

## 2 What Are Multimodal Systems? Why Do They Need Evaluation? - Theoretical Background

The following section introduces relevant concepts and definitions. After that, the cognitive foundations of multimodal interaction are briefly described. Next current evaluation methods are introduced and reviewed with respect to their appropriateness regarding multimodal systems.

### 2.1 Modality, Media and Multimodal Systems - Definitions and Terminology

This sub-section will discuss several conceptualisations of the term *modality*. On this basis a definition of multimodal systems is adopted. This definition is used throughout the remainder of this thesis.

Several definitions of modality can be found in the relevant literature. These can be broadly categorized into three general groups: Physiological/human-centred definitions, technology/system-centred definitions, and definitions incorporating both views.

The term modality has its origins in physiology and is, according to Charwat (1992), defined as

"perception via one of the three perceptual channels. You can distinguish the three modalities: visual, auditive, and tactile (physiology of senses)."

Thus, the three senses sight, hearing, and touch, correspond to the three perceptual channels. Thereby the terms visual and auditive refer to the perception and the sensory modalities; the terms optical and acoustical refer to physical (and not physiological) parameters (Schomaker et al., 1995). According to Charwat's (1992) definition only three different modalities respectively three different human senses can be distinguished. Although the aforementioned senses are nowadays those with the highest relevance for human-computer-interaction (HCI), at least three more senses (smell, vestibular, taste) are defined in physiology.

Another definition of modality is offered by Bernsen (2008):

"Modality is a way of representing information in some physical medium. Thus, a modality is defined by its physical medium and its particular "way" of representation."

With this definition, Bernsen (2008) moves away from the physiological understanding of modality: Modalities according to Bernsen (2008) refer to *ways of information representation* rather than to the human sense. Thus, he broadens the term modality: Humans use many ways to represent information and these different ways of information representation may refer to the same sensory modality (e.g. images and text are different ways of representation but both refer to vision). A multimodal system in Bernsen's sense is a system employing at least two different modalities ("ways of information representation") for input and/or output. In contrast, Bernsen defines a

unimodal system as a system using the same modality for input and output. This conceptualization of modalities and multimodal systems is insofar rather unconventional, as devices offering a traditional graphical user interface (GUI) only, are multimodal since they offer texts and graphic as output and haptics as input. An example of a unimodal system in Bernsen's sense is a system offering only spoken language as input and output.

Another system-oriented conceptualisation of modality is presented by Nigay and Coutaz (1995). They too expand the strict physiological view but unlike Bernsen, they focus on the *interaction technique* rather than on the way of information representation. They posit a modality as the combination of an *interaction language* ( $L$ ) and a *physical input or output device* ( $d$ ), which can be formalized as a tuple  $\langle d, L \rangle$ . Examples for interaction languages, as proposed by Nigay and Coutaz (1995), are direct manipulation, gestures, or pseudo natural speech. Thus interaction modalities on a smart-phone could be  $\langle \text{touchscreen, gestures} \rangle$  or  $\langle \text{microphone, speech} \rangle$ .

Yet another definition is given by Oviatt (2002) defining a multimodal system as systems processing

“two or more combined user input modes— such as speech, pen, touch, manual gestures, gaze, and head and body movements— in a coordinated manner with multimedia system output. [...] Such interfaces eventually will interpret continuous input from a large number of different visual, auditory, and tactile input modes [...]”.

Thus, in Oviatt's definition modalities refer to *input modes*, in contrast to the above mentioned definitions which explicitly refer to the term modality. However, as no guidance is given on what defines an input mode, this definition is rather unhelpful but lends itself as a good example for a strict technology-driven definition of multimodality, which is not expanding the strict physiological view but neglecting it. According to Baber (2001), only the combination of the technical system-oriented view, (which focuses on interaction techniques, input device and output devices) and the user-oriented view (which focuses on human perception), will be useful to investigate multimodal human machine interaction. Typically, a multimodal system employs different interaction techniques and a user needs to have different sensory modalities to interact with such a systems. However, most of the definitions, as those presented exemplarily above, focus either on one or the other perspective.

Möller et al. (2009) take both views into account, stating that

“multimodal dialogue systems are systems which enable human-machine interaction through a number of media, making use of different sensory channels.”

The understanding of the term *media* in the scientific community is, in contrast to the term modality, mostly uniform. Media is associated with the physical realization respectively presentation of information via input and output devices (cf e.g. Bernsen, 1997; Gibbon, Moore, & Winski, 1998; Hovy & Arens, 1990; Jokinen & Raike, 2003; Sturm, 2005).

Möller et al. (2009) elaborate their definition further as follows:

“These channels may be used sequentially or in parallel, and they may provide complementary or redundant information to the user.”

With this part of the definition, Möller et al. (2009) refer to Nigay and Coutaz (1993) who developed a design space for multimodal systems along the three dimensions (1) *level of abstraction*, (2) *usage of modalities* and (3) *fusion*. Level of abstraction represents the technical level, on which information of the different input and output devices is processed. Speech input may be processed as a signal, a sequence of phonemes or as parsed sentences bearing meaning. Usage of modalities means the temporal availability of the modalities: While some systems allow for parallel usage, other systems only offer sequential interaction. The third dimension, fusion, describes if and how the information of the different modalities is combined. Based on the design space Nigay and Coutaz (1993) identified four different types of multimodal systems:

- **Exclusive interactions:** The system offers different modalities but usage is only sequential (one modality at one time). Fusion is absent.
- **Alternate interactions:** The system offers different modalities. Like for exclusive interactions, the modalities can only be used sequentially but they can be related to each other. Fusion, the combination of the possible input data, is implemented.
- **Simultaneous interactions:** The modalities can be used in parallel (simultaneously). Fusion is absent.
- **Synergistic interactions:** The modalities can be used in parallel and the information can be related to each other.

In the following the term *multimodal system* is used as defined by Möller et al. (2009). The systems used as material in the presented studies offered just sequential input and rudimentary fusion modules, as a result for the most part only exclusive interactions were possible.

## 2.2 Cognitive Foundations of Multimodal Interaction and Assumed Advantages of Multimodal Systems

By providing multiple communication channels, multimodal systems are assumed to support human information processing by using different cognitive resources. This assumption is largely based on cognitive theories postulating multiple, modality-specific processing resources (e.g.; Baddeley & Hitch, 1974; Baddely, 2003; Paivio, 1986; Wickens, 1984, 2002).

According to the working memory theory proposed by Baddeley, different types of information refer to different cognitive resources. Baddeley's working memory model includes three components: the *central executive*, the *visual-spatial sketchpad* and the

*phonological loop*. Hence, the short-term storage of visual-spatial information (e.g. colors and shape) is the visual-spatial sketchpad, whereas the phonological loop is the short term storage for auditory-verbal data. It has to be noted, that visual information (e.g. written text) might be recoded into verbal information by sub-vocal articulation and consequently be stored in the phonological loop. The central executive is the controlling unit, monitoring and where required adjusting thinking processes and actions. Later, a fourth component, the *episodic buffer*, was added. The *episodic buffer* is a multimodal component and represents an interface to the long-term memory.

Also, regarding the long term memory, the coding of information into mental representations is assumed to be modality specific to a large extent. *Dual coding theory* by Paivio (1986) postulates two largely independent cognitive systems: the *imaginal, non-verbal systems* and the *verbal systems*. As Baddeley's phonological loop, the verbal system processes verbal information whereas the imaginal system, analogue to the visual-spatial sketchpad, processes visual-spatial information. According to Paivio (1986), findings of neuro-psychology support these assumptions. It was shown, that dependent on the type of information (verbal vs. spatial) different brain areas are active. Still, these systems are connected. This explains why multimodal presentations, e.g. verbal-auditory paired with visual information, can be superior to unimodal presentations. The dual coding leads to higher recognition and recall performance.

A third modality specific theory is the *Multiple Resource Theory (MRT)* by Wickens (2002, cf. Section 3.2.1). This theory proposes three different *processing stages*, two *different response codes*, two different *perceptual modalities* and two different *input codes*. Moreover, two *visual channels* are suggested. In contrast to the theories mentioned above, Wickens does not only differentiate between the information processing modalities but also between sensory modalities. For each of the different stages, modalities and response codes, different cognitive resources are assumed. MRT predicts that tasks accessing the same resources are very difficult to be performed in parallel. Or the other way around: Timesharing, the splitting of attention between two tasks, is easier when the necessary information is presented via two modalities instead of one modality.

In summary the redundant information presentation and the splitting of information to several channels reduces the overall cognitive load experienced by the user. With lower cognitive load, errors are less likely and the interaction gets more robust (Qvarfort, 2004).

Additionally to a more robust interaction, multimodality may also enhance a system's flexibility, its naturalness, and its efficiency (Hedecke, 2000; Höllerer, 2002; Oviatt, 1999; Qvarfort, 2004). With a higher degree of freedom the user is free to choose his/her preferred interaction modality with regards to situation, task and context. This higher flexibility is assumed to increase the systems inclusiveness (Jokinen & Raïke, 2003) and efficiency; however, regarding the latter also the opposite is reported (Sturm, 2005). The hypothesis regarding naturalness stems from the observa-

tion of human-human communication being multimodal, e.g. verbal human communication has a visual part through lip-reading, mimicry or gestures (Schomaker et al., 1995). Due to the possibilities to use richer natural languages and new flexible ways of interaction, multimodality has the potential to realize the *system-as-an-agent* metaphor proposed by Jokinen (2009). Jokinen describes such agents as interaction partners mediating between the user and the application, rather than as a tool that is used to perform certain tasks. Consequently, the gulf between user and system can be minimized by adapting the system to the user's natural characteristics (Norman, 1986).

While empirical findings support the above assumptions - multimodal systems have indeed been shown to be more natural, more efficient, more reliable and more robust (e.g. Oviatt, 1996; Oviatt et al., 2000, Cohen, McGee, & Clow, 2000; Burke et al., 2006) - it has to be noted that these benefits are not an inherent property of all multimodal systems. Oviatt (1999) points out that all these advantages are mediated through the design of the interface and the usage context; multimodality was shown to be especially beneficial in situation with high workload and high task complexity (Oviatt, Coulston & Lunsford, 2004). Moreover, considering the above presented MRT (Wickens, 2000) a multimodal system can also be inferior compared to a unimodal system, e.g. if additional speech input is necessary in verbally demanding situations like in air traffic control or call centres. In addition, a higher cognitive load due to more degrees of freedom may occur (Schomaker et al., 1995). Furthermore, the different modalities may interfere with each other (Schomaker et al., 1995): When presenting identical information via two modalities (e.g. reading and listening to the same text simultaneously) a synchronization problem can arise (Schnotz, Bannert, & Seufert, 2002). Additionally, if different modalities refer to the same cognitive resources, task performance may even decrease (Oviatt, 1996). Thus, it is not surprising that it has been shown, that making a system multimodal by just adding a further modality to a unimodal system may not necessarily lead to an improvement (Oviatt, 1999). Consequently, evaluation of the interface is an indispensable issue.

### 2.3 Quality and Usability Evaluation Methods

The following section will give an overview of established and well-known evaluation methods. Advantages and disadvantages regarding their appropriateness and suitability for multimodal systems will be reviewed. Please note, that in the literature all the described methods are often labelled as *usability* evaluation methods rather than as *quality* evaluation methods. While those constructs are partly overlapping, they are not identical. Hence, before introducing the evaluation methods, the relationship between quality and usability will be discussed.

### 2.3.1 Quality vs. Usability

Most of the described methods were developed to measure usability in the narrow sense as described by the International Organization for Standardization (ISO) with the standard 9241-11 (ISO 9241-11, 1998). Here usability is defined by the three factors efficiency, effectiveness, and satisfaction. While user satisfaction is mentioned in this standard, the focus of early usability evaluation was focused on the efficiency and effectiveness of the system. For instance in a meta-analysis by Hornbæk and Law (2007), user satisfaction was found to be the usability factor which was assessed least frequently. However, in this thesis the user-experienced satisfaction is considered an essential part of usability (cf. Section 3.3). Usability itself is one important, but not the only aspect of quality. Quality, as understood in this thesis, is the result of the user's appraisal of the perceived capabilities of the system to support the user's individual goals.

The first part of this conceptualisation is based on the definition of Jekosch (2000, as cited in Möller, 2005). It implies that quality is an inherently "subjective" concept; it is a result of the user's individual perceptual and judgemental processes. Please note that, Jekosch (2000) original definition suggests that users appraise the perceived entity in comparison with a desired entity. This part of the definition was not adopted as it implies a rather resource-extensive cognitive process involving mental comparison of the features of the perceived and the desired system. Findings from cognitive psychology imply that the brain is rather lazy and avoids resource-intensive processing: Judgements are often biased and are based on heuristics or on intuition (Kahneman, 2003). Moreover, the original definition indicates that the user knows how the desired system should be. This is also debatable: While users may know their goals, they may not know how exactly a systems needs to be designed in order to fulfil those goals. Thus, the definition by Jekosch may apply to the quality of speech and voice signals, the context were it has been developed for, but not for rather complex multimodal systems.

Hence, the second part is based on Hassenzahl's work (Hassenzahl & Roto, 2007). Here, quality is the related to the fulfilment individual goals. Those goals can be either *do-goals*, for example "making a phone call", or *be-goals*, "being related to somebody". Do-goals are derived from the higher-level be-goals (Hassenzahl & Roto, 2007). For example, missing somebody may lead to the desire to communicate this person. Making the phone call is than the do-goal, the feeling of being related to this person is the be-goal

### 2.3.2 Evaluation Methods

With usability being understood as one of many quality aspects, usability evaluations can be considered as a subgroup of quality evaluations.

According to Preece et al. (1994) evaluation methods can be distinguished using five different criteria. The first distinction refers to the question addressed with the evaluation and comprises four different categories. Evaluation studies may be conducted,

- to see if the system is good enough,
- to compare two alternative and see if one system is better than another one,
- to get the system closer to the real world,
- to see how well the system is working in the real world or,
- to see if the systems complies to certain standards.

Depending on the question addressed with the evaluation, it has to be decided in which stage of the development cycle the evaluation should take place. Here, *formative, process-oriented evaluation* can be distinguished from *summative, goal-oriented evaluation*. Formative evaluation can already take place in early development cycles without a prototype and aims to improve the system as part of the iterative design process. For summative evaluation, an advanced prototype is necessary, as summative evaluation is typically carried out to assess the quality of a late version of the system.

Another distinction criterion is the level of user involvement with *user-centred, empirical methods* on the one side, and *expert-centred, analytical-formal methods* on the other side. Especially formative evaluations are often conducted in early phases of the system development without users. Elements of the interface and their consequences are analysed and modelled by experts. Consequently, neither users nor a running prototype are necessary. These methods are often also labelled as *inspections* (Holzinger, 2005). User-centred, empirical methods are methods, which are observing and “measuring” user’s reactions towards the interface. Measurements collected in this manner, are assumed to represent the system’s quality. A prototype, with which the user can interact, is at least to a certain extent necessary (Sturm, 2005). Another term for such methods is *testing* (Holzinger, 2005).

The type of data collected can be *qualitative* and *quantitative data*. Quantitative data are numerical, abstract data. Abstract means that such numbers do not directly represents the meaning of the measured date (Witt, 2001). Typically, this kind of data is analysed using statistical methods. Examples are questionnaires or time measurements. Qualitative data cannot be quantified in numbers, and its analysis is usually interpretative as applying statistical methods is not possible. However, it is often possible to transform qualitative data into quantitative data. For instance, free text answers are usually qualitative data, but they can be converted into quantitative data by first analysing them with respect to the opinion stated in the text. Then the number of positive, negative and neutral answers can be counted, the counts are forming a quantitative data set which can be analysed statistically.

In the context of usability and quality evaluation, *direct* and *indirect* measurements can be distinguished (Seebode et al., 2009, Möller et al., 2009). Direct measurements are assessed directly from the user and are a direct representation of the quality as perceived by the user. Indirect measurements refers to interaction parameters or psychophysiological parameters, these kinds of measurements cannot be interpreted as direct assessments of perceived quality or perceived usability – in the best case such data is correlated with the quality perceptions but might as well be unrelated to the user's judgement (c.f. Hornbæk & Law, 2007; Naumann & Wechsung, 2008). Additional characteristics to categorize evaluation methods are according to Dix et al. (1993):

- **The style of the evaluation.** It refers to the setting, laboratory or field of the study. While lab studies offer a more controlled setting eliminating interfering variables, field studies lead to a higher naturalness.
- **The level of information.** It describes how abstract the gathered information is. Low level information is very specific, e.g. if the wording of a specific prompt is understandable. High level information is more general for instance if the system is usable.
- **The immediacy of the answer.** It describes whether the data is assessed during or after the interaction, the latter possibly being influenced by memory biases.
- **The intrusiveness of the answer.** This characteristic is directly related to immediacy, as asking questions during the interaction is rather intrusive and might affect the user's behaviour.
- **The resources required.** Resources comprise factors like time, money, effort, equipment, and manpower.

In Table 2.1 (adapted from Dix et al., 1993) established methods are categorized, based partly on the categories by Dix et al. (1993) and on additional own categories like advantages and disadvantages and appropriateness for different modalities.

In the following, established methods are described and discussed regarding their suitability for the evaluation of multimodal systems. At first, expert-centred analytical-formal methods are presented, followed by user-centred empirical methods.



Table 2.1. Classification of different evaluation methods

	Cognitive Walkthrough	Heuristic Evaluation	Model-Based Evaluation	Experiment	Interviews, Questionnaires	Protocol Analysis, Thinking Aloud (TA)
Stage of development process	Throughout	Throughout	Design	Throughout	Throughout	After prototyping
User involvement	No	No	No	Yes	Yes	Yes
Style	Laboratory	Laboratory	Laboratory	Laboratory	Laboratory/field	Laboratory/field
Level of information	Low	High	Low	Low and high	High	Low and high
Data type	Qualitative	Qualitative	Quantitative	Qualitative /quantitative	Qualitative /quantitative	Qualitative
Is method obtrusive?	No	No	No	Yes	No	Yes
Amount of Time required	Medium	Low	Medium	High	Low	High
Material resources	Low	Low	Low	High	Low	Low to high
Expertise	High	Medium	High	High	Low	Medium to high
Task-oriented?	Yes	No	Typically yes	Yes and no	Yes and no	Yes and no
Appropriateness for different modalities	Auditory, visual and haptic/ limited for multimodal systems	Auditory, visual and haptic/ limited for multimodal systems	Auditory, visual and haptic/ limited for multimodal systems	Auditory, visual and haptic and multimodal systems	Auditory, visual and haptic and multimodal	Auditory visual and haptic/ limited for multimodal systems TA is not useful for auditory modality
Main disadvantages	Only „assumed“ problems, high expertise necessary	Only „assumed“ problems, high expertise necessary, appropriate heuristics are often not available	High expertise	High expertise necessary	Development of questionnaires is very resource-consuming	TA is often difficult for users and not appropriate for speech input Working prototype is necessary
Main advantages	Low resources required	Low resources required	Low resources required	Data is generalizable	Easy to use	Data is generated by real users

### *Cognitive Walkthrough*

The *Cognitive Walkthrough* is a task-based, expert-centred, analytical method (Holzinger, 2005) based on explorative learning and problem solving theory (Wharton et al., 1994). It takes into account that user often learn the interface by exploring it, instead of reading the manual.

Experts, usually designers or psychologist, analyse the functionalities of the interface based on a description of the system, a description of the tasks the end user will carry out, a list of actions necessary to perform the tasks, and a description of user and usage context (Wharton, et al. 1994). Critical information is recorded by the experts using a standardized protocol. The procedure itself involves the following five steps (Wharton et al., 1994):

- Definition of inputs for the walkthrough (e.g. identifying the users, defining the tasks to evaluate, describing the interface in detail)
- Calling in the experts
- Analysing the action sequences for each task
- Protocolling critical information
- Revising the interface

The biggest advantages of the Cognitive Walkthrough are, as for almost all formative-analytical methods, that end users as well as an implemented system are not necessary. Disadvantages are the quite low level of information, only the ease of learning is investigated (Wharton et al., 1994). Moreover, a Cognitive Walkthrough might be very time consuming for complex systems. As multimodal systems are usually more complex than unimodal systems, due to more degrees of freedom offered by multiple modalities, the Cognitive Walkthrough, in its classical form, is rather unattractive for such systems. Moreover, the Cognitive Walkthrough is strictly task-based and will only be able to evaluate the ease-of-use of an interface rather than its joy-of-use.

### *Heuristic Evaluation*

The term “heuristic” is derived from the Greek “heureskein“ and means “to find” or “to explore” something (Holzinger, 2005). *Heuristic Evaluation* is a method of the so-called *Discount Usability Engineering*, a resource conserving, pragmatic approach proposed by Nielsen, aiming to overcome the argumentation that usability evaluation is too expensive, too difficult and too time consuming.

In a Heuristic Evaluation, several experts check if the interface complies with certain usability principles (heuristics). To ensure an independent, unbiased judgement of every evaluator, they do not communicate to find an aggregated judgement until each of them investigated the interface on his/her own (Nielsen, 1994). Result of a Heuristic Evaluation is a list of usability problems and the respective explanations. Addi-

tionally, problems might be judged according to their frequency and pertinence. According to Nielsen, three to five experts will find 60-70 % of the problems, with no improvements for more than ten evaluators (Nielsen, 1994). However, this statement has repeatedly caused disputes; research provides support (Virzi, 1992) as well as contrary findings (Woolrych & Cockton, 1992; Spool & Schroeder, 2001).

The Heuristic Evaluation is a cheap and quick to apply method and can be conducted throughout the whole development cycle (Holzinger, 2005). However, to the authors' knowledge, established usability heuristics tailored to multimodal systems do not exist.

### *Review-Based Evaluation*

For a *Review-Based Evaluation*, existing experimental findings and principles are employed to provide a judgment. Relevant literature has to be analysed in order to approve or disapprove the design of the interface (Dix et al., 1993). Hereby, the context of the respective studies has to be carefully considered. To prevent a confirmatory bias not only the similarities, but also the dissimilarities between the interface to be evaluated and the studies serving as a basis for the evaluation have to be taken into account (Gerhard, 2003).

Review-Based Evaluation is faster and more economical than conducting an own experiment. But wrong conclusions might be drawn if the selection of the considered studies is not done with the required prudence. Additionally, the vast majority of studies are addressing very specific problems, making it difficult to generalize the results to other interfaces and vice versa, a specific interface is difficult to evaluate with the results of another specific interface (Gerhard, 2003).

### *Model-Based Evaluation*

For *Model-Based Evaluation*, on a very general level, two different approaches can be distinguished. The first approach has its origin in cognitive psychology and focuses on the cognitive process while interacting with an interface; the other approach is rooted in the engineering domains and is focusing on the prediction of user behaviour patterns. Within both approaches, user models are employed for the predictions.

Methods of the first approach are usually addressing low-level parameters like task execution time, memory processes or cognitive load (cf. Engelbrecht, Quade, Möller, 2009) and are largely bottom-up oriented. Starting point to define user models are theories and findings from cognitive psychology. Examples are the methods GOMS (Goals, Operator, Methods, Selection rules; Card, Newell & Moran, 1983), the Cognitive Complexity Theory (CCT) by Kieras and Polson (1985), or ACT-R (Adaptive Control of Thought-Rational) by Anderson and his group. (e.g. Anderson & Lebiere, 1998).

With the method GOMS, the interaction with a system is reduced to basic elements, which are *goals*, *methods*, *operators* and *selection rules*. Goals are descriptions

of the goals or sub-goals of the user, what he/she intends while using the system. Operators are the actions offered by the system to accomplish these goals. Methods are well-learned sequences of sub-goals and operators suitable to achieve a goal (John & Kieras 1996). Selection 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 (John & Kieras, 1996).

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, because of this the definition of multiple selection rules is required. The EPIC framework by (Kieras & Meyer, 1997) 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 (Kieras & Meyer, 1997).

As all these cognitive models are grounded well in theory, they provide useful insights in user behaviour. Although cognitive modelling is an active research field, so far it has not been received particularly well by usability practitioners and only rarely finds its way into non-academic evaluations (Engelbrecht, Quade, Möller, 2009; Kieras, 2003). Reasons are their often-high complexity (Kieras, 2003) and possibly the aforementioned low level of the information possible to gain with cognitive modelling.

Thus the engineering-based, statistically-driven approach attempts to provide more high level information, e.g. if the user is "satisfied" with the system, and therefore rather utilizes top-down strategies. Here, user models are usually defined based on real user data and are not necessarily linked to cognitive theories (cf. Engelbrecht, Quade, Möller, 2009). Most of these methods and algorithms were developed for spoken dialogue systems, with PARADISE (Paradigm for Dialogue System Evaluation; Walker et al., 1997) likely being the most widespread one. PARADISE uses linear regression to predict user satisfaction based on interaction parameters such as task success or task duration. Other approaches are the MeMo (Mental Models) workbench using a probabilistic model of user behaviour, which includes a rule engine derived from empirical data in order to predict user behaviour (Engelbrecht, 2012). But like cognitive models, the engineering models are rather of academic interest than a widespread usability evaluation method amongst practitioners. Even though Model-Based Evaluation, like all the expert-centred methods above, can be conducted in very early design stages and is cheaper than testing with real users. The main disadvantages might be the very high expertise necessary (Holzinger, 2005) and, regarding multimodal interaction, the lack of theories to use for Model-Based Evaluation, as these processes are not well understood so far. For instance, the factors deter-

mining why users choose one modality over another have just recently been identified (cf. Chapter 6).

A model-based approach explicitly tailored to multimodal system is the PROMISE framework by Beringer and colleagues (2002), an extension of Walker's PARADISE. However, studies applying PROMISE are very seldom, possibly because some of the parameters are relatively ill defined (e.g. the way of interaction), and it is not specified how they should be assessed. Just recently (Kühnel, 2012) proposed a well-defined set of interaction parameters for multimodal interaction yielding reasonable prediction performance (>50% accuracy) for user judgements.

### *Protocol Analyses and Thinking Aloud*

For *Protocol Analyses*, user behaviour is captured using video, audio and log-files. A prominent method is *Thinking Aloud*. Here participants are asked to verbalize and loudly utter their thoughts (Holzinger, 2005). This might be done during the interaction or after the interaction as retrospective Thinking Aloud. For the latter, the user is confronted with video recordings of the test session and is asked to comment on them. Although retrospective Thinking Aloud is less intrusive than online Thinking Aloud, it might probably be affected by memory biases. Another version is the plural Thinking Aloud involving multiple participants using the system together. Therewith, the unnaturalness of this method should be reduced. Hackman and Biers (1992) confirmed that Thinking Aloud in double-teams is beneficial, yielding better results compared to single user Thinking Aloud.

This method can be used for free exploration of the interface as well as for conducting concrete tasks. Though it is often perceived as unnatural and confusing (Lin, Choong & Salvendy, 1997) and often the experimenter has to repeatedly advise the participants to actually think aloud as the constant verbalising can be difficult and effortful for the users (Hegner, 2003). Further problems are the systematic biases due to social desirability. Moreover, for interfaces offering speech input, the non-retrospective version of this method is inappropriate, as Thinking Aloud and speaking to the system simultaneously is not possible (Hegner, 2003).

### *Experiments*

Experimental evaluation investigates specific aspects of the interaction behaviour under controlled conditions (Sturm, 2005). In the simplest experimental design, one hypothesis is formulated and two experimental conditions, differing only regarding the manipulated factor to be investigated (independent variable), are set-up (Dix, et al. 1993). All differences occurring in the measured variables (dependent variable) are attributed to the manipulations of the independent variable (Dix, et al. 1993). Experiments allow collecting high quality data as interfering variables are controlled and/or eliminated. Experiments provide, if carried out carefully, causal inference. Thus, experiments are essential to establish and verify theories (Hegner, 2003); accordingly,

experiments are a useful method for evaluation of multimodal interfaces. However, with experiments being strongly controlled, user behaviour might be rather unnatural (Sturm, 2005). In the worst case, results can be an experimental artefact. Another drawback is the high amount of resources required to set up and conduct a proper experiment.

### *Interviews and Questionnaires*

*Questionnaire and interviews* are indispensable to measure the users' judgements of the system (Holzinger, 2005) as interaction data will not necessarily reflect the users' perceptions (e.g. Naumann & Wechsung, 2008). Interviews and questionnaires are often used to assess user satisfaction, emotions or attitudes towards the system. If reliable questionnaires or interviews are available, they are relatively cheap and easy to employ. Thus, the probably most common technique applied in user-centred evaluations are questionnaires, but a standardized and validated questionnaire addressing the evaluation of multimodal systems is still not available. Even for speech-based systems the probably most common questionnaire, the Subjective Assessment of Speech Interfaces questionnaire (SASSI; Hone & Graham, 2000), still lacks final psychometric validation. Please note that the SASSI is not suitable for spoken dialogue systems, as only input but not output quality is assessed. However, the ITU-T Rec. P.851, the ITU's recommendation regarding quality evaluation of telephone-based spoken dialogue systems (ITU-T Rec. P.851, 2003), proposes an extended version of the SASSI, including items covering output quality (Möller, Engelbrecht, & Schleicher, 2008).

As standardized, well-validated questionnaires tailored to multimodal system are rare, self-made questionnaires or questionnaires developed for unimodal systems are often employed. Both approaches are problematic: Self-constructed questionnaires are usually not properly validated (Larsen, 2003a) and questionnaires developed for unimodal systems may not provide valid and reliable results for multimodal systems. A detailed comparison of relevant questionnaires is presented in Chapter 4.

A notable exception is the SUXES method presented by (Turunen et al., 2009), which is, as explicitly stated by the authors, addressing multimodal systems. SUXES aims to measure user expectation and user experience with different pre- and post-test questionnaires. Constructs measured with SUXES are speed, pleasantness, clearness, error free use, robustness, learning curve, naturalness, usefulness, and future use. It is based on the SERVQUAL method (Parasuraman, Zeithaml, & Berry, 1988), initially developed to assess perceived quality of service in service and retailing companies. Even though, the original publication of the SERVQUAL method includes psychometric validation, for SUXES no such data is available. Hence, the reliability and validity of the constructs measured with SUXES is not confirmed as the constructs measured with SUXES (see above) do not match the constructs assessed with SERVQUAL (Reliability, Assurance, Tangibles, Empathy, Responsiveness). Moreo-

ver, SERVQUAL itself has been heavily criticized on a conceptual as well as on a methodological level (Buttle, 1996; Nyeck et al. 2002). Major critique points, also highly relevant for SUXES, are related to the measurement of expectations before interacting with the system (Buttle, 1996). For example several authors (e.g. Kahneman & Miller, 1986; for a comprehensive discussion see Buttle, 1996) state, that expectations are formed after interacting with a system or service and not before. Additionally, expectations may be affected by a social desirability bias as user may want to comply with the “I have high expectations” social norm (Buttle, 1996). Moreover, asking for expectation may induce expectation, which would not have been relevant without questioning for them. Additionally, costumers tend to adapt their expectation to their actual experience. Thus, if applying SUXES experimenters need to keep in mind that asking for expectations may alter them and that participants may not have expectations before the usage of a system. Furthermore, a psychometric validation of SUXES is necessary to ensure if the constructs, which admittedly have high face validity, are actually statistically reliable and valid.

## 2.4 Chapter Summary

Although multimodal systems have been around for more than 25 years now and besides the rapidly increasing technical developments in this area, evaluation methods and design guidelines are still rare and evaluation of multimodal systems is considered as problematic (Jokinen, 2008). Even though the HCI literature provides a wide range of evaluation most of them were developed to assess unimodal graphical user interfaces and will not necessarily be useful for multimodal systems. As mentioned in the respective sections, most of the established methods presented above, are not instantly usable for multimodal evaluation.

The formal-analytical methods need theories of multimodal interaction, which are just emerging by now (Kühnel, 2012). The empirical methods lack the measurement instruments (e.g. questionnaires). Thus, it is not surprising that most studies evaluating multimodal interfaces employ empirical methods, using either self-constructed questionnaires, which are not or only little validated, or adapt standardized questionnaires, which were initially developed for GUI-based interfaces, and which are also not validated for multimodal systems (e.g. Bauckhage et al., 2002; Baillie, et al., 2002; Bernsen & Dybkjaer, 2004; Bornträger, et al., 2003; Damianos, et al., 2000; DeAngeli et al. 1998, Hemsén, 2004; Höllerer, 2002; Qvarfordt, Jönsson, & Dahlbäck, 2003; Sturm, 2005).

Consequently, the constructs measured are quite diverse and the results are hardly comparable. Thus, the first step towards a unified evaluation approach for multimodal interaction is a unified framework of quality aspects of multimodal interaction. The following chapter will present such a framework based on the work of Möller et al.

(2009) and Wechsung et al. (2012a), and identify measurements methods for each aspect.



An Evaluation Framework for Multimodal Interaction

Determining Quality Aspects and Modality Choice

Wechsung, I.

2014, XIII, 191 p. 33 illus., 19 illus. in color., Hardcover

ISBN: 978-3-319-03809-4