The process of evaluating movements from a qualitative perspective

before the quantitative, enabled us to detect three users' expressions with success, and to express other three in an intelligible way, at the first try. The patterns are improving, and are already integrated with the new vision component.

Working with children is being very rewarding, their excitement when we visit them to make some more testing (plays), convinces us that their expectations are being satisfied, and that consequently our development method is being well succeeded, neither less the cultural probes experiment tells us that not all new innovative methods are suited for the context of this project.

The interactive-drama framework that is still under development (and is expected to be fully functional by the end of the project) is expected to offer us some good references for the criteria that will be used to evaluate the effectiveness of our affective system

4 Planned program of research

4.1 Element 1: Criteria for usable affective interaction systems

OFAI will complete the survey of scenarios applicable to existing affective systems in collaboration with WP7. A main goal of this work is the identification of systematic relations between application scenarios for affective agent architectures and their possible designs.

KTH will continue to explore success criteria for affective interaction wrt to specific design aims. We will do this in the context of developing a dramatic affective game which is being developed (see deliverable d9c for a short description of the game). We want to evaluate how different game components contribute to the overall game experience in general and a dramatic experience in particular. This evaluation studies performed will build on the study of three games that KTH performed during the year mentioned in 3.4.a.

4.2 Element 2: Evaluation metrics for criteria

OFAI will continue their work on quantitative and qualitative dialog analysis and evaluation methods and application to the mixed HHI and HMI corpus. The goal is a context-sensitive collection of discourse markers for determining relationships in dialog (e.g. dominance).

The development of the dramatic game (see 4.1) will also involve looking at suitable metrics for operationalising the criteria important for the scenario. For instance, what is a suitable metric for dramatic tension? While it is fairly easy to plot a curve of dramatic tension (indeed this is what many drama systems use as guidance see e.g. Mateas & Stern, 2003) it is less clear what that actually tells us about the users experience. Hence, further investigations are needed.

4.3 Element 3: Existing user - centred methods for design and evaluation

As previously stated in 3.3.c the University of Leeds continues to work on a simple taxonomy that can be used as the basis for considering usability evaluations. Related to this, a review of

primary research on affective computing is in progress that will serve to ‘populate’ the taxonomy. This work will be made available shortly in the form of HUMAINE project deliverable d9g.

The result received from Cantoches new process for designing ECAs seems to be more positive and we want to continue to apply this process to our new clients. We want to focus our work on 2 aspects:

- Finalisation of a general process and methods for the creation of an ECA. We want to analyse the different cases we had this year and to write a general methodology we can apply for the rest of our clients.
- Evaluation of the ECA by collecting user information directly on the application in order to improve the methodology.

4.4 Element 4: New methods for design and evaluation

a. Sensual evaluation instrument

During the last year of HUMAINE our goal is to use the SEI in a real systems development process to evaluate its usefulness as a tool for designers to gain early feedback on the affective qualities of their designs. As of yet we do not have such a project secured but we are having discussions with several parties.

b. WoZ test bed

At OFAI a project in which the test bed will be fitted in will be brought to a testing stage in mid-2007.

c. Extended think aloud protocol

In July 2006 IMPERIAL/GERG conducted an experiment on the think aloud protocol. We started with 3 research questions which we are now delineating in further work. Below we give a description of each question and our future plans in pursuing it.

1. **Are users expressive while thinking aloud?** At the moment we are undergoing a rating study in an attempt to ‘objectively’ verify the non-verbal expressions shown by users of the experiment. The context will be removed by having 12 non-French speaking participants to rate the expressivity of the 320 usability and experience incidents collected during the study. Participants will rate the arousal of the expression shown in the clips from negative to positive. The question we aim to answer with this rating study is whether users are perceived being expressive in the lab and if their expressions are reliable. Mapping this question back to usability testing, can usability experts reliably perceive users’ non-verbal expressions?
2. **How reliable are users’ expressions in the lab?** Although in many instances, the user’s action corresponded to his or her expression, we found that expression as such has to be approached with some caution. Users were ironic when expressing dislike. They joked about options they found in bad taste and usability problems were at times fun. After the rating study is complete, we are

planning to locate incidents which we believe were rated inappropriately. A rater who is unaware of the context may perceive a user who is laughing while criticising the interface to be having a positive experience. By putting emphasis on these kinds of incidents, we should better understand what kind of scenarios demand further attention from us by looking deeper into the context of the interaction.

3. **Can we retrieve user experience data with this method given its artificial setting?** Our experience with this method leads us to believe that usability experts can expect to obtain more than performance information. We are now undergoing a qualitative analysis in which we are grouping the user experience incidents found in the experiment into relevant semantic categories. This process will inform us on the kinds of strategies people draw from, e.g. cultural values, when projecting themselves onto an avatar.

d. Psychometric assessment model/approach

As stated in 3.4.d the Leeds group has undertaken a review of the use of psychometric assessments of emotion-related computing systems and, particularly, usability assessments. This will be completed shortly and made available in the form of HUMAINE project deliverable d9f.

5 References

- Boren, M.T. & Ramey, J. (2000). Thinking Aloud: Reconciling Theory and Practice. *IEEE Trans. Prof. Comm.*, 43, 261-278.
- Cooper, A. (1999). *The Inmates Are Running the Asylum: Why High-Tech Products Drive Us Crazy and How to Re-store the Sanity*, SAMS, 1999.
- DeBono, E. (1985) *Six Thinking Hats*. Little, Brown & Company, Boston, MA
- Djajadiningrat, J. P., Gaver, W. W., and Fres, J. W. (2000) Interaction relabelling and extreme characters: methods for exploring aesthetic interactions, *Proceedings of the conference on Designing interactive systems: processes, practices, methods, and techniques*, Pages: 66 – 71, 2000, ACM Press New York, NY, USA
- Fischer, Kerstin. 2006. The Role of Users' Preconceptions in Talking to Computers and Robots. In *How People Talk to Computers, Robots, and Other Artificial Communication Partners*, edited by K. Fischer. SFB/TR 8 Spatial Cognition. pp. 112-130.
- Gaver, B., Dunne, T., and Pacenti, E. (1999) Cultural probes, In *Interactions*, Volume 6, Issue 1, pp. 21-29, ACM Press, New York, USA.
- Isbister, K., Höök, K., Sharp, M., Laaksolahti, J. (2006): The sensual evaluation instrument: developing an affective evaluation tool, In *proceedings of CHI06*, pp. 1163-1172.
- Isbister, K., Höök, K., Laaksolahti, J., Sharp, M. (2007): The sensual evaluation instrument: developing a trans-cultural self-report measure of affect, Accepted to a special issue of *International Journal of Human Computer Studies* on “Evaluation Evaluating Affective Interactions: Innovative Approaches and Future Directions” (forthcoming).
- Kopp, Stefan. 2006. How People Talk to a Virtual Human. Conversations from a Real-World Application. In *How People Talk to Computers, Robots, and Other Artificial Communication Partners*, edited by K. Fischer. SFB/TR 8 Spatial Cognition. pp. 101-111.
- Mateas, M., Stern, A., (2003). *Facade: An Experiment in Building a Fully-Realized Interactive Drama*, Game Developers Conference, Game Design track, March 2003
- Morel B., (2004), *Agent Culture: Recruiting a Virtual Employee: Adaptive and Personalized Agents in Corporate Communication*, Lawrence Erlbaum
- Murray, J. (1998), *Hamlet on the Holodeck: The Future of Narrative in Cyberspace*, MIT Press
- Oulasvirta, A., Kurvinen, E., and Kankainen, T. (2003) Understanding contexts by being there: case studies in bodystorming Personal and Ubiquitous Computing, Volume 7 , Issue 2 (July 2003) Pages: 125 – 134, Springer-Verlag, London, UK.
- Payr, Sabine. 2006. Seriously Socially Situated Agents. In *Proceedings EMCSR 2006*, edited by R. Trapp. Vienna: OGKS.
- Porzel, Robert. 2006. How Computers (Should) Talk to Humans. In *How People Talk to Computers, Robots, and Other Artificial Communication Partners*, edited by K. Fischer. SFB/TR 8 Spatial Cognition. pp. 7-37.

Preece, J., Rogers, Y. Sharp, H., Benyon, D., Holland, S. and Carey, T. (2002) Human-Computer Interaction, Addison Wesley

Rank S., Petta P. (2006a): Comparability is Key to Assess Affective Architectures, in Trapp R. (ed.), Cybernetics and Systems 2006, Austrian Society for Cybernetic Studies, Vienna, pp.643-648.

Rank S., Petta P. (2006b): Clarifying criteria: comparability and evaluation, WP9 Session (evaluation) at the Humaine WP10 Workshop, <http://emotion-research.net/ws/wp10/presentation-materials/StefanRank-PaoloPetta-wp10ws-comparability-and-evaluation-final.pdf>

Rettig, M. (1994) Prototyping for tiny fingers, Communications of the ACM, volume 37, issue 4, pp. 21-27.

Rienks, Rutger, and Dirk Heylen. 2006. Dominance Detection in Meetings Using Easily Obtainable Features. In Selected Papers MLMI 2005. Heidelberg: Springer.

Sengers, P., Boehner, K., Gay, G., Kaye, J., Mateas, M., Gaver, B., and Höök, K Experience as Interpretation, In Proceedings of CHI 2004 workshop on “Cross Dressing and Border Crossing”, organised by Ron Wakkary, Thecla Schiphorst, Jim Budd, 2004.

Spolin, V. (1963), Improvisation for the Theater, Third Edition, Northwestern Univesity Press: 3 – 47

Suchman, L. (2002). Replicants and Irreductions: Affective encounters at the interface, published by the Centre for Science Studies, Lancaster University, Lancaster LA1 4YN, UK, at <http://www.comp.lancs.ac.uk/sociology/papers/Suchman-Replicants-and-Irreductions.pdf>

Sundström, P., Ståhl, A., Höök, K., (2007): In situ informants exploring an emotional mobile messaging system in their everyday practice, Accepted to a special issue of International Journal of Human Computer Studies on “Evaluation Evaluating Affective Interactions: Innovative Approaches and Future Directions” (forthcoming).

Wallis, Peter. 2005. Believable Conversational Agents: Introducing the Intention Map. In Proceedings of WS "Creating Bonds with Humanoids" at AAMAS'05, edited by C. Pélauchaud, E. André, S. Kopp and Z. Ruttkay. University of Utrecht. Available from <http://www.dcs.shef.ac.uk/~peter/Wallis05-3.html> (09 May 06).