

Chapter 2

Evaluating Multimodal Systems

As seen in the previous chapter, multimodal systems are well-established—at least in the research community studying HCI. But during the development process and—at the latest—once the system is built, methods for quantitative assessment are needed. In fact, as systems are usually meant to fulfill certain needs of, assist or even replace the human, the user's perspective needs to be considered from the very start of conceiving a system. This approach of taking into account the potential user group and its characteristics has culminated in the user-centered or participatory design approach (Schuler and Namioka 1993), namely involving the user in the design process. Some accepted methods for user-centered design, such as the cognitive walkthrough, are discussed below, please refer to Vredenburg et al. (2002) for a survey of user-centered design practice.

But while established design methods exist that can be (partially) transferred to the context of multimodal systems, so far, evaluations, as a part or rounding off of the design process, have been mostly individual undertakings (Möller et al. 2010a). Furthermore, only “few commonly accepted practices and standards” exist (Gibbon et al. 2000), among those the iterative design approach or simulation studies. According to Gibbon et al. (2000) the evaluation of multimodal systems is challenging due to several reasons:

- No standard benchmark databases exist, although there are benchmarks for the evaluation of single components, such as the speech recognizer.
- It is difficult to record under normalized and reproducible conditions.
- Evaluation criteria are unclear and qualitative aspects play a significant role.
- Qualitative aspects are difficult to measure and user studies are costly.

Dybkjær et al. (2004) have concluded that “the state of the art in spoken multimodal and mobile system's usability and evaluation remains uncharted to a large extent” but that “with the technical advances and market growth in the Spoken Language Dialogue System (SLDS) field, evaluation and usability of uni-modal and multimodal SLDSs are becoming crucial issues.”

Since it is the aim of this work to contribute to the continuing research on the evaluation of multimodal interactive systems important concepts related to this are

introduced and an overview of established tools is given in this chapter. Related work in the area of evaluation in general and of multimodal interactive systems in particular is presented and discussed. A taxonomy of quality aspects aiming to structure concepts related to the evaluation of multimodal interactive systems is presented.

2.1 Evaluation Concepts

Speaking about evaluation the notion of quality might be considered as the headline under which the different concepts and aspects can be examined. Quality has been defined by Jekosch (2005) as

“The result of appraisal of the perceived composition of a unit in comparison to its desired composition.”

This involves a perception and judgement process by the user and emphasises the need to expose the user to the ‘composition’ in question—which leads, in the case of interactive systems, to the necessity of interaction tests. During or after those tests different measures can be applied that should allow an insight into the final judgement of the system by the user.

One approach to the assessment of quality is the system-centered view of developers who often presume that it is

“The collective effect of service performance which determines the degree of satisfaction of the user”

in terms of ‘Quality of Service’ (QoS) (ITU-T Rec. E.800, 1994). And while it is surely true that a reduced performance would have a negative impact on the user’s satisfaction—at least below a certain threshold—one cannot assume a strict cause-and-effect relationship. A high QoS is not sufficient to guarantee user satisfaction. In telecommunications, the term ‘Quality of Experience’ has recently been used for describing all aspects, including and beyond QoS, which finally result in the acceptability of the service (ITU-T Rec. P.10, 2007).

In the field of human–computer interaction, the focus has been for a long time on a system’s usability (ISO Standard 9241–Part 11, 1999):

“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.”

Effectiveness describes the accuracy and completeness with which specified users can reach specified goals in particular environments, while efficiency refers to the effort and resources required in relation to the accuracy and completeness achieved (ISO Standard 9241–Part 11, 1999). Satisfaction is defined as “the freedom from discomfort, and positive attitude to the use of the product” (ISO Standard 9241–Part 11, 1999). The context of use finally takes up the “characteristics of the

users, tasks and the organizational and physical environments” (ISO Standard 9241–Part 11, 1999).

According to Nielsen, learnability—the capability of a system to enable the user to learn how to use it (ISO/IEC Standard 9126, 2001) [or ease with which users can start effective interaction (Dix et al. 2003)] – is also an important aspect of usability (Nielsen 1994). Lately, the term intuitiveness has come into focus: the extent to which the user is able to interact with a technical system effectively by applying knowledge unconsciously (Mohs et al. 2006).

Davis (1989) coined the expression ‘ease of use’, describing the degree to which users assume that the usage of a system will be without effort. In 2004 Nielsen expressed his hopes that the notion of ‘ease of use’ will be joined by the concept of ‘joy of use’, thus promoting an attention shift towards the pleasantness of interactions (Nielsen 2004) similar to the extension of QoS to QoE.

Recently, ‘User eXperience’ (UX) appears to be the new catchword. In (ISO DIS Standard 9241–Part 210:2010, 2010) UX has been defined as

“A person’s perceptions and responses that result from the use or anticipated use of a product, system or service.”

The relation between usability and user experience has been discussed in detail by Bevan (2009) and Law et al. (2009) among others.

As has become apparent in the section above there are numerous aspects of quality that might be the target of an evaluation. Furthermore, different terms are often used for the same construct, and they are measured using the same metrics. In Möller et al. (2010a) a taxonomy of quality aspects of multimodal human–machine interaction has been proposed in order to better understand and differentiate between these general constructs currently used when speaking about assessment or evaluation. The taxonomy is the joint work of different researchers, among them the author of this book.

The taxonomy consists of three layers:

1. Quality factors related to the user, the system and the context that have an impact on interaction behavior and thus on perceived quality.
2. Interaction performance aspects describing user and system performance and behavior.
3. Quality aspects related to quality perception and judgment. The color gradient applied to the third layer indicates the differentiation of hedonic and pragmatic aspects.

The third layer can be taken as QoE, while both, the quality factors as well as the interaction performance aspects relate to QoS (Möller et al. 2010c).

Satisfaction is not named explicitly in the taxonomy but is addressed by the joy-of-use component of usability—as is UX, which is not listed either (see Fig. 2.1). The wrapping-up of UX in joy-of-use is analogous to the interpretation of UX as “an elaboration of the satisfaction component of usability” discussed in Bevan (2009). Joy-of-use specifically addresses those aspects of a user interface that appeal to a

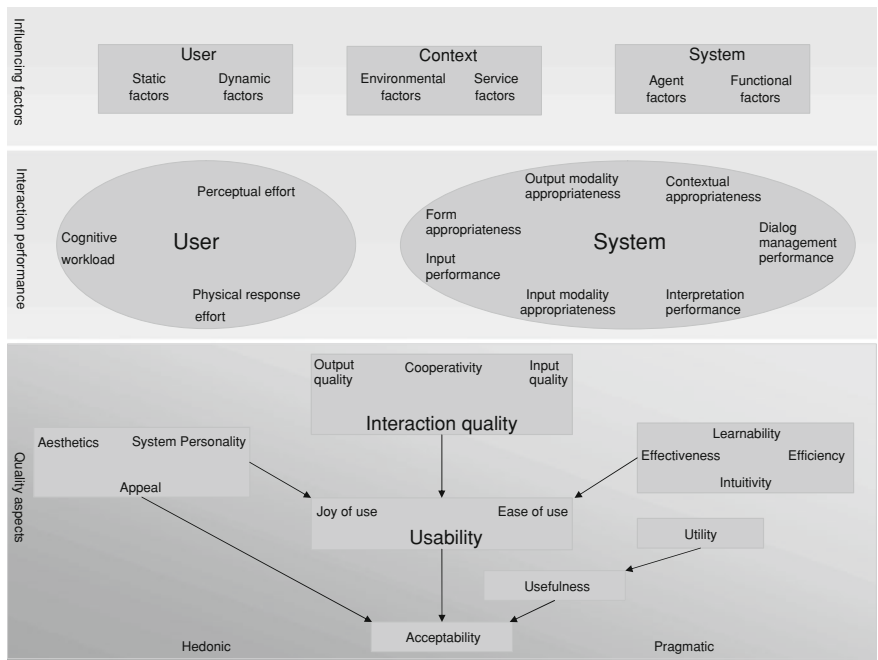


Fig. 2.1 Taxonomy of quality aspects of multimodal human–machine interaction according to Möller et al. (2010a)

person’s desire of pleasure—aspects that are fun, original, interesting, engaging, and cool.

According to the authors the taxonomy may serve at least three different purposes. System developers may search for the interaction performance and quality aspects they are interested in and find the appropriate evaluation metrics in the paper. The taxonomy could serve as basis for systematic efforts to collect evaluation data. And the constructs and influencing factors, once identified, can serve as targets for an automatic or semiautomatic evaluation.

Implications for this book—I

As stated in Chap. 1 the research questions can be roughly organized based on the structure of the multimodal system. But each contribution reported here can be localized as well in the taxonomy. Thus, in this work the different layers and boxes of the taxonomy will be addressed. Interaction performance aspects related to user behaviour and system performance are proposed and evaluated concerning their applicability to describe and predict perceived quality in Chap. 4. The quantification of output and input quality will be described in Chaps. 5 and 6, respectively. And quality aspects, related mostly to interaction quality and its predecessors input and output

quality, are examined concerning their interrelations in Chap. 7. This work can thus be seen as a first comprehensive application of the taxonomy of quality aspects of multimodal systems.

The next section will give an overview of the most important sets of tools for the evaluation of multimodal systems necessary as prerequisites for the tasks described.

2.2 Evaluation Methods

According to Gibbon et al. (2000) evaluation can be divided into evaluation on the component- or system-level. Component-level evaluation would lead to a certain QoS. Information on component-level evaluation can be found, for example, in Cole et al. (1997) and López-Cózar Delgado and Araki (2005) but will not be discussed in detail here. As explained above, a high QoS is not sufficient to guarantee user satisfaction and finally acceptance. Therefore, an evaluation of the overall system in terms of system-level evaluation is necessary. As these evaluation techniques usually involve user tests, either as direct evaluation or during data collection, they are costly. Nevertheless, the approach most commonly applied and often most effective is the user-centred diagnostic evaluation on the system-level (Sweeney et al. 1993). System-level evaluation techniques have been classified by Balbo et al. (1993) as being either predictive or experimental. This approach has been expanded by Gibbon et al. (2000) to also include expert evaluation. The structure of this section follows this classification.

2.2.1 *Predictive Evaluation*

User behaviour and performance variables can be predicted based on empirical observations or a theoretical model. The model is usually based on a detailed description of the proposed design and a task analysis. These methods can be applied early in the design phase as they do not require a system implementation. As usability predictions can be obtained from calculations or simulations once the model is built, variations of the design can be explored by making changes to the model. Thus, revise-and-evaluate-iterations can be accomplished quickly. On the other hand, data collection and the development of models necessary for the prediction can be just as time-consuming as user tests. Furthermore, a final user test is required to cover aspects of usability not addressed by the model and to ensure that no critical issues have been overlooked.

According to Kieras (2003) there are currently three main approaches to model-based evaluation: task network models, cognitive architecture models and GOMS models. The first are based on a network of processes with assigned completion times. Workload and resource parameters can be attached to the processes and performance predictions are obtained by Monte-Carlo simulations. Cognitive architecture

models consist of a set of hypothetical interacting perceptual, cognitive and motor components of the human. Thus, a simulated human interacts in a simulated task environment. These representations of theories on human psychological functions are primarily used in basic research projects. GOMS (Goals, Operators, Methods, and Selection rules) models represent the knowledge of procedures a user must dispose of to be able to operate a system. Amongst the model-based approaches GOMS models are the most widely used in interface design.

In the area of multimodal systems predictive models have been used, for example, by Mellor and Baber (1997) to model transaction time of different systems. A first step towards formalized multimodal interaction has been described by Suhm et al. (2001). The authors used the results of a user study on a multimodal dictation system to build a performance model of (recognition-based) multimodal interaction that predicts input speed including time needed for error correction. While both models seem to work fine for simple metrics such as transaction or error correction time the constraint of this kind of model-based prediction is obvious: how the predicted performance metrics relate to user satisfaction or other quality aspects is not clear.

Assuming that user satisfaction is the ultimate measure of a system's success the PARADISE framework (PARAdigm for dialogue System Evaluation) (Walker et al. 1997)—for spoken dialogue systems—tries to predict this metric from performance parameters. Those parameters are collected during user evaluations (or annotated afterwards). The most significant predictors for user satisfaction from a large set of variables are determined using multivariate regression analysis. The weighting factor denotes the respective importance of each parameter for user satisfaction. Possible predictor variables are classified as task-based success measures or cost measures. Cost measures in turn are composed of efficiency and qualitative measures. Beringer et al. (2002) proposed an adaptation of the PARADISE framework for multimodal systems: PROMISE, Procedure for Multimodal Interactive System Evaluation. Both, PARADISE and its application to multimodal interactive systems will be discussed in more detail in Chap. 4.

2.2.2 Experimental Evaluation

Although an automatic evaluation of systems based on interaction parameters is tempting, it could never completely replace experimental evaluations due to the reasons stated above. To collect data for an approach similar to PARADISE, or for a final user test complementing a model-based evaluation an experimental evaluation is necessary. Here, real users are involved, and the tasks accomplished, as well as the environment the study is carried out in, should mirror the reality the system has been designed for. Participants should represent the target group according to all characteristics of the users that could influence their interaction behavior and quality judgment. These include static (e.g., age, gender, native language) as well as dynamic characteristics (e.g., motivation, emotional status). Much of what has been widely accepted concerning usability engineering methods, such as the acquisition of

participants and the study design, can be transferred from the areas of spoken dialogue system evaluation or evaluation of graphical user interfaces. A detailed description of experimental evaluations of multimodal systems can be found in Bernsen and Dybkjær (2009).

For such an evaluation, three main approaches exist: the user study with a system prototype, a simulation study and iterative design or rapid prototyping. During a user study quantitative measures as well as qualitative data such as observations can be gathered and used not only for the evaluation but also to fill a database of multimodal interaction for later benchmark tests (Gibbon et al. 2000). But, as a working system has to be implemented and real users are involved, it is a time-consuming and expensive approach. Furthermore, the threshold for actually applying the findings from these evaluations and rebuilding the system accordingly is high.

To avoid at least part of the development costs it is possible to replace the system or parts of the system (e.g., the speech recognizer) with a human being, a so-called Wizard-of-Oz (WOz) while the users believe they are interacting with a fully functional system. The WOz technique has been described and discussed in detail in Dahlbäck et al. (1993) for spoken dialogue systems. An extension of the WOz mechanism to the analysis of multimodal interfaces and a set of requirements for a generic multimodal WOz platform has been presented in Salber and Coutaz (1993).

The iterative design approach (often relying on rapid prototyping) describes a (re)design, implementation and user testing cycle that allows a fast exploration of detailed implementation issues (Nielsen 1993a). This method has been applied in the development of multimodal systems, for example in the EMBASSI project (Rapp and Strube 2002). For spoken dialogue systems a rapid dialogue prototyping methodology had been described in Bui et al. (2004), later extended to multimodal dialogue systems (Ailomaa et al. 2006). Further work by Dumas and colleagues includes a rapid prototyping platform (McGee-Lennon et al. 2009) and a toolkit (Dumas et al. 2009a) as well as SMUIML, a markup language (Dumas et al. 2010).

Independent from the method used for experimental evaluation usability issues can be found by analyzing the interaction and the user perceptions assessed with questionnaires or during interviews. A methodology to assess user experience of multimodal dialogue systems (SUXES) is described by Turunen et al. (2009a). SUXES is a complete procedure, starting with an introduction to the evaluation and a background questionnaire. This is followed by an introduction to the application and the assessment of expectations of the users. Then, the user experiment is carried out, and the user experience is assessed. The questionnaires rely on a set of nine statements, related to speed, pleasantness, clearness, error free use, robustness, learning curve, naturalness, usefulness and future use. Interaction parameters are not analyzed. The method addresses the question to which degree the expectations are met by the actual experience with the system. The authors claim that the method is efficient and ‘particularly suitable for iterative development’.

In Möller et al. (2010a) established questionnaires are discussed concerning their appropriateness for assessing the multimodal quality aspects introduced above (see Fig. 2.1). It is argued that the AttrakDiff (Hassenzahl et al. 2003) and the System Usability Scale (SUS) (Brooke 1996) cover most aspects (Learnability, Effective-

ness, Efficiency, Aesthetics, System Personality and Appeal) at least partly. Further questionnaires examined are the Software Usability Measurement Inventory (SUMI) (Kirakowski and Corbett 1993) and the questionnaire for Subjective Assessment of Speech System Interfaces (SASSI) (Hone and Graham 2000). While the SASSI would have to be adapted to be used for multimodal systems, the authors do not recommend the use of the SUMI for the evaluation of multimodal systems.

Implications for this book–II

It has become apparent that there is currently no questionnaire developed specifically for multimodal systems. Thus, the problem of which questionnaire to use had to be addressed in preparation to each study analyzed in the following chapters. As will be discussed in the corresponding chapters, both, existing questionnaires currently under development at the home institution and un-validated questionnaires designed specifically for each problem were utilized.

2.2.3 Expert Evaluation

Once a prototype is built, instead of inviting users the system can also be tested by a group of experts (experienced professionals). Possible methods are the Cognitive Walkthrough and heuristic evaluation. During a Cognitive Walkthrough at least one expert follows a previously determined ‘optimal’ path to accomplish a posed task. At every step along this path the expert controls whether the next step would be obvious to a novice user. Heuristic evaluation is one of the most cost-effective methods to identify usability issues during the design process. General, interface-specific and product-specific guidelines have been proposed, for example for GUIs (Nielsen 1993b), spoken dialogue systems (Cohen et al. 2004; Bernsen and Dybkjær 2000) or software (ISO Standard 9241–Part 110, 2006). Those lists are used by a group of experts to identify and classify usability problems. Thus, a heuristic evaluation yields not only a list of problems but also indications of how to solve each problem.

According to Gibbon et al. (2000) at least three experts are necessary to identify about half of the usability problems. The more experts are involved, the more problems are found but the more expensive is the evaluation. Efficiency can be improved by assigning not only usability experts but also domain, and usability-domain experts.

A drawback of these methods is that they are well-suited to identify problems but less appropriate for a quantification of system quality—necessary, for example, for a comparison of systems or system versions.

2.3 Summary

In this chapter the theoretical foundations for the evaluation of multimodal systems have been briefly summarized. The concepts of different quality aspects crucial to this work have been introduced. The taxonomy of multimodal quality describing the interrelation between those concepts will be used in the following chapters to further structure, analyze and discuss the research questions and the answers found. It has become apparent that many open questions remain and that the problems identified in the early 2000s have not been solved completely. There is, for example, no validated questionnaire assessing the quality of multimodal systems in existence yet. Depending on the subject addressed, different ways will be described to circumvent this problem where possible.

As it is the aim of this work to point out possible approaches to quantifying quality aspects of multimodal interactive systems, most results presented in the following have been achieved relying on experimental evaluation methods, such as user studies.



<http://www.springer.com/978-3-642-29601-7>

Quantifying Quality Aspects of Multimodal Interactive
Systems

Kühnel, C.

2012, XVI, 188 p., Hardcover

ISBN: 978-3-642-29601-7