4th International Workshop on Perceptual Quality of Systems (PQS 2013)
2-4 September 2013, Vienna, Austria

User Experience in Human-Robot Interactions

Kristiina Jokinen [1,2] *Graham Wilcock* [2]

[1]University of Tartu, Tartu, Estonia
[2]University of Helsinki, Helsinki, Finland
kjokinen@ut.ee, graham.wilcock@helsinki.fi

Abstract

This paper describes our experiments concerning human-robot interaction and its evaluation in order to measure the user's communication experience with the robot. We assume that the user's evaluation of the interaction reflects their participation in the interaction as a participant, and correlates with their own multimodal signaling as well as with their perception of the robot's communicative capability. The work contributes to evaluation methodology of intelligent situated agents, as we explore the role and effect of the user's own activity in successful communication experience, and ultimately in their evaluation of interactive systems, instead of focusing solely on task completion. This also adds to our understanding of how the interlocutors perceive interaction, and of the important cognitive processes that underlie communication experience in general.
Index Terms: multimodal interaction, evaluation, user expectation, experience

1. Introduction

Present day technologies enable fairly robust human-machine interactions, and various spoken dialogue systems, human-robot interactions, and related real-time communication scenarios support growing interest in high-tech communication systems in general. Natural language interaction is an important aspect of these applications, and it is assumed that this directly contributes to the usability of the system. Besides speech and text-based interfaces, multimodal communication is also considered useful with situated agents and intelligent environments, since many of the emotional and social signals which make the communication fluent and effortless, are conveyed non-verbally.

However, naturalness of the communication between human users and situated agents suffers from the same problems that are common to other communication systems and applications: difficulties in distinguishing important pieces of information, inappropriate interaction strategies, bad connections, etc. These issues make interaction rigid and less fluent, and the technology is often of insufficient quality in order to deliver communication that is required in social contexts: interaction is not only about exchanging information but also about building shared context and mutual understanding. The user's experience in these situations may also be difficult to measure, since people seldom spontaneously describe such situations as natural, and when they do, it may be by comparison with traditional input/output interfaces rather than with their experience of the robot as such.

When interaction technologies are developed, the situational context is also important: the user's communication experience varies in different situations. Good user experience is affected by basic usability and quality issues such as ease of use, transparency of interface, technical robustness, etc. It is likely that the same multimodal activity which seems to have a positive correlation with the participant's experience in human-human interactions ([11], [14]), also positively affects the user's experience when this interacts with situated agents.

There are many proposals to achieve a good definition of user experience. In this paper, the focus is on the user's communication experience with the WikiTalk implementation on a humanoid Nao robot ([22], [4]). The work explores experiential aspects of the communication experience via a questionnaire and the user's observed behaviour in the different situations. The purpose of this empirical research was to combine social-psychological and technological elements in the assessment of a human-robot interactive system, and to provide a wide and holistic picture of the communication experience phenomena. A particular goal was to explore the hypothesis that multimodal activity predicts user satisfaction in human-robot interactions.

The paper contributes to the evaluation methodology of intelligent situated agents, by emphasizing the user's own role and activity in the interaction with the agent. Instead of focussing solely on task completion and efficient communication of factual information, the paper starts from the user's experience of the progress of communication with the robotic partner, and checks how this is related to the user's evaluation of the system. It is assumed that the user's participation in the interaction is reflected in their evaluation of the interaction, i.e. the more actively engaged in the interaction the user is, the more positive experience they have of the interaction. Engagement is related to the user's own multimodal signaling and to their perception of the signals emitted by the intelligent agent, correlating to their communicative capability. Consequently, the partner is valued and perceived as having qualities, such as helpfulness and pleasantness, which are highly rated in the evaluation scores.

The paper is structured as follows. Section 2 surveys recent work on system evaluation related to user experience and gives an overview of relevant research concerning user experience and interaction. Section 3 describes Constructive Conversations and relates them to the open-domain dialogue system WikiTalk. The evaluation data and the annotations are described in Section 0. Section 5 presents the results concerning expectations and experience as well as experimentation with evaluation classification concerning multimodal features. Finally, Section 6 provides discussion and interpretation of the results, and draws conclusions and describes plans for future work.

2. User experience and Interaction

The ISO standard ([5]) focuses on the cognitive concepts of a person's perceptions and responses which result from the use or anticipated use of a product, system, or service. In this paper, we follow this definition and use the expectation-experience methodology as outlined in [12]. The user's expectations about

10.21437/PQS.2013-24

the interaction with the intelligent agent are related with the user's experience with the agent and the experience is assumed to be more positive the more these expectations are met. We further assume that in human-robot interactions, the expectations concern the communicative capability of the robot, and if the user perceives the interaction as smooth, easy, and engaging, the system is also perceived as engaging and easy to interact with, and, consequently, it receives positive evaluation. If it is possible to measure the user's engagement in the interactive situations, it may also be possible to measure their experience to the extent that it is possible to estimate their likely future use of the system.

The main question in evaluating use experience is what are the relevant concepts which best represent the psychological processes and enable measurements of the true characteristics of user experience. Concepts such as involvement ([24]), presence ([19]), immersion ([7]) and flow ([5]) have been used to describe experience and creative state of mind especially in digital game playing. In the PIFF framework [21], the concepts of presence, involvement and flow are integrated, and the five experiential dimensions of the communication experience are specified as emotional involvement (being immersed and captivated by the discussion, enjoying and impressing), active participation (related to interactivity, having impact and control of the discussion, being aroused and cognitively spontaneous in the situation, perceived emotional interdependence), reciprocity (perceived behavioural interdependence, attention allocated to the other participants, affective understanding), co-presence (naturalness and smoothness of the discussion, sharing the space with partners), and group cohesion (perceived message understanding, trust, attraction and commitment to the in-group). The latent variables of physical presence, involvement, social presence, cognitive evaluation, and emotional outcomes are measured using questionnaire items (observed variables) and from these, subcomponents of user experience in games are extracted via factor analysis.

From the view-point of communicative competence, the user's experience is linked to natural interaction and cooperation so that the system is perceived as having an ability to communicate in a successful and intuitive manner. In interaction research, the participants' cooperation seems to correlate with their synchronous communicative activity, i.e. harmonious and coordinated behaviour by the interlocutors' seems to coincide with the speaker's positive experience of the communicative situation in general. Concepts such as alignment ([18]), engagement ([10]), and entrainment ([15]) have been used to study this kind of synchrony and cooperation. In our experiments we used five evaluation categories related to the communicative aspects of the interaction. They measure the robot partner's perceived expressiveness and responsiveness, technical aspects of the interface, usability, and overall view of the system, and are related to multimodal behaviours which are independently annotated in the video data (see Table 2 in Section 4).

It is important to emphasize that naturalness in interaction technology does not necessarily mean similarity to human-human interaction but rather, it can be described by the notion of *affordance*. In the context of natural language communicating systems, affordance was introduced by [8], following [17], and it refers to the intuitive and natural design of the system interface, so that the system's communication is readily comprehensible to the user. The participants should be able to perceive the conveyed information as it is intended to be understood in the interaction, and it is crucial that emotional and social signals are also perceived as part of the social interaction and addressed in a way that contributes to fluent and effortless communication experience. The available communication channels and methods, although advanced in their present state, are not sufficient for maintaining fluent verbal and non-verbal conversation, especially in situations where the partners could freely talk about interesting open-domain topics and accompany their presentation with appropriate gesticulation and emotional signals. Instead, affordance relates natural interaction to intuitive communication principles which invite the user to make appropriate and intended use of the system: how to enable communication goals to be met, and how to deal with the affective aspects of conversation. The "naturalness" of the interaction can be understood in a creative manner across the novel channels and interaction modes, by taking into account new behaviors that emerge in order to compensate the limitations of the channels with respect to full human-human like communication.

3. Constructive Conversations

Social and sociable robots refer to the robots which can communicate with humans in a socially correct way, in order to perform a task ([2], [3]). Their communicative repertoire can be fairly rich due to various multimodal technologies that enable them to recognize the user's spoken and multimodal utterances and also produce their own output using speech, gestures and other modalities. Consequently, interactions with social robots resemble interactions between humans, and we can assume that affordable interfaces to social robots effectively deploy multimodal technology as well as human natural communication strategies. It must be emphasized however, that HRI applications are complex integrated systems, and their evaluation results are not solely determined by the interaction design and the user interface: performance failures may be due to the robot's software system or malfunctioning of the robot's sensory system, rather than an issue with the user interaction (cf. [20]). On the other hand, such malfunctioning need not even be noticed by the human partner as a problem and dialogue strategies may enable conversation to continue, so that malfunctioning may, in fact, be perceived as a sign of the robot's creative and autonomous behavior that makes it an "intelligent" conversational partner.

If we regard face-to-face situations in a holistic manner, we can mimic interactions by utilizing our knowledge of the significant social interaction rules and multimodal cues. We can integrate psychological, social, and technological elements that affect subjective experience, and provide a wide and holistic picture of the phenomena of communication experience. In the optimal case, technology can enhance social interaction, and help the partners to achieve mutual targets. For instance, in human-robot communication, the participants' (multimodal) reactions can be recognized as pertinent signals that contribute to mutual understanding, and although they will add to the complexity of the situation, they also help to convey information that is required to achieve communication goals.

As a starting point for metrics for natural interaction, we have used the Constructive Dialogue Model (CDM, [8]). It uses the basic enablements for communication, Contact, Perception, Understanding, and Reaction, and models them in the computer agent's behaviour in human-computer interaction. The agent's awareness of the partner is mediated through the enablements of Contact and Perception: the agent needs to be within a space where interaction with the partner is possible (speech audible,

gestures visible, group configuration supporting interaction, etc.) and the agent also needs to perceive the partner communicating or intending to communicate – if these enablements are not fulfilled, the agent must signal to the partner that communication is not possible unless the enablements are fulfilled. The level of "awareness" depends on the level of the agent's autonomous behaviour and its role in the situation, but it is important that the agent is aware of the social space in which the interaction takes place, and can also signal its understanding to the partner, e.g. by appropriate distance. The enablements of Understanding and Reaction concern the agent's engagement in the interaction. The agent processes the partner's communicative signals and creates a meaning of the partner's contribution in the context. Cognitive processing also includes deliberation and generation of one's own reaction in order to provide feedback to the partner and to advance one's own communicative goals.

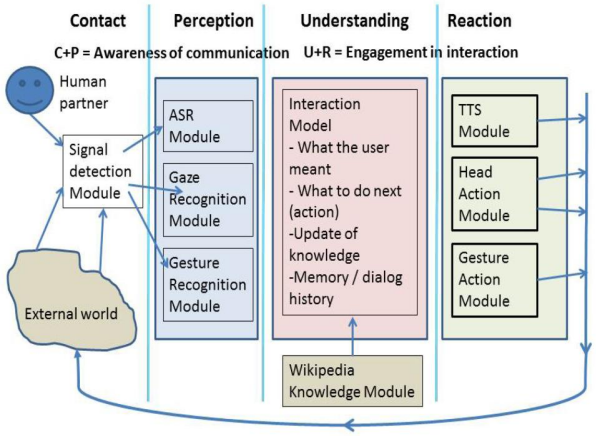

Figure 1 Basic enablements of communication (Contact, Perception, Understanding, Reaction) in relation to the general architecture of an interactive system. From [13].

In applying the basic enablements of communication to human-computer interactions, the architecture should reflect the basic enablements so as to provide opportunities for signalling possible problems and misunderstandings to the partner, and to apply strategies that help the agent to maintain the enablements through interactive situations. Figure 1 depicts a general system architecture and how the communicative enablements are applied to the system modules (from [13]). The Contact-Perception awareness can be interpreted as an emergent property of the agent's signal detection and signal processing modules, while Understanding is related to the agent's reasoning about the input given the current context, and Reaction to the execution of the communicative goal (possibly combining physical acts as well). In the current implementation of WikiTalk, Understanding is a state-based module which coordinates conversations and topic management with respect to the Wikipedia articles. The Reaction components refer to the robot's speech and motor control engines through which the communicative actions are executed.

4. Data

We use the *eNTERFACE-2012 Nao-Human Interaction Data* which contains videos of human-robot spoken interactions using the Nao WikiTalk system ([22]). The robot uses Wikipedia as its knowledge base, and this allows it to hold open domain conversations on any desired topics (see [22], [4], [13], [16]for more details about WikiTalk and its implementation on the Nao robot). To construct expressive presentations of the requested information, the robot uses various multi-modal signals to help the user to follow the discourse: e.g. nodding to give feedback to the user, hand gesturing to direct the user's attention to a discourse topic and New Information, and body posture to distinguish speaking posture from the listening posture.

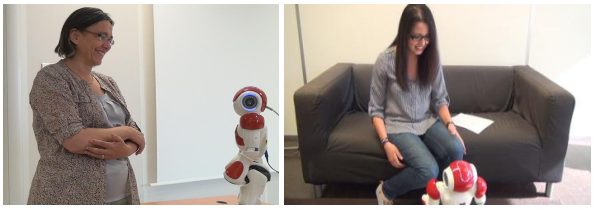

Figure 2 Users interacting with the Nao robot.

The data consists of human-robot interactions as shown in Figure 2. The users interacted with three slightly different versions of the Nao WikiTalk: version 1 implements only face following but no gestures; version 2 has head nods as well as hand gesturing, while version 3 used only various types of hand gestures. Altogether twelve users (7 men, 5 women) interacted with the robot, resulting in a corpus of 12 x 3 = 36 interactions. The users, aged 20-40, were given short instructions of how to interact with the Nao robot, and their task was just to explore its capability to converse about Wikipedia articles. The users had no previous experience of the robot, although they knew about it and were familiar with multimodal interaction technology.

Table 1 Annotation categories (n=6), distinct feature values (n=46), feature values selected for the experiments (n=16), and sample features.

Annotation category	Distinct features	Used features	Example features
Gaze (focus of attention)	3	3	**toRobot, toInstructor, toBackground**
Perceived emotional expression	17	6	**amused, disappointed, interested, sad, satisfied, uncertain**, …
Body engagement	7	1	**leansTowardsRobot**, leansAwayRobot, …
Hand gestures	8	0	singlehandWaving, bothhandWaving, beatGesture, …
Facial expression	5	2	**smile, laugh**, frown, …
Robot's perceived appropriateness	6	4	**ok, offThePoint, odd, okByChance**, …

The corpus was manually annotated concerning multimodal categories which were considered important to describe the user's conversational behavior and their multimodal signals in particular, related to their comprehension of the message and displaying their affective state (Table 1). Of all the possible annotation features we selected for the experiments only those which occurred more than 9 times in the corpus (i.e. 9/36 = 1/4 of the interactions). The column "used features" in Table 1 refers to the number of these pruned features, and the boldfaced features in the examples mark the selected ones. It is interesting that none of the user's hand gesture features are frequent enough, and only one type of body movement was annotated frequently enough. This may be due to the setup in which the users listen to the robot rather than present long contributions themselves.

For the actual evaluation, we followed the methodology outlined in [12]. This compares user expectations with the user's actual experience of the system. The users were first given a questionnaire that recorded their expectations, and after each interactive session they were given a similar questionnaire, but this time to record their experience with a particular system version. The questionnaire was linguistically adapted to suit to the different evaluations: future expectations used future tense (e.g. *I expect to find Nao's hand and body movements creating curiosity in me)* while the experience statements used past tense (e.g. *Nao's hand and body movements created curiosity in me).* We considered possible priming effect of the expectation questionnaire neglible, since the focus was on how the user experienced a category rather than whether they noticed it. We used five evaluation categories (Table 2), and each category consists of 5-8 evaluative statements. The users were asked to score the statements using a 5-point Likert scale (1 = total disagreement, 5 = total agreement with the statement).

Table 2 Evaluation categories (for user experience)

Expressiveness	the robot's lively behavior and clear and expressive presentation
Responsiveness	the robot's quick and appropriate reaction
Interface	technical aspects of the speech interface
Usability	the robot's performance with respect to the open-domain conversation task
Overall	general aspects of the interaction design

5. Evaluation results and experiments

The average expectations and experiences for the five evaluation categories are shown in Figure 3. The highest expectations concern *Usability*, and the lowest *Expressiveness*, in accordance with the assumption that the users expect interactions to be useful and functional, but not very natural.

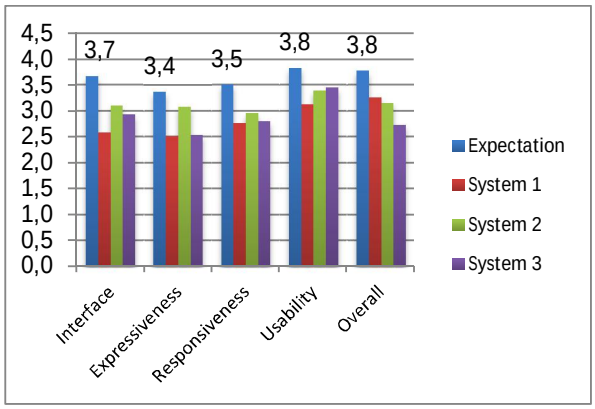

Figure 3 Average expectations in all evaluation categories (5-point Likert scale). The mean values of user expectations are shown.

Table 3 shows how the expectation values correlate with the experience values. The version 3 has highest correlation with the Overall expectations, the version 2 with Usability expectations, and version 1 with Expressive expectations. This fits with the intuitive views of the different versions: version 3 with gestures only addresses the user's expectations of the general aspects of interactive systems, version 2 with nodding and gesturing fulfils usability expectations, and the basic version 1 with only face following relates to expectations of the system expressiveness.

Table 3 Correlations along the five different evaluation categories between expectations (E) and the three system versions (s1-s3), as well as between the different evaluation systems (s3, s2, s1).

	Interface	Expressive	Responsive	Usability	Overall
E-3	0,305	0,129	-0,362	0,139	**0,414**
E-2	0,442	0,212	-0,057	**0,692**	0,351
E-1	-0,122	**0,331**	0,284	0,306	0,135
s3-s2	0,460	0,430	0,265	**0,554**	**0,560**
s3-s1	-0,020	0,359	**0,615**	**0,607**	0,145
s2-s1	0,278	**0,942**	0,159	**0,642**	0,379

We also ran paired t-tests (alpha = 0.05) on each evaluation category to see if there are statistically significant differences between expectations and experience ($p < 0.05$). With respect to the simplest version 1, the user's expectations in all categories are significantly higher than their experience, whereas with the version 3, the users' expectations are significantly higher than their experience for *Expressiveness* and *Responsiveness*, but their expectations for Usability, Interface, and Overall did not differ statistically from their experience. With the version 2, the user's expectations were significantly higher only for *Responsiveness*, while in other categories the user's experience did not differ significantly from their expectations (see more details in [1]).

Our original hypothesis, that the evaluation of the interaction reflects the user's participation in the interaction and correlates with the user's multimodal signaling, was formulated as a classification problem: how well do the annotated multimodal features distinguish the different interaction sessions with respect to user's evaluation. We first created an *interaction instance vector*, which consists of the frequencies of the multimodal features as independent variables, and the evaluation category as the class variable to be predicted. The frequencies are normalized with respect to the length of the interaction (measured in seconds), while the class variable (= the evaluation category) is an average of the evaluative statement scores for the evaluation category in regard to the interaction and the user. We used Weka [23] for the classification, and its instance filter was used to discretize the numeric class variables into nominal attributes (3-bin and 5-bin discrete classes). The experiments were conducted via 10-fold cross-validation.

Table 4 shows the percentage correct and weighted average f-measure results for classifying multimodal feature instances into the five evaluation categories (Interface, Expressiveness, Responsiveness, Usability, Overall), using instance vectors that contain all 16 multimodal annotation features. In 5-bins, the multimodal features best predict the evaluation category Usability. Next come Responsiveness and Interface, if we look at the %-correct values, but if we look at the f-measure, Interface is paired with Overall. In 3-bins, the best predicted categories are Overall and Interface, with more than 50% correct. Since the differences are not statistically significant, we cannot draw firm conclusions, but can nevertheless find some corroboration to our hypothesis that the annotation features have some predictive power in the evaluation. For instance, the best features for the evaluation category Interface describe the user's emotional state and perception of the appropriateness of system behavior, i.e. the user's inner psychological state and the perceived quality of the interaction seem to positively correlate with the evaluation.

Table 4 SVM classification on all 16 annotation features (standard deviation in parenthesis)

Evaluation category	Percent correct		Weighted average F-measure	
	5-bin	3-bin	5-bin	3-bin
Overall	37.50 (22.71)	**52.33 (20.99)**	**0.30 (0.21)**	**0.44 (0.23)**
Usability	**42.92 (21.76)**	42.58 (18.98)	**0.33 (0.22)**	0.33 (0.21)
Expressiveness	39.00 (21.25)	47.92 (19.26)	0.29 (0.20)	0.39 (0.19)
Responsiveness	**40.58 (15.43)**	30.67 (24.27)	0.27 (0.13)	0.27 (0.24)
Interface	**40.42 (20.22)**	**51.00 (18.28)**	**0.30 (0.22)**	**0.40 (0.19)**

We also used Weka's CfsSubsetEval attribute selection algorithm with the best first search strategy (greedy hill-climbing with backtracking) to select a subset of the best predictive annotation features. The algorithm considers the predictive ability of each feature together with the mutual redundancy of the features, and the preferred subsets are those that correlate highly with the class but have low inter-correlation among the selected features. Using 10-fold cross-validation, the best features for each evaluation category are as shown in Table 5.

Table 5 Best features for the evaluation categories.

Evaluation category	Best features selected
Expressiveness (5)	gazedParticipant-atInstructor, emotionalExpression-Amused, emotionalExpression-Disappointed, emotionalExpression-Satisfied, appropriateness-odd
Responsiveness (5)	gazedParticipant-atInstructor, emotionalExpression-Amused, emotionalExpression-Disappointed, emotionalExpression-Sad, emotionalExpression-Satisfied
Interface (6)	emotionalExpression-Amused, emotionalExpression-Disappointed, emotionalExpression-Satisfied, appropriateness-okByChance, appropriateness-offThePoint, appropriateness-odd
Usability (5)	gazedParticipant-atInstructor, emotionalExpression-Amused, emotionalExpression-Sad, emotionalExpression-Satisfied, appropriateness-okByChance
Overall (8)	gazedParticipant-atInstructor, emotionalExpression-Amused, emotionalExpression-Disappointed, emotionalExpression-Sad, appropriateness-OK, appropriateness-odd, facialExpression-smile, facialExpression-laugh

The SVM results with the selected best feature sets for each evaluation category are given in Table 6. The pruning of the features seems to improve Overall, Usability and Expressiveness evaluation, which are all predicted significantly better with the selected features than with the full feature set. The pruning seems to have especially bad effect on Responsiveness, which is understandable since the selected best features are related to the user's observed emotional status while it seems more likely that the system's own behavior plays a crucial role in predicting responsiveness (however, annotation did not include Nao robot's gesturing). Some multimodal features can obviously function as important signals for the evaluation categories, lending support for our original hypothesis.

Table 6 SVM classification on the best feature set for each evaluation category (standard deviation in parentheses)

Evaluation category	Percent correct	Weighted average F-measure
Overall	**49.92 (19.08)**	**0.40 (0.20)**
Usability	**53.08 (22.49)**	**0.44 (0.26)**
Expressiveness	**53.25 (15.84)**	**0.41 (0.19)**
Responsiveness	23.58 (18.84)	0.17 (0.16)
Interface	45.67 (11.75)	0.31 (0.10)

6. Discussion and future work

In this paper, we have discussed user experience in interactive systems and empirically studied the user's multimodal behaviour and its impact on the user's experience of the interaction in spoken human-robot interactive situations. We explored how certain features of multimodal behaviour can predict the user's perception of the interaction and found statistically significant correlations between the users' multimodal activity and their evaluation of the interactive system. The annotated multimodal features give some indication of the success of interaction: the evaluation categories Interface and Usability can be predicted with sufficient accuracy based on the multimodal features, but the categories Expressiveness and Overall only slightly.

The paper contributes to the evaluation methodology of interactive spoken dialogue systems by putting emphasis on the user's own communicative activity and cognitive processing of the interaction. Starting with observations of the user's multimodal activity, the evaluation methodology measures the impact on the user's assessment of the interactions with the system. The user's communicative activity is correlated with the evaluation of the system: it is assumed that the user's experience of the interaction as a participant in the communicative situation (rather than just an observer of the system's behavior) relates to their assessment of the success of the interaction. The user's perception of the communicative capability of the interactive system is input to their cognitive processing, and forms the basis for conscious system evaluation: the recognized communicative signals by the system can support the user's experience of the system being capable of maintaining natural communication.

The assessment of the success of the interaction does not simply lead to the future use of the system, or positive responses imply motivated system use. Experience is actively generated in the repetitive interactive situations, and the basic psychological processes that are related to the user's attention, motivation and cognition, direct the user's behaviour in the evaluation situations, too. They affect the user's perception and operate in the conscious evaluation of the interaction: we perceive certain behavioural signals and selectively focus our attention on the stimuli that motivate and interest us. Conscious evaluation of the system includes only fraction of the perceptions that may carry interesting and meaningful information, as a selection process filters out many signals that the user thinks are irrelevant or noise, and also those which the user may not consciously perceive at all as their attention is focussed on something else.

Evaluation is thus ultimately based on cognitive capability, and in this we come close to the psychological research framework PIFF ([21]) which studies experiences in mediated communication situations, especially in game playing. Although interactions with robots differ from game interactions, we can assume that the same cognitive principles apply to the user's perception of the interaction, whether this takes place in a game or situated agent environment. An important concept is engagement, the partner's involvement, commitment, and presence in the interactive situations. In games, engagement is defined as the game's ability to captivate and enthrall the players so that they will come back to playing the game again, while interactions with social agents tend to interpret engagement with respect to the interlocutor's communicative activity, involvement in the task completion and development of conversational topics. [5] points out, in his studies of the notion of flow in games, that intrinsically rewarding behaviours are experienced as more enjoyable, and they are more likely to be repeated. Concerning the use of practical service applications, this may not depend so much on the user's intrinsic motivation than on their practical need to complete a task, so the repetitive use of the system is dictated by other reasons than intrinsic motivations crucial for game playing. However, the user's attitudes and past experiences have an impact on the experiential process, and thus also on the system evaluation. Recurrence of positive experiences with a service agent or social robot can engage the user and direct their perception of the agent's communicative capability, leading to successful interaction and positive usability evaluation.

Our experimental results show that the users' experiences are versatile in nature, and they seem to point to a promising way ahead in studying user experience in interactive systems. However, more investigations on the theory of perception and cognition are needed, as well as experimental studies on how the various parameters work and influence each other in different communicative settings. For instance, the predictive power of the multimodal annotation features with respect to the evaluation categories needs more accurate studies, and systematic analysis of large multimodal datasets containing versatile activities.

Future work will thus concern experimenting with larger quantities of interaction data and exploring the best sets of multimodal signals for evaluation. It is also useful to refine the set of annotation features and to study different feature combinations, so as to deepen our knowledge of the human perception capability. Moreover, the features used in the current experiments can be automatically extracted from the video and speech signals (using motion trackers it may even be possible to use online data in real situations), and it is expected that in this way evaluation and the user's experience prediction can be automated. Finally, it will be interesting to compare the notion of engagement in game and social robot interactions, and thus further develop general views about human engagement in a wide range of interactive situations.

7. Acknowledgements

Thanks go to the participants of the eNTERFACE 2012 Summer School and to the annotators of the video files.

8. References

[1] Anastasiou, A., Jokinen, K., Wilcock, G. "Evaluation of WikiTalk - user studies of human-robot interaction". Procs of 15th International Conference on Human-Computer Interaction (HCII 2013), 2013.

[2] Breazeal, CL. *Designing Sociable Robots*, MIT Press, 2002.

[3] Breazeal C., Scassellati, B., "A context-dependent attention system for a social robot". Procs of the 16th International Joint Conference in Artificial Intelligence (IJCAI 99), Stockholm, 1146–1151, 1999.

[4] Csapo, A., Gilmartin, E., Grizou, J., Han, J., Meena, R., Anastasiou, D., Jokinen, K., Wilcock, G. "Multimodal conversational interaction with a humanoid robot", Procs of the 3rd IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2012), Kosice, 667-672, 2012.

[5] Csikszentmihalyi, M. *Beyond boredom and anxiety*. Jossey-Bass Publishers, San Francisco, 1975.

[6] ISO DIS 9241-210:2010. Ergonomics of human system interaction. Part 210: Human-centred design for interactive systems. International Standardization Organization (ISO). Switzerland.

[7] Jennett, C., Cox, AL, Cairns, P, Dhoparee, S, Epps, A, Tijs, T, Walton A. "Measuring and defining the experience of immersion in games". *International Journal of Human-Computer Studies* 66:641-661, 2008.

[8] Jokinen, K. *Constructive Dialogue Modelling – Speech Interaction and Rational Agents*. John Wiley & Sons, Chichester, UK, 2009.

[9] Jokinen, K. "Rational Communication and Affordable Natural Language Interaction for Ambient Environments". In Nakamura, S., Geunbae Lee, G., Mariani, J., Minker, W. (Eds.) *The Second Workshop on Spoken Dialogue Systems Technology. Springer Lecture Notes in Computer Science*. Vol. 6392: 163-168, 2010.

[10] Jokinen, K. "Turn taking, Utterance Density, and Gaze Patterns as Cues to Conversational Activity". *Proceedings of the International Conference on Multimodal Interaction (ICMI-2011) Workshop on Multimodal Corpora for Machine Learning*, Alicante, Spain, 2011.

[11] Jokinen, K. "Explorations in the Speakers' Interaction Experience and Self-assessments". *Procs of the 24th International Conference on Computational Linguistics (COLING)*. Mumbai, India, 2012.

[12] Jokinen, K., Hurtig, T. "User expectations and real experience on a multimodal interactive system". *Procs of Interspeech*, Pittsburgh, USA, 2006.

[13] Jokinen, K., Wilcock, G. "Multimodal Open-domain Conversations with the Nao Robot". *Procs of the 4th International Conference on Spoken Dialogue Systems (IWSDS)*. Ermenonville, France, 2012a.

[14] Jokinen, K., Wilcock, G. "Multimodal Signals and Holistic Interaction Structuring". *Procs of the 24th International Conference on Computational Linguistics (COLING*). Mumbai, India, 2012b.

[15] Levitan, R., Gravano, A., Hirschberg, J. "Entrainment in Speech Preceding Backchannels", in *Procs of ACL 2011*, 113-117, 2011.

[16] Meena, R., Jokinen, K., Wilcock, G. "Integration of gestures and speech in human-robot interaction". *Proceedings of the 3rd IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2012)*, Kosice, 673-678, 2012.

[17] Norman, D. A. *The psychology of everyday things*. Basic Books, New York, 1988.

[18] Pickering, M., Garrod, S. "Towards a mechanistic psychology of dialogue", *Behavioral and Brain Sciences* 27: 169-226, 2004

[19] Schubert T, Friedmann F, Regenbrecht H. "The experience of presence: Factor analytic insights. Presence-Teleoper". *Virtual Environment* 10:266-281, 2001.

[20] Steinfeld, A., Fong, T., Kaber, D., Scholtz, J., Schultz, A., Goodrich, M. Common Metrics for Human-Robot Interaction, *Human-Robot Interaction Conference*, 2006.

[21] Takatalo, J., Häkkinen, J., Kaistinen, J., Nyman, G. "Measuring user experience in digital gaming: Theoretical and methodological issues". *Proceedings of the IS&T/SPIE Symposium on Electronic Imaging*, San Jose, California USA.1-13, 2007.

[22] Wilcock, G. "WikiTalk: a spoken Wikipedia-based open-domain knowledge access system", *Procs of the COLING 2012 Workshop on Question Answering for Complex Domains*, India, 57-69, 2012.

[23] Witten, I., Frank, E., Hall, M. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2011.

[24] Zaichkowsky J.L. "Measuring the involvement construct". *Journal of Consumer Research*, 12:341-52, 1985.