The final publication is available at Springer via DOI: 10.1007/s12193-011-0088-y

Measuring the Quality of Service and Quality of Experience of Multimodal Human-Machine Interaction

Abstract

Quality of Service (QoS) and Quality of Experience (QoE) have to be considered when designing, building and maintaining services involving multimodal human-machine interaction. In order to guide the assessment and evaluation of such services, we first develop a taxonomy of the most relevant QoS and QoE aspects which result from multimodal human-machine interactions. It consists of three layers: (1) The quality factors influencing QoS and QoE related to the user, the system, and the context of use; (2) the QoS interaction performance aspects describing user and system behavior and performance; and (3) the QoE aspects related to the quality perception and judgment processes taking place within the user. For each of these layers, we then provide metrics which are able to capture the QoS and QoE aspects in a quantitative way, either via questionnaires or performance measures. The metrics are meant to guide system evaluation and make it more systematic and comparable.

Keywords *Quality Assessment, Multimodal Interfaces, Usability*

The final publication is available at Springer via DOI: 10.1007/s12193-011-0088-y

1 Introduction

Whereas the quality of unimodal interfaces and multimedial interactive systems has been addressed for several decades, the quality of multimodal human-machine interaction is a relatively new topic. The reason is that multimodal dialogue systems have only recently reached a level of maturity which allows for a wide-spread application. Examples include information kiosks at airports or train stations, navigation systems, media guides, entertainment and education systems, or smart home environments [1][2][3][4][5].

In the frame of this paper, we define multimodal dialog systems as computer systems with which human users interact on a turn-by-turn basis, using several different modalities for information input and/or receiving information from the system in different modalities. By interaction modality, we mean the sensory channel used by a communicating agent to convey information to a communication partner, e.g. spoken language, intonation, gaze, hand gestures, body gestures, or facial expressions [1]. These channels may be used sequentially or in parallel, and they may provide complementary or redundant information to the user [7]. It is commonly expected that the use of multiple modalities makes better use of the human cognitive resources, and will thus result in a lower cognitive load on the user during the interaction [8]. In addition, the use of different modalities may provide better recognition and interpretation performance on the system side, in particular in adverse environments where ambient noise and illumination degrade the information input performance. Information provided by the system can better be tailored to the user and the situational context if several output channels are available. Thus, multimodal systems have some principle advantages over comparable unimodal systems.

In order to reach high quality and usability, each system ideally passes several assessment and evaluation cycles during its development: Individual components (such as a speech or gesture recognizer) are assessed as to whether they provide sufficient performance; design concepts are evaluated with respect to their functional requirements as well as with respect to the expected user experience; initial prototypes are tested in terms of mock-ups or through Wizard-of-Oz simulations; preliminary system versions are evaluated in tests with “friendly” users; and roll-off systems are evaluated with their first customers. For IP-based

services, network performance is determined, verified and controlled in regard to providing and processing information for/of the multimodal interface. Such huge efforts should lead to systems and services which provide a high quality to their users; however, the high percentage of unsuccessful innovations in this area shows that the quality of multimodal dialogue systems is still limited, and users might recur to well-established unimodal systems instead. The fact that a substantial number of commercial systems reach their customers without a thorough evaluation is in our experience only partially due to the lack of time and financial effort spent in such evaluations. Rather, we think that a significant part of the problem is due to insufficient evaluation techniques.

Despite several efforts made in the past either for multimodal systems [6][9] or for the general system development process [10], most evaluations are still individual undertakings: Test protocols and metrics are developed on the spot, with a particular system and user group in focus, and with limited budget and time. As a result, we see a multitude of highly interesting – but virtually incomparable – evaluation exercises, which address different aspects of quality, and which rely on different evaluation criteria. Already in 1998, Gray and Salzman [11] reviewed papers investigating usability evaluation methods and concluded that most of these studies are misleading. As a result they recommended to follow a strict experimental approach. Although the paper was widely discussed in the community [12], efforts to standardize evaluation methods were apparently only seldom made as presented in a meta-analysis reviewing 180 studies [13]. In his paper the author criticizes the diversity in measuring user satisfaction. In particular he states that not employing standardized questionnaires leads to severe difficulties in comparing different studies [13]. Also [14] discuss the large variety of different usability evaluation methods and the resulting lack of understanding of each approach.

An evaluation criterion commonly used by system designers is *performance*: To what degree does the system provide the function it has been built for. A collection of such performance criteria can result in *Quality of Service (QoS)*, i.e. “the collective effect of service performance which determines the degree of satisfaction of the user” [15]. QoS can be viewed both in terms of the prerequisites, i.e. the influencing factors of the system, and in terms of the resulting performance metrics.

The final publication is available at Springer via DOI: 10.1007/s12193-011-0088-y

Obviously, a certain level of system performance is necessary to fulfill the user's needs, as it is stated in the above definition. However, QoS does not determine user satisfaction in terms of a strict cause-and-effect relationship. On the contrary, user perception and satisfaction comes in a multitude of different aspects, each of which may (or may not) be influenced by the performance of individual service components. An established way to partly deal with this is to find relationships between single system factors and user perception by applying standardized test – this is named "User-perceived QoS" or "Quality of Perception" by ETSI [16] and is limited to user perception [17].

In telecommunications, the term *Quality of Experience (QoE)* has recently been used for describing all such aspects which finally result in the acceptability of the service [18]. In other areas (such as speech or sound quality engineering), the term "quality" is used instead, being defined as the "result of appraisal of the perceived composition of a unit with respect to its desired composition" [19]. Quality is thus far more than performance – it is the degree of fulfillment of the user's expectations and needs. Note that – following this definition – the measurement of quality requires a perception and a judgment process to take place inside the user. Thus, measurement of QoE usually relies on interaction experiments with real or test users and subsequent analysis of both, questionnaire data and performance measures. In contrast to this QoS can be quantified by a person external to the interaction process, e.g. by the system developer, by relying solely on performance measures.

For interactive systems based on the speech modality alone (so-called spoken dialogue systems), efforts have been made to come up with a standard set of QoS and QoE metrics. These include interaction parameters describing the performance of system components and the behavior of user and system during the interaction [20][21] as well as questionnaires for collecting quality judgments [20][22][23]. This is a distinguishing feature compared to other taxonomies defined for the general case of human computer interaction, which typically define components of usability on a rather abstract level, i.e. not defining instrumentation (e.g., [24][25]). The taxonomy presented here follows the former approach, that is, to systematise concepts and parameters which can be measured in an interaction, focusing on the special case of multimodal systems. For multimodal dialogue systems, there is only one proposal for standard metrics

The final publication is available at Springer via DOI: 10.1007/s12193-011-0088-y

[26][24]. One reason for this is a lack of understanding of the relevant QoS and QoE aspects of such systems.

Our final aim is to bundle evaluation efforts so that the insights gained in individual campaigns can be applied to other systems in the future. A prerequisite to this is to agree on which aspects to evaluate, and how to evaluate. As a first step towards this aim, we propose a taxonomy of QoS aspects (system factors and performance aspects of the interaction) and QoE aspects (related to human perception and judgment of the interaction). This taxonomy is described in Section 2. In the subsequent Sections 3 – 5, we provide definitions and examples for the individual items of the taxonomy. As far as they are available, we provide metrics for quantifying them in the evaluation. Finally, Section 6 defines the next steps which are necessary to take full benefit of the approach.

2 Taxonomy of QoS and QoE aspects

Our taxonomy is based on a similar one developed for spoken dialogue systems in [27], but considers a broader range of input and output modalities (such as touch input, gesture recognition, audio-visual speech output, etc.) and more quality aspects. It is intended to serve as a framework for evaluation, and not as a precise protocol: Whereas standardized protocols such as the ones followed by DARPA or Blizzard [28][29] are very helpful for the development of core technologies, they provide little insight into the appropriateness of this technology for a system to-be-developed. Via a framework, in turn, developers are able to tailor evaluation to their individual needs.

The use of the taxonomy is threefold: System developers may search for the QoS and QoE aspect they are interested in and find the appropriate evaluation metrics. It could also serve as a basis for a systematic collection of evaluation data. And finally, if no evaluation metric is available for a given quality aspect, necessary methodological research can be identified.

The taxonomy consists of three layers, two of them addressing QoS and one addressing QoE, see Fig. 1:

The final publication is available at Springer via DOI: 10.1007/s12193-011-0088-y

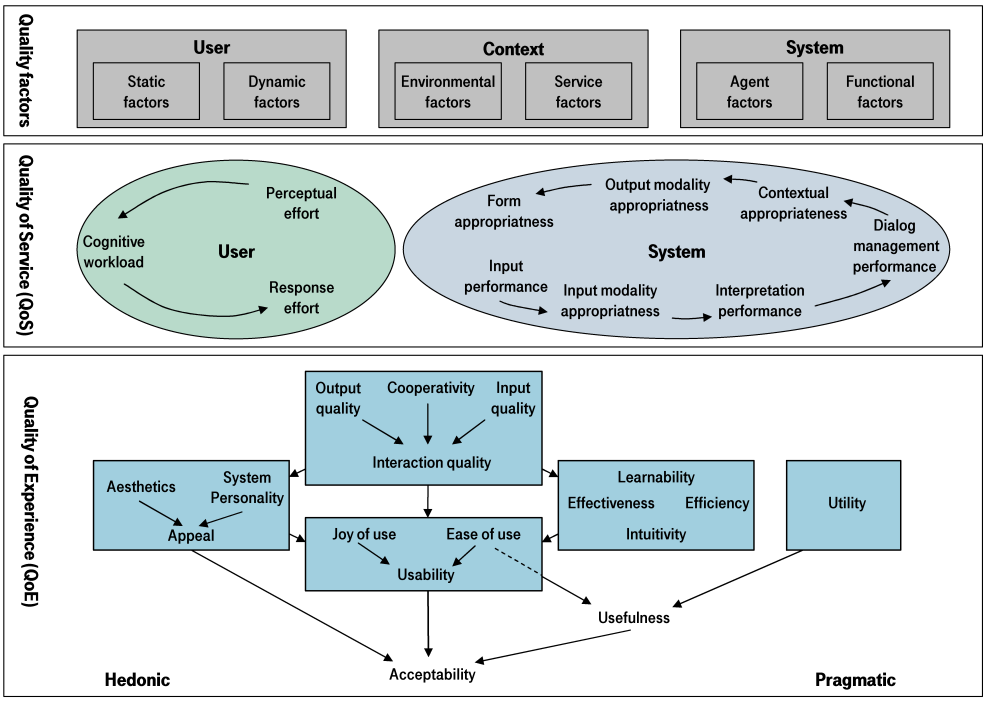

Figure 1: Taxonomy of QoS and QoE aspects of multimodal human-machine interaction, adapted from [X].

- Quality factors influencing QoS and QoE;
- QoS interaction performance aspects describing user and system performance and behavior;
- QoE aspects related to the quality perception and judgment processes. As a result of the interaction de-scribed on the second layer, quality "happens" as a multidimensional perceptual event in a particular context of use.

The layers are necessary to make a clear distinction between QoS and performance aspects (which can be determined by an external observer) on the one hand, and QoE and quality aspects (which have to be acquired from the user) on the other. By using three instead of two (QoS and QoE) layers we could better differentiate between the influencing factors (which are defined by the service and its users) and the resulting interaction performance aspects (which can be measured during the interaction).

It is obvious from Fig. 1 that there are relationships between quality factors, QoS interaction performance aspects, and QoE aspects. These relationships are mostly not one-to-one and can vary in their strength depending on the system, user and context. Inside each layer, however, relations between aspects are better defined and therefore indicated as far as possible.

The final publication is available at Springer via DOI: 10.1007/s12193-011-0088-y

As a result of the relationships, there are also dependencies between the associated metrics. Sometimes, metrics may point into contradicting directions, and the system developer has to carefully review the metrics in order to not take wrong design decisions. Metrics may also be combined in order to obtain estimations of global quality aspects, but care should be taken in the integration; simple linear combinations of different metrics might not be adequate.

Please note, that we propose questionnaires and performance metrics here as only the combination of both can give a comprehensive and valid picture for several reasons: a) Users might be influenced not to honestly report their experience; b) a combination will give the system developer more insight what system factors influenced noteworthy user ratings; c) for most of the QoE aspects described there exist no valid and reliable metrics for the case of multimodal systems, so a mixture can help to interpret the results better.

Also note that the metrics defined will not automatically result in better systems. For efficiently developing systems which provide a high quality to their users, usability engineering methods have to be used during the entire system lifecycle, involving proper analysis, design, prototyping, expert evaluation, user evaluation, and feedback cycles [24]. The metrics defined here will certainly help to quantify the progress made during the usability evaluation cycle, and to uncover the characteristics of the system, the user and the context of use which are relevant for achieving high quality. However, we think it is not possible to relate individual concepts and related metrics to each step of the system lifecycle, as the exact concepts to be measured and the results expected will largely depend on the type of system to be developed.

In the following sections, we provide definitions and examples for the individual items of the taxonomy. As far as they are available, we also present and explain metrics for quantifying them in the evaluation.

3 Quality factors

Quality factors exercise an influence on QoE through the interaction performance aspects. They include the characteristics of the user, the system and the context of use which have an impact on perceived quality.

The final publication is available at Springer via DOI: 10.1007/s12193-011-0088-y

3.1 User factors

User factors include all characteristics of the user which carry an influence on his/her interaction behavior and quality judgment. Some of these characteristics are static (e.g. age, gender, native language), others change dynamically from interaction to interaction, or even within an interaction (e.g. motivation, emotional status). Because systems cannot be designed for an individual user, users are commonly classified into groups which are relevant for the purpose of the evaluation. Such a classification can e.g. be performed on the basis of

- perceptional characteristics (e.g. visual or auditory impairments, typical for elderly users),
- behavioral characteristics (e.g. left-/ right-handed users, accented vs. non-accented speech),
- experience (e.g. with the system under investigation, with similar systems, with the task domain, with technology in general),
- motivation (for using the system), and
- individual preferences, capabilities or knowledge.

These characteristics carry an influence not only on the interaction behavior (and thus the interaction performance), but also on the quality which is influenced by the reference internal to the user (i.e. the “desired composition” in the definition of Section 1).

With particular relevance for multimodal information and communication technologies (ICT), Hermann et al. [30] and Naumann et al. [31] developed a scheme which classifies users according to their affinity towards technology, their general interaction methods and capabilities (cognitive capabilities, problem-solving strategies, purposefulness), as well as (less importantly) their domain knowledge, language competence, age, and orientation towards social norms. It results in seven user groups which show a distinct behavior towards and experience with such systems. However, there exists currently no screening questionnaire to classify users according to this scheme. Instead, there are several different questionnaires to assess user aspects like computer anxiety [32], computer literacy [33], attitudes towards computers [34] computer self-efficacy [35], computer experience [36], mental abilities [37] and so on. This situation has to be considered as quite problematic, as the different questionnaires do not only cover varying domains, ranging from basic electric or electronic devices to current

The final publication is available at Springer via DOI: 10.1007/s12193-011-0088-y

mobile multimedia interfaces, but the most validated ones are also the oldest, and thus maybe not applicable 20 years after creation and validation, as experience with technology and expectations constantly change.

3.2 System factors

System factors are the characteristics of the system as an interaction partner (agent factors) and those related to its functional capabilities (functional factors). Agent factors include the technical characteristics of the individual system modules (speech, gesture and/or face recognition; multimodal fusion, dialogue management, multimodal fission, etc.) as well as their aesthetic appearance [27]. Functional factors include the functionalities covered, the type of task (well-, or ill-structured, homo- or, heterogeneous, see [38], the number of available tasks, the task complexity, the task frequency, and task consequences (particularly important for security-critical systems) [27]. For less task-directed systems the domain characteristics gain importance (e.g. education or entertainment systems). Both agent and functional factors are commonly specified in advance, resulting in specification documents which list the characteristics, but mostly in a qualitative, not in a quantitative way. Most agent factors have to be specified by the system developer, however, aesthetics can usually better be specified by design experts or experienced salesmen. Functional factors can best be specified by domain experts; they may be the outcome of concept testing phases or focus group discussions where potential users try to find core and secondary functions which should be implemented in the final system, and weight them according to their importance for the later usage scenario.

3.3 Context factors

Context factors consist of two types. The so-called environmental factors capture the physical usage environment (home, office, mobile, or public usage), as well as any transmission channels involved in the interaction. These characteristics include any space, acoustic, and lighting conditions which might exercise an influence on the performance of the system or on the behavior of the user. The usage environment may also include potential parallel activities of the user; such activities have to be taken into account when evaluating the system, as they may reduce the cognitive resources which can be dedicated to the interaction with the system under test. A second class of context factors are the so-called service

The final publication is available at Springer via DOI: 10.1007/s12193-011-0088-y

factors, i.e. non-physical characteristics of the system and its usage which may carry an influence on how the user judges upon its quality, like access restrictions (from where the system can be reached), the availability of the system (restricted opening hours), any security or privacy issues resulting from the use of the system, and the resulting costs. The latter are very important for the final acceptance, as the user will try to find a balance between the value provided by the system and the price s/he is willing to pay for it.

Like the system factors, context factors are usually specified prior to system design, and mostly in a qualitative way. Usage contexts cannot always be anticipated by the developers, and sometimes it makes sense to use task analysis methods as in [24] in order to find out about the user's functional needs in specific usage contexts.

4 QoS interaction performance aspects

It is during the interaction when the perception and judgment processes forming quality take place. Interaction performance aspects are organized into two cycles (one for the system and one for the user), their order reflecting the processing step they are located in. These cycles and the respective interaction performance aspects are described in the following sub-sections.

4.1 System interaction performance aspects

System interaction performance aspects can be quantified with the help of interaction parameters, which are either logged during the interaction or annotated by an expert afterwards. While these interaction parameters are not directly linked to the perceived quality their interpretation can offer useful information to system developers. They are particularly useful to assess and compare the performance of the involved technologies, such as recognizers or fusion and fission components. Thus, a meaningful parameter set can only be selected (or defined) under consideration of the specifics of the technologies actually used. However, some general principles apply at each processing stage, which will be named in this section. For further reference, a list of multimodal interaction parameters has been

The final publication is available at Springer via DOI: 10.1007/s12193-011-0088-y

published in [24] and a first application for an evaluation of a multimodal dialogue system in [X].

4.1.1 Input performance

Input performance can be quantified e.g. in terms of its accuracy or error rate, as it is common practice for speech, gesture recognizers and facial expression recognizers. To compute those parameters an expert has to transcribe the user utterance or the handwriting input, or annotates the facial expressions and gestures concerning beginning and end as well as the interpretation of these. Typically, the number of correctly determined words, gestures or expressions, of substitutions, of insertions, and of deletions is counted. These counts can be divided by the total number of words, gestures or facial expressions in the reference to produce *error rates*. These measures will mostly be computed separately for each modality, an exception being the case of signal level fusion. In contrast to fusion on the semantic level, which has to be evaluated as part of the interpretation performance, fusion on the signal level can be considered as forming part of the input performance. The performance of the signal level fusion can then be measured by comparing the fusion results with the recognition results obtained with each modality separately (see [38] for examples).
In addition to that, the degree of coverage of the user's behavior (vocabulary, gestures, facial expressions) as well as the system's real-time performance are indicators of system input performance. The real-time performance can be captured in terms of *system feedback delay* and *system response delay* - measured from the end of user input to the beginning of system feedback, such as the display of loading status of a web page, or the beginning of the system response. Concerning special multimodal input like face detection, person or hand tracking, see [39] for metrics and even a corpus to evaluate system components in a comparable way.

4.1.2 Input modality appropriateness

Multimodal dialogue systems may offer a set of modalities for user input. These modalities may be used sequentially, simultaneously or compositely. Depending on the content, the environment and the user it can be determined (e.g. guided by modality properties as described in [40]) if the offered input modalities are appropriate for every given turn in a given context. For example, spoken input is

The final publication is available at Springer via DOI: 10.1007/s12193-011-0088-y

inappropriate for secret information like a PIN when it occurs in public spaces. Appropriateness can be annotated per modality or – in the case of composite input – for the multimodal input as a whole. In the first case, each modality can be appropriate or inappropriate. In the second case, the multimodal input can be appropriate, partially appropriate or inappropriate. From the annotations, rates of appropriate input modalities can be calculated by dividing through the number of times the system asked the user for input.

4.1.3 Interpretation performance

The performance of the system to extract meaning from the user input can be quantified in terms of accuracy when a limited set of underlying semantic concepts is used for meaning description. Example: Counting of the errors in filling the correct attribute-value pairs on the basis of an expert-derived measure of the "correct" interpretation. For independent input modalities, the *concept error rate* is typically calculated [21]. However, in the case of multimodal input, also the performance of the modality fusion component should be considered, as fusion on the semantic level can help to reduce the impact of recognition errors. This gain in accuracy (or *fusion gain*) can be evaluated by comparing the fused result with the results of the recognition modules for the different modalities.

4.1.4 Dialogue management performance

Depending on the function of interest, different metrics can be defined for the dialogue manager. Its main function is to drive the dialogue to the intended goal; this function can be assessed only indirectly, namely in terms of dialogue success (see below). In addition, goals should be achievable in an efficient and elegant way. Efficiency can be indicated by the *dialog duration* or the *number of dialog turns*. In addition, the *query density* can be computed as the quotient of unique concepts introduced by the user and the total number of concepts input to the system. Regarding the elegance of the dialog, several metrics are listed in [21]. These include the average *user turn duration* and *system turn duration* as well as the number of user input modality changes and system output modality changes. In addition, counts of special actions can be collected, such as the *number of system questions*, the *number of diagnostic system error messages*, the *number of timeouts*, the *number of help requests* or help messages or the *number of cancel attempts*. The dialogue manager's ability to correct misunderstandings can e.g. be

The final publication is available at Springer via DOI: 10.1007/s12193-011-0088-y

quantified by computing the *user and system correction rate* as the number of turns concerned with corrections in relation to the total number of turns.

4.1.5 Contextual appropriateness

The system response should be appropriate in a given context, where appropriateness can be defined based on Grice's Cooperativity Principle [41] and quantified in terms of violations of this principle [40]. Based on this, the *contextual appropriateness* parameter defined in [21] involves first an expert rating the appropriateness of each system response, and second calculating the rate of appropriate responses among all system responses.

4.1.6 Output modality appropriateness

As for the input side, *output modality appropriateness* can be checked on the basis of modality properties as defined in [40], taking into account the interrelations between simultaneous modalities. In the case of multimodal output, the assignment of an output modality to a system response is the task of the output modality fission. Its performance can be judged by an expert: has the appropriate modality been chosen for every bit of information? As for contextual appropriateness, the rate of appropriate modalities can then be determined.

4.1.7 Form appropriateness

Refers to the surface form of the output provided to the user. For example: Form appropriateness of spoken output can be measured via its intelligibility, comprehensibility, or the required listening effort. The appropriateness of an Embodied Conversational Agent (ECA) can be assessed by its ability to convey specific information, including emotions, turn-taking backchannels, etc. The synchrony of the output modalities – especially in the case of ECAs – can be measured in terms of the *lag of time* of each modality compared to the other modalities [24].

4.2 User interaction performance aspects

On the user's side, interaction performance can be quantified by the effort required from the user to interact with the system, as well as by the freedom of interaction. Aspects include:

The final publication is available at Springer via DOI: 10.1007/s12193-011-0088-y

Perceptual effort: Effort required decoding the system messages, understanding and interpreting their meaning [42] e.g. listening-effort or reading effort. Metrics: Borg's category-ratio scale [43].
Cognitive Workload: Specification of the costs of task performance (e.g. necessary information processing capacity and resources) [44]. There are several ways to measure workload. A simple and cheap option is to use questionnaires like the Nasa-TLX [45] or the RSME [46]. A more elaborate measure are psycho-physiological parameters like pupil diameter [47] or test stetting employing the dual task paradigm. An overview of methods assessing cognitive workload is given in [44].
Physical response effort: Physical effort required to communicate with the system. Example: Effort required for entering information into a mobile phone. Metrics: Questionnaires, e.g. [22].

5 QoE aspects

So far, we have limited ourselves to QoS, both in terms of influencing factors and of performance metrics. However, the ultimate aim of a system developer should be to satisfy the user's needs. According to Hassenzahl et al. [48], the user evaluation of a system is influenced by pragmatic and hedonic quality aspects. These quality aspects have to be evaluated with the help of real or test users providing judgments on what they perceive. Such judgments can be seen as "direct" QoE measurements. In addition to that, "indirect" QoE measurements can be obtained by logging user behavior and relating it to perceived QoE [49][50].

5.1 Interaction quality

This term relates to the quality of the pure interaction between user and system, disregarding any other aspects of system usage. It includes the perceived input quality, the perceived output quality, as well as the system's cooperativity.
Input quality relates to the *perceived* system understanding and input comfort; in contrast to input performance, it reflects how the user thinks that s/he is understood by the system. Studies have shown that the relationship between e.g. perceived system performance and actual performance is only weak at best [51]. One underlying reason is that the user frequently does not know which concepts

The final publication is available at Springer via DOI: 10.1007/s12193-011-0088-y

have been understood by the system, unless there is a direct feedback. In other cases, this might only be detected in further stages of the dialogue, or a misunderstanding might even never be detected during the interaction, potentially resulting in task failure unnoted by the user.

Output quality refers to the perceived system understandability, and to its form appropriateness. This includes whether the meaning of a system message can be discerned, and whether the form supports meaning extraction by the user. Meaning extraction may be limited by the output performance of the system (e.g. legibility of characters on the screen, intelligibility of synthesized speech), but goes significantly beyond this by taking into account the content of the system message in the immediate dialogue context, which will bear on the perceived transparency of the system.

Output quality is strongly related to system cooperativity, as the user can judge the system's support in reaching a joint goal mainly via the system output messages. Beyond the perceived quality of the system output, cooperativity includes the distribution of initiative between the partners (which may be asymmetric because of different roles and expectations), the consideration of background knowledge of the user and the system, and the ability for repair and clarification.

Questionnaires have been developed to quantify a variety of interaction quality aspects. For speech-based interfaces, the framework provided in ITU-T Rec. P.851 [22] captures the most relevant aspects such as the input and output quality, the perceived speed/pace of the interaction, and the smoothness and naturalness of the interaction. It is mainly based on the SASSI [23] questionnaire which has been developed for systems with speech input capabilities, and which has been extended towards the output quality. The framework is currently being extended towards multimodal systems.

5.2 Usability

According to the ISO definition, usability is the "extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" [52]. Although this is probably the most common definition, relevant literature offers a large variety of different additional definitions. Apart from this it has to be mentioned that the term User

The final publication is available at Springer via DOI: 10.1007/s12193-011-0088-y

eXperience (UX) got increasingly popular during the last decade. The ISO standard 9241-210:2010 [53] defines UX as ”a person's perceptions and responses that result from the use or anticipated use of a product, system or service“. According to [54] this definition permits three different interpretations of the term UX: First of all UX can be understood as “an umbrella term for all the user’s perceptions and responses […]”. Secondly, UX can be understood as a different may be even as counter concept to usability as historically the focus of usability is mainly on performance. The third interpretation sees UX as “an elaboration of the satisfaction component of usability”. In order to be consistent with the other concepts defined on this layer of the taxonomy and to have a more fine-grained distinction, we adopted the last view and we will thus consider two aspects of usability: The Ease of Use, which is influenced by the mentioned consequences of interaction quality, and the Joy of Use, which is often associated with UX. Joy of Use depends not only on the quality of the interaction; hedonic aspects like the appeal or the “personality” of the system will largely influence this sub-aspect. Both, Ease of Use and Joy of Use may determine the satisfaction of the user, which can be considered as a consequence of good usability.

Our concept of usability follows [54] and incorporates hedonic as well as pragmatic qualities, thus we go far beyond a strict performance-related approach. We do not use the term User Experience within the taxonomy, as the definition provided by ISO is very broad and allows for different interpretations (see previous paragraph). Instead, the last layer tries to provide a structured picture of the UX sub-concepts.

Following the definition above questionnaires need to measure both, Joy and Ease of Use. Although several questionnaires measure the “Ease of Use-part” of usability only few include the Joy of Use. An affect scale is included in the SUMI [55] the PSUQ [56] and the CSUQ [56] additionally measure frustration. The AttrakDiff’s “attractiveness” scale, measuring pragmatic (Ease of Use) as well as hedonic qualities (Joy of Use), is probably closest to our conception of usability [57].

Apart from the questionnaires presented above other suitable methods to assess Joy and Ease of Use include qualitative approaches like the Repertory Grid Technique [58] the Valence Method proposed by Burmester [59] and the UX Curve [60]. The Repertory Grid Technique has its origin in the psychology of

The final publication is available at Springer via DOI: 10.1007/s12193-011-0088-y

personal constructs by Kelly [58]. Constructs in Kelly's sense are bipolar (dis)similarity dimensions [61]. According to Kelly [58][54] every human owns an individual and characteristic system of constructs, through which he/she perceives and evaluates his/her experiences. The Repertory Grid Technique aims to assess these individual constructs using two phases. In the elicitation phase the persons is presented with triads of the relevant object (e.g. three websites as in [62]) and is asked to verbalize what two objects have in common and how they differ from the third. This way bipolar constructs in the form of a semantic differential are generated. These bipolar constructs are later used as the rating scale for all constructs. The result is an individual construct-based description of the objects [61].

The Valence Method [85] is a two-phase measure based on the theoretical work by [63]. In the first part, the users are asked to set positive and negative valence markers while exploring the interface for up to eight minutes. The whole session is videotaped. In the next phase, the marked situations are presented to the participants again while the interviewer is asking which design aspect was the reason for setting the marker. The laddering interviewing technique is employed to uncover the underlying need by repeating question why a certain attribute was mentioned until the affected need is identified.

The main limitation according to the authors [59] is that it is currently recommendable for first usage situations only as the number of markers increases substantially if the product is already known to the user: Valuable insights in form of quantitative data (number of positive and negative markers) and qualitative data (interviews) can be gained with this method, one disadvantage probably being the relatively high resources required.

The UX Curve [60] is used to retrospectively assess the system's quality over time. Participants are asked to draw curves describing their perceptions of the system's attractiveness, the system's ease of use, the system's utility, as well as the usage frequency and their general experience over time. Also users should explain major changes in the curves. According to the authors, the UX curve allows to measure the long-term experience and the influences that improve or decrease the perceived quality of the experience. A similar methods is offered with iScale [64]. Again users are asked to draw a curve reflecting their experience.

The final publication is available at Springer via DOI: 10.1007/s12193-011-0088-y

However, iScale is an online survey instrument while the UX curve was developed for face-to-faces setting. Also the assessed dimensions differ.

5.3 Ease of use

The perceived ”Ease of Use“ describes the extent to which users assume that the usage of a system will be effortless [65]. Relevant determinants for this construct are the aspects described in the ISO 9241 standard [52], namely efficiency and effectiveness and, moreover, learnability and intuitivity.

The effectiveness refers to the accuracy and completeness with which specified users can reach specified goals in particular environments [52]. Efficiency is the effort and resources required in relation to the accuracy and completeness achieved [52]. The vast majority of standardized usability questionnaires cover these two constructs. Examples are the QUIS [66], the SUS [67], the IsoMetrics Usability Inventory [68], the AttrakDiff [57], the SASSI [23], and the USE [69]. It has to be noted that the questionnaires subscales are not necessarily named efficiency or effectiveness. The SASSI subscale “speed” is strongly related to efficiency, the scale pragmatic-qualities on the AttrakDiff refers to both, efficiency and effectiveness. Also performance data can be used to operationalize these aspects: task duration might serve as an efficiency measures, tasks success as an effectiveness measure. Also learnability, the ease with which novice users can start effective interactions and maximize performance is covered by most usability questionnaires. This can not be said for intuitivity, the degree the user is able to interact with a system effectively by applying knowledge unconsciously. Intuivity might be associated to constructs covered by established questionnaires like familiarity or self-descriptiveness. However despite intuitivity being often considered as an important determinant of a products quality, it is not as commonly included in usability evaluations as the above mentioned aspects are. Only recently questionnaires, specially focusing on intuitivity, have been developed [70][71].

Other methods regularly used for the evaluation of Ease of use are expert-oriented procedures like the Cognitive Walkthrough [72] and modelling approaches e.g. GOMS [73]. The Cognitive Walkthrough is rooted in theories of explorative learning. Experts, usually designers or psychologist, analyse the system’s

The final publication is available at Springer via DOI: 10.1007/s12193-011-0088-y

functionalities based on a description of the interface, the tasks, the action necessary to perform the task and information about the user and the usage context. Critical information is recorded by the experts using a standardized protocol.

With the method GOMS the interaction with a system is reduced to basic elements, which are goals, methods, operators and selection rules. Goals are the basic goals of the user, what he/she wants to achieve while using the system [bonnie john]. Operators are the actions offered by the system to accomplish the goals. Methods are well-learned sequences of sub-goals and operators suitable to achieve a goal [73] . Selections rules apply if several methods are possible and reflect the user's personal preferences. These four basic elements describe the procedural knowledge necessary to perform the tasks. This knowledge is applied to the design to check if the system provides methods for all user goals, furthermore execution times of well-trained, error-free expert users can be predicted.

In case of multimodal systems, GOMS analyses can become quite extensive due to the complexity of such systems. As multimodal systems allow for parallel, serial or combined usage of different modalities multiple methods for one goal are possible which than require the definition of multiple selection rules. The EPIC framework by [74] is a more sophisticated architecture better suitable for predicting execution times for interactions with multimodal systems, however, EPIC is first and foremost a research system and thus not focused on being a tool for evaluation purposes [74].

5.4 Joy of Use

According to Schleicher and Trösterer [75] "Joy of Use" is the conscious positive experience of a systems quality. Important determinant is the products aesthetic. Although the term aesthetic often refers to visual aesthetics only, we go along with the following, broader definition proposed by Hekkert [76]. Thus "aesthetic is the pleasure attained from sensory perception, as opposed to anesthetic. An experience of any kind […] comprises an aesthetic part, but the experience as a whole is not aesthetic." [76]. This definition implies that aesthetic can be experienced through all our senses, respectively modalities, and is not limited to

The final publication is available at Springer via DOI: 10.1007/s12193-011-0088-y

visual aesthetics. The system's personality refers to the users' perception of the system characteristics originating from the current combination of agent factors and surface form. This includes system factors like chosen gender, appearance or voice for embodied conversational agents, wording of voice prompts, structure, color and icon scheme for graphical displays, that should exhibit a consistent and application-adequate personality (also called character) [77][78][79].

The appeal is a result of the aesthetics of the product, its physical factors and the extent to which the product inherits interesting, novel, and surprising features [48][80].

It is noteworthy that there is an ongoing debate concerning the relationship between hedonic qualities, the aspects related to Joy of Use, and pragmatic qualities, the aspects related to Ease of Use. While some findings provide evidence for the claim that "what is beautiful is usable" [81] and for the underlying assumption that Joy of Use and Ease of Use are interdependent; other studies could not confirm these results [82]. Hassenzahl [83] suggests that these ambiguous results are caused by different approaches in understanding and measuring aesthetics. Accordingly, a variety of methods is available to measure "Joy of Use" related aspects but before deciding on a measurements method it has to be defined which aspect should be assessed. The questionnaire proposed by Lavie and Tractinsky [84] is suitable for measuring the visual aesthetics, but not for aesthetics perceived via other modalities. The AttrakDiff [57] measures hedonic qualities on a higher level and is not limited to unimodal interfaces. For measuring hedonic qualities during the interaction the Joy of Use-Button [75] and psycho-physiological parameters are available options with the latter being the most resource-intensive method.

Another well validated and widely used instrument is the Self-Assesment Manikin [85], which measures the arousal, pleasure and dominance linked to affective reactions on three non-verbal scales.

If the aim is to measure specific emotions LemTool [86] and PrEmo [87] can be used. However, both tools are so far only validated for specific application areas only: Lemtool for websites [86] and PrEmo for non interactive products [87].

However, although a wide range of methods assessing hedonic, affective qualities are nowadays available a recent review by [88] indicates that questionnaires, more

The final publication is available at Springer via DOI: 10.1007/s12193-011-0088-y

specifically the hedonic qualities subscales of Hassenzahl's Attrakdiff [48] and the Self Assessment Manikin by [85], are by far the most popular instrument.

5.5 Utility and usefulness

In order to judge whether a system is useful, we have to compare the functional requirements of the user with the functions offered by the system. Utility answers the question: Can a specific user resolve his personal task with the help of the system [89]? Usefulness relates this to the usability: How well can a user resolve the task, considering the effort spent in the interaction, but also the joy experienced herein? Questionnaires focusing on utility and usefulness are rather rare. However usefulness as understood in our framework is a subscale in the PSSUQ as well as in the CSUQ [56]. Those questionnaires are identical except that PSSUQ was developed to use after a usability test and is thus addressing specific tasks whereas the CSUQ asks about the general system and is suitable for surveys [56].

5.6 Acceptability

The term acceptability describes how readily a user will actually use the system. There are some disputes about whether acceptability is still a part of QoE, as it may be represented as a purely economic measure, relating the number of potential users to the quantity of the target group [90]. Thus, we consider it to be a consequence of the quality aspects and the influencing factors described in the taxonomy.

Unfortunately, it is still ill understood how acceptability is really influenced by QoE aspects; user needs as well as economic considerations may be dominant in certain fields and rule out the advantages of good QoE, at least temporarily. For example, Chateau et al. [50] have postulated that the relative importance of factors on acceptability varies roughly from decade to decade, and that the price may significantly outperform perceived quality, as long as the latter is still above a certain threshold. In-depth and long-term analyses of the relationship between QoE and acceptability are thus needed to further clarify the value of high QoE for the success on the market.

On of the most influential approach in determining acceptance is the Technology Acceptance Model (TAM) [91]. Theoretical base of the model is the theory of reasoned actions (TRA) by Ajzen and Fishbein [92]. According to the TRA actual

The final publication is available at Springer via DOI: 10.1007/s12193-011-0088-y

behavior is determined by behavioral intentions. Behavioral intentions are dependent on attitudes towards the behaviors and subjective norms. In terms of the TAM the attitudes towards the behavior are the perceived usefulness and the perceived Ease of Use. Subjective norms are not included in TAM. It is important to note that the perceived usefulness, described in the TAM, does not exactly match our understanding of these terms. More precisely, perceived usefulness is defined as "the degree to which a person believes that using a particular system would enhance his or her job performance". Thus the subscale usefulness in the TAM questionnaire is not to be used to measure the conceptualization of usefulness given in Chapter 5.3. However to measure acceptance the questionnaire is helpful although it will not provide as detailed information about the system's quality as other questionnaires mentioned.

As Acceptability is also affected by social constraints beyond the experienced system's quality, it is a complex and multidimensional QoE aspect only assessable in the field or even in the market with valid results. In our opinion Acceptability is primarily a topic for marketing and not for interaction research.

6 Conclusions and future work

The presented taxonomy provides definitions of factors and aspects as well as information about their relationships. On this common ground, comparable evaluation can be performed, its results can be identified and categorized, and metrics for specific purposes (or the lack of those) can be identified. A first example of an application of the taxonomy – however without relating it to QoS and QoE – is the evaluation of multimodal interfaces for intelligent environments presented in [X].

Still, the presented metrics are not exhaustive, and many of the established metrics are not sufficiently evaluated for their application to multimodal systems. For example, there is no standardized questionnaire available for assessing the interaction quality of multimodal applications. As current systems cover a wide range of modalities, applications and domains, we anticipate that an open framework will be needed to enable meaningful evaluation for specific contexts. Also, concerning performance measures describing HCI, interaction parameters are already defined for multimodal interfaces [24]. However, there exist only

The final publication is available at Springer via DOI: 10.1007/s12193-011-0088-y

preliminary results concerning the relationships between these measures and aspects of QoS and QoE [X].

The taxonomy has been developed on the basis of own evaluation exercises, as well as on the basis of a broad literature survey. The attribution of concepts to the different layers was guided by the definitions cited above, as well as by the intuition of usability engineers in our department. It would be good to validate this classification with further usability experts from other domains, so that the taxonomy becomes more generic and stable. This could e.g. be done in a card-sorting experiment where usability experts have to organize QoS and QoE concepts according to certain rules.

Up to now, practical application examples of the taxonomy are still very limited, and they are limited to a few systems in the telecommunication sector. In order to further substantiate the taxonomy, our aim is to classify further evaluations described in the literature according to the taxonomy. This will help to identify further evaluation metrics for individual QoS and QoE aspects, and to validate them on a larger basis of systems and user groups. Furthermore, we will use such evaluations to systematically identify relationships between quality factors on the one hand, and QoS and QoE aspects on the other. As soon as such relationships are identified, it will be able to judge the impact of a certain quality factor in advance, and to design the system accordingly.

Overall, we aim at an integration of the taxonomy described here in the usability engineering lifecycle. We expect that for the early steps of the lifecycle (analysis, design, prototyping) a list of quality factors can be defined which facilitate the proper specification of all characteristics of the systems which are relevant for its quality. For the later steps of the lifecycle (expert evaluation and user texting), the metrics defined for the QoS and QoE layers will help to capture relevant aspects so that adequate conclusions can be drawn for the (re-) design of the system.

As stated above, the taxonomy is designed as a framework facilitation evaluation, but not as a strict model which could be implemented in order to predict acceptability. Still, we see the potential that algorithmic relationships can be defined between quality factors as input elements, QoS interaction performance metrics as mediating elements, and QoE aspects as output results. The relationships could be described by deterministic linear or non-linear formulae, or by a probabilistic framework such as a Hidden Markov Model. Although some of

the concepts addressed in the taxonomy are still immature (such as aesthetics, appeal), others such as interaction quality are well operationalized so that meaningful predictions of quality aspects get within reach. The definition of algorithmic relationships will be very helpful both for a target-oriented design and optimization, as well as for the online adaptation of future multimodal dialogue systems.

7 References

[1] L. Lamel, S. Bennacef, J.L. Gauvain, H. Dartigues, and J.N. Temem, "User Evaluation of the MASK Kiosk", *Speech Communication* 38, pp. 131–139, 2002.

[2] N.O. Bernsen, L. Dybkjær, and S. Kiilerich, "Evaluating Conversation with Hans Christian Andersen", in *Proc. LREC 2004*, vol. 3, pp. 1011–1014, Lisbon, 2004.

[3] W. Wahlster, *SmartKom: Foundations of Multimodal Dialogue Systems*, Springer, Berlin, 2006.

[4] M. Turunen, J. Hakulinen, O. Ståhl, B. Gambäck, P. Hansen, M. Rodriguez Gancedo, R. Santos de la Camara, C. Smith, D. Charlton, and M. Cavazza, "Multimodal and Mobile Conversational Health and Fitness Companions", *Computer Speech and Language* 25, pp. 192–209, 2011.

[5] J. Cassell, S. Kopp, P. Tepper, K. Ferriman, and K. Striegnitz, "Trading Spaces: How Humans and Humanoids use Speech and Gesture to Give Directions", in *Conversational Informatics* (T. Nishida, ed.), John Wiley & Sons, New York NY, pp. 133–160, 2007.

[6] D. Gibbon, I. Mertins, and R.K. Moore, Eds., *Handbook of Multimodal and Spoken Dialogue Systems*, Kluwer Academic Publishers, Boston MA, 2000.

[7] J. Coutaz, L. Nigay, D. Salber, A.E. Blandford, J. May, and R.M. Young, "Four easy pieces for assessing the usability of multimodal interaction: The CARE properties", in *Human-Computer Interaction, Proc. Interact 1995* (K. Nordby, P.H. Helmersen, D.J. Gilmore, and S.A. Arnesen, eds.), Chapman & Hall, London, pp. 115–120, 1995.

The final publication is available at Springer via DOI: 10.1007/s12193-011-0088-y

[8] C.D. Wickens, “Multiple resources and mental workload”, *Human Factors: The Journal of the Human Factors and Ergonomics Society* 50, pp. 449–455, 2008.

[9] N.O. Bernsen and L. Dybkjær, *Multimodal Usability*, Human–Computer Interaction Series, Springer, Berlin, 2010.

[10] ISO 25000, “Software Engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Guide to SQuaRE”, *International Organization for Standardization*, Geneva, 2005

[11] W.D. Gray and M.C Salzman, “Damaged merchandise? A review of experiments that compare usability evaluation methods”, *Human-Computer Interaction, 13*(3), pp. 203–261, 1998.

[12] G.M. Olson and T.P. Moran, “Commentary on ‘*Damaged* Merchandise?’” *Human Computer Interaction, 13*(3), pp. 263–323, 1998.

[13] K. Hornbæk, “Current practice in measuring usability: Challenges to usability studies and research”, *International Journal of Man-Machine Studies 64*(2), pp. 79–102, 2006.

[14] H.R. Hartson, T.S. Andre, and R.C. Williges, “Criteria for evaluating usability evaluation methods”, *International Journal of Human-Computer Interaction 15*(1): pp. 145–181, 2003.

[15] ITU-T Rec. E.800, *Terms and Definitions Related to Quality of Service and Network Performance Including Dependability*, International Telecommunication Union, Geneva, 1994.

[16] ETSI TR 102 643, *Human Factors (HF); Quality of Experience (QoE) requirements for real-time communication services*, European Telecommunications Standards Institute, Sophia Antipolis, 2010.

[17] P. Brooks and B. Hestnes, “User measures of quality of experience: why being objective and quantitative is important”, *IEEE Network* 24, pp. 8–13, 2010.

[18] ITU-T Rec. P.10, *Vocabulary for Performance and Quality of Service*, International Telecommunication Union, Geneva, 2007.

[19] U. Jekosch, *Voice and Speech Quality Perception. Assessment and Evaluation*, Springer, Berlin, 2005.

[20] N. Fraser, “Assessment of Interactive Systems”, in *Handbook on Standards and Resources for Spoken Language Systems* (D. Gibbon, R. Moore, and R. Winski, eds.), Mouton de Gruyter, Berlin, pp. 564–615, 1997.

[21] ITU-T Suppl. 24 to P-Series Rec., *Parameters Describing the Interaction with Spoken Dialogue Systems*, International Telecommunication Union, Geneva, 2005.

[22] ITU-T Rec. P.851, *Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems*, International Telecommunication Union, Geneva, 2003.

[23] K.S. Hone and R. Graham, “Towards a Tool for the Subjective Assessment of Speech System Interfaces (SASSI)”, *Natural Language Engineering 6*(3/4), pp. 287–303, 2000.

[24] J. Nielsen, *Usability Engineering*, Academic Press, Boston MA, 1993

[25] D. Alonso-Ríos, A. Vázquez-García, E. Mosqueira-Rey, and V. Moret-Bonillo, “Usability: A Critical Analysis and a Taxonomy”, *International Journal of Human–Computer Interaction 26(1),* 53–74, 2010.

[26] S. Möller, C. Kühnel, and B. Weiss, “Extending Suppl. 24 to P-Series towards multimodal systems and services”, *International Telecommunication Union*, Geneva, Source: Deutsche Telekom Laboratories, ITU-T SIG12 Meeting 18–27 May 2010.

[27] S. Möller, *Quality of Telephone-Based Spoken Dialogue Systems*, Springer, New York NY, 2005.

[28] D. Pallett, J. Fiscus, W. Fisher, and J. Garofolo, “Benchmark Tests for the DARPA Spoken Language Program”, in *Proc. DARPA Human Language Technology Workshop*, Princeton, pp. 7–18, 1993.

[29] C.L. Bennett, and A.W. Black, “The Blizzard Challenge 2006”, in *Proc. Blizzard Satellite Workshop to Interspeech 2006*, Pittsburgh, 2006.

[30] F. Hermann, I. Niedermann, M. Peissner, K. Henke, and A. Naumann, “Users Interact Differently: Towards a Usability-Oriented Taxonomy”, in *Interaction Design and Usability, Proc. HCII 2007* (J. Jacko, ed.), vol. 1, pp. 812–817, Springer, Heidelberg, 2007.

[31] A.B. Naumann, F. Hermann, M. Peissner, and K. Henke, „Interaktion mit Informations- und Kommunikationstechnologie: Eine Klassifikation von Benutzertypen“ [Interaction with Information and Communication Technology:

A Classification of User Types], in *Mensch & Computer 2008: Viel Mehr Interaktion* (M. Herczeg and M.C. Kindsmüller, eds.), Oldenbourg Wissenschaftsverlag, München, pp. 37–45, 2008. (http://mc.informatik.uni-hamburg.de/konferenzbaende/mc2008/konferenzband/mc2008_05_naumann.pdf)

[32] R.K. Heinssen, C.R. Glass, and L.A. Knight, "Assessing computer anxiety: Development and validation of the Computer Anxiety Rating Scale", *Computers in Human Behavior*, 3(1), pp. 49–59, 1987.

[33] P.J.A. Van Vliet, M.G. Kletke, and G. Chakraborty, "The Measurement of Computer Literacy – A Comparison of Self-Appraisal and Objective Tests", *International Journal of Human-Computer Studies* 40(5), pp. 835–857, 1994.

[34] T. Richter, J. Naumann, and N. Groeben, „Attitudes toward the computer: Construct validation of an instrument with scales differentiated by content", *Computers in Human Behavior* 16, pp. 473–491, 2000.

[35] D.R. Compeau and C.A Higgins, "Computer self efficacy: Development of a measure and initial test," *MIS Quarterly*, pp. 189–211, June 1995.

[36] B. Smith, P. Caputi, and P. Rawstorne, "The development of a measure of subjective computer experience", *Computers in Human Behavior* 23(1), pp. 127–145, 2007.

[37] E. Kalbe, J. Kessler, P. Calabrese, R. Smith, A.P. Passmore, M. Brand, and R. Bullock, "DemTect: a new, sensitive cognitive screening test to support the diagnosis of mild cognitive impairment and early dementia", *International Journal of Geriatric Psychiatry* 19, pp. 136–143, 2004..

[38] R. Lopez-Cozar Delgado and M. Araki, *Spoken, multilingual and multimodal dialogue systems – Development and Assessment*, Wiley, 2005.

[39] D. Mostefa, M.-N. Garcia, and K. Choukri, "Evaluation of Multimodal Components within CHIL: The Evaluation Packages and Results", in *Proc. LREC 2006*, pp. 915–918, Genoa, 2006.

[40] N.O. Bernsen, "Multimodality in Language and Speech Systems – From Theory to Design Support Tool", in *Multimodality in Language and Speech Systems* (B. Granström, D. House, and I. Karlsson, eds.), Kluwer Academic Publishers, Dordrecht, pp. 93–148, 2002.

[41] H. Grice, “Logic and Conversation”, in *Syntax and Semantics, Vol. 3: Speech Acts* (P. Cole, and J.L. Morgan, eds.), Academic Press, New York NY, pp. 41–58, 1975.

[42] P.G. Zimbardo, *Psychologie*, Springer, Berlin, 1995.

[43] G. Borg, “Psychophysical Bases of Perceived Exertion”, *Medicine and Science in Sports and Exercise* 14, pp. 377–381, 1982.

[44] D. De Waard, *The Measurement of Drivers' Mental Workload*, PhD thesis, University of Groningen, Haren, 1996.

[45] S.G. Hart and L.E. Staveland, “Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research”, in *Human Mental Workload* (P. Hancock and N. Meshkati, eds.), North Holland, Amsterdam, pp. 139–183, 1988.

[46] F.R.H. Zijlstra, *Efficiency in work behavior. A design approach for modern tools*. PhD thesis, Delft University of Technology, Delft University Press, Delft, 1993.

[47] J.T. Cacioppo, L.G. Tassinary, and G.G. Berntson, Eds., *Handbook* of *psychophysiology*, 3rd edition, Cambridge University Press, New York NY, 2002.

[48] M. Hassenzahl, A. Platz, M. Burmester, and K. Lehner, “Hedonic and Ergonomic Quality Aspects Determine a Software's Appeal”, in *Proc. CHI 2000*, Den Haag, pp. 201–208, 2000.

[49] R.L. Mandryk, K. Inkpen, and T.W. Calvert, “Using Psycho-physiological Techniques to Measure User Experience with Entertainment Technologies”, *Behaviour and Information Technology 25*(2), pp. 141–158, 2006.

[50] N. Chateau, L. Gros, V. Durin, and A. Macé, “Redrawing the Link Between Customer Satisfaction and Speech Quality”, in *Proc. 2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems*, Berlin, pp. 88–94, 2006.

[51] K. Hornbæk and E.L. Law, “Meta-analysis of correlations among usability measures,” in *Proc. CHI 2007*, ACM, New York NY, pp. 617–626, 2007.

[52] ISO 9241-11, “Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs). Part 11: Guidance on Usability”, *International Organization for Standardization*, Geneva, 1999.

[53] ISO DIS 9241-210:2010. “Ergonomics of human system interaction – Part 210: Human-centred design for interactive systems” (formerly known as 13407). *International Organization for Standardization (ISO*). Switzerland.

[54] N. Bevan, “What is the difference between the purpose of usability and user experience evaluation methods”, in *Proc. UXEM'09 (INTERACT'09)*, Uppsala, 2009.

[55] J. Kirakowski and M. Corbett, “SUMI: The Software Usability Measurement Inventory”, *British Journal of Educational Technology 24*(3), pp. 210–212, 1993.

[56] J.R. Lewis, “IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use”, *International Journal of Human–Computer Interaction* 7, pp. 57–78, 1995.

[57] M. Hassenzahl, “The interplay of beauty, goodness, and usability in interactive products”. *Human-Computer Interaction* 19, pp. 319–349, 2004.

[58] G.A. Kelly, *The Psychology of Personal Constructs.* Norton, New York NY, 1955.

[59] M. Burmester, M. Mast, K. Jäger, and H. Homans. “Valence method for formative evaluation of user experience”, in *Proc. DIS '10*, ACM, New York, NY, pp. 364–367, 2010.

[60] S. Kujala, V. Roto, K. Vaananen-Vainio-Mattila, E. Karapanos, and A. Sinnela, “UX Curve: A Method for Evaluating Long-Term User Experience.” *Interacting with Computers*, 2011, doi: 10.1016/j.intcom.2011.06.005.

[61] M. Hassenzahl and R. Wessler, “Capturing design space from a user perspective: The repertory grid technique revisited”, *International Journal of Human-Computer Interaction 12*(3&4), pp. 441–459, 2000.

[62] M. Hassenzahl and T. Trautmann, “Analysis of web sites with the repertory grid technique”, In *Proc. CHI 2001* ACM, New York NY, pp. 167–168, 2001.

[63] M. Hassenzahl, S. Diefenbach and A.S. Göritz, “Needs, affect, and interactive products - Facets of user experience”, *Interacting with Computers 22*(5), pp. 353–362, 2010.

[64] E. Karapanos, J.-B. Martens, and M. Hassenzahl, “On the Retrospective Assessment of Users’ Experiences Over Time: Memory or Actuality?“, in *CHI*

2010 extended abstracts on Human factors in computing systems. Atlanta, ACM Press, pp. 4075–4080, 2010.

[65] F. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology", *MIS Quarterly 13*(3), pp. 319–340, 1989.

[66] J. Chin, V. Diehl, and K. Norman, "Development of an instrument measuring user satisfaction of the human-computer interface", in *Proceedings of SIGCHI 1988*, pp. 213–218, 1988.

[67] J. Brooke, "SUS: A "quick and dirty" usability scale", in *Usability Evaluation in Industry* (P. Jordan, B. Thomas, B. Weerdmeester, and I. McClelland, eds.), Taylor & Francis, London, pp. 189-194,1996.

[68] G. Gediga, K.-C. Hamborg, and I. Düntsch, "The IsoMetrics Usability Inventory: An operationalisation of ISO 9241-10", *Behaviour and Information Technology* 18, pp. 151–164, 1999.

[69] A.M. Lund, "Measuring usability with the USE questionnaire", *Usability and User Experience* 8, STC Usability SIG Newsletter, 2001.

[70] D. Ullrich and S. Diefenbach, „INTUI. Exploring the Facets of Intuitive Interaction", in *Mensch & Computer 2010 Interaktive Kulturen*, 2010.

[71] A. Naumann, J. Hurtienne, J.H. Israel, C. Mohs, M.C. Kindsmüller, H.A. Meyer, and S. Husslein, "Intuitive Use of User Interfaces: Defining a Vague Concept", in *Engineering Psychology and Cognitive Ergonomics, Proc. HCII 2007* (D. Harris, ed.), vol. 13, LNAI 4562, pp. 128–136, Springer, Heidelberg, 2007.

[72] P.G. Polson, C. Lewis, J. Rieman, and C. Wharton, "Cognitive Walkthroughs: A Method for Theory-Based Evaluation of User Interfaces", *International Journal of Man-Machine Studies* 36, pp. 741–773, 1992.

[73] B.E. John and D.E. Kieras, "Using GOMS for user interface design and evaluation: which technique?" *ACM Trans. Comput.-Hum. Interact. 3*(4), pp. 287–319, 1996.

[74] D.E. Kieras and D.E. Meyer, "An overview of the EPIC architecture for cognition and performance with application to human-computer interaction", *Human-Comput. Interact. 12*(4), pp. 391–438. 1997.

[75] R. Schleicher and S. Trösterer, “The 'Joy-of-Use'-Button: Recording Pleasant Moments While Using a PC”, *Human-Computer Interaction – INTERACT 2009* (Vol. 5727/2009), Springer, Heidelberg, 2009.

[76] P. Hekkert, “Design Aesthetics: Principles of Pleasure in Product Design”, *Psychology Science*, Vol. 48, pp. 157–172, 2006.

[77] K. Isbister and C. Nass, “Consistency of personality in interactive characters: Verbal cues, non-verbal cues, and user characteristics”, *International Journal of Human-Computer Studies 53*(2), pp. 251–267, 2000.

[78] A. Marcus, “Principles of effective visual communication for graphical user interface design”, in *Human-computer interaction: toward the year 2000* (R. Baecker, J. Grudin, W. Buxton, S. Greenberg (eds), pp. 425–441, Morgan Kaufmann, San Francisco CA, 1995.

[79] F. Mairesse, M. Walker, M. Mehl, and R. Moore, “Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text”, *Journal of Artificial Intelligence Research* 30, 2007.

[80] H. Stelmaszweska, B. Fields, and A. Blandford, “Conceptualising User Hedonic Experience”, in, *Proc. ECCE-12, Living and Working with Technology* (D.J. Reed, G. Baxter, and M. Blythe, eds), EACE, York, pp. 83–89, 2004.

[81] N. Tractinsky, A.S. Katz, and D. Ikar, “What is beautiful is usable”, *Interacting with Computers 13*(2), pp. 127–145 (2000)

[82] S. Mahlke and G. Lindgaard, “Emotional experiences and quality perceptions of interactive products”, in *Proceedings of the 12th international Conference on Human-Computer interaction: Interaction Design and Usabilit,. Lecture Notes in Computer Science* (J.A. Jacko, ed.), Springer-Verlag, Berlin, Heidelberg, pp. 164–173, 2007.

[83] M. Hassenzahl, “Aesthetics in interactive products: Correlates and consequences of beauty”, in *Product experience* (H. N. J. Schifferstein and P. Hekkert, eds.), Elsevier, San Diego CA, pp. 287–302, 2008.

[84] T. Lavie and N. Tractinsky , “Assessing dimensions of perceived visual aesthetics of web sites”, *International Journal of Human-Computer Studies* 60, pp. 269–298, 2004.

[85] M.M. Bradley and P.J. Lang, “Measuring emotion: The Self-Assessment Manikin and the semantic differentia”, *Journal of Behavior Therapy & Experimental Psychiatry* 25, pp. 49–59, 1994.

[86] G. Huisman and M. Van Hout, "The development of a graphical emotion measurement instrument using caricatured expressions: the LEMtool", in *Emotion in HCI – Designing for People : Proceedings of the 2008 International Workshop* (C. Peter, E. Crane, M. Fabri, H. Agius, and L. Axelrod, eds.), Fraunhofer Verlag, Stuttgart, pp. 5–7, 2008.

[87] P.M.A. Desmet, "Measuring emotions: development and application of an instrument to measure emotional responses to products", in.*Funology: From Usability To Enjoyment* (M.A. Blythe, K. Overbeeke, A.F. Monk, and P.C. Wright, eds), Kluwer Academic Publishers, Norwell MA, pp. 111–123, 2004.

[88] J.A. Bargas-Avila and K. Hornbæk, "Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience", in *Proc CHI 2011*, pp. 2689–2698, 2011.

[89] N. Bevan, "Usability is Quality of Use", in *Proc. HCII 1995*, pp. 349–354, Elsevier, Amsterdam, 1995.

[90] Eurescom Project P.807 Deliverable 1, "Jupiter II – Usability, Performability and Interoperability Trials in Europe", *European Institute for Research and Strategic Studies in Telecommunications*, Heidelberg, 1998.

[91] F.D. Davis, "User Acceptance of Information Technology System Characteristics, User Perceptions and Behavioral Impacts" *International Journal of Man-Machine Studies 38*(3), pp. 475–487, 1993.

[92] I. Ajzen and M. Fishbein, *Understanding Attitudes and Predicting Social Behavior*, Englewood Cliffs, Prentice Hall, New York, 1980.