Int. J. Human-Computer Studies 75 (2015) 21–34

Contents lists available at ScienceDirect

Int. J. Human-Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Modeling input modality choice in mobile graphical and speech interfaces ☆

Stefan Schaffer *, Robert Schleicher, Sebastian Möller

Quality and Usability Lab, Telekom Innovation Laboratories, Technische Universität Berlin, Ernst-Reuter-Platz 7, D-10587 Berlin, Germany

ARTICLE INFO

Article history:
Received 15 October 2013
Received in revised form
9 October 2014
Accepted 19 November 2014
Communicated by P. Mulholland
Available online 28 November 2014

Keywords:
Multimodal human–computer interaction
Modality choice
User modeling
Simulation
Automatic usability evaluation

ABSTRACT

In this paper, we review three experiments with a mobile application that integrates graphical input with a touch-screen and a speech interface and develop a model for input modality choice in multimodal interaction. The model aims to enable simulation of multimodal human–computer interaction for automatic usability evaluation. The experimental results indicate that modality efficiency and input performance are important moderators of modality choice. Accordingly, we establish a utility-driven model that provides probability estimations of modality usage, based on the parameters of modality efficiency and input performance. Four variants of the model that differ in training data are fitted by means of Sequential Least Squares Programming. The analysis reveals a considerable fit regarding averaged modality usage. When applied to individual modality usage profiles, the accuracy decreases significantly. In an application example it is shown how the modality choice mechanism can be deployed for simulating interaction in the field of automatic usability evaluation. Results and possible limitations are discussed.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

1.1. State of the art mobile graphical and speech interfaces

Recent developments in human–computer interaction show that increasing numbers of dialog systems offer more than one option to enable user input. Smartphones or navigational systems often come with an additional Voice User Interface (VUI). However, the Graphical User Interface (GUI) is still the more common input modality in those systems.

Touch screens are now a standard input method for the GUI. If an additional VUI is integrated into a system where only a GUI was previously available, a multimodal dialog system (MDS) is built. Examples of such systems are Apple's iPhone that is extended with the speech interface Siri (Apple, 2011) and Android smart phones provided with "google voice search" (Franz et al., 2006). Both systems enable use of selected spoken commands for specific interactions. Employing the VUI often saves interaction steps or time. A novice user deploying this benefit, however, will not know exactly if or how a speech-based interaction is possible. Experience is needed and, if the user in not accustomed to the system, reasoning and decision-making processes increase the cognitive load (Sweller, 1988). One possible way to avoid this additional load would be to make any modality possible at any point in the interaction. Input modalities would then be processed sequentially and independently (Nigay and Coutaz, 1993). ("Sequential processing" here means that users may perform system input only consecutively regardless which modality they use. "Independent processing" means that modalities are interpreted separately and no semantic fusion is performed.) In doing so, a system integrating a GUI and VUI would provide a graphical input element like a button and a voice input for each interaction. The previously mentioned smart phone examples fit only partly into this category of systems, as they integrate speech interaction only for very specialized tasks like menu navigation or keyboard typing.

In the domain of consumer products almost no sequential independent multimodal systems (SIMS) have appeared on the market so far that allow any interaction using any modality. An unsolved issue with voice input is that direct interaction via spoken commands still do not work sufficiently well because actual automatic speech recognition (ASR) modules cause too many errors, e.g., by processing extraneous background noise as user input. Accordingly, push-to-talk or "open microphone in combination with key-word-spotting systems" (OKS) are used to enable voice input. However, these implementations need at least one additional step during the interaction to activate ASR.

Assuming further improvements in ASR technology during the next years, future SIMS might have no need for push-to-talk or OKS. OKS might be implemented in an effective manner minimizing

☆This paper has been recommended for acceptance by P. Mulholland.

* Corresponding author.
E-mail address: stefan.schaffer@gmail.com (S. Schaffer).

http://dx.doi.org/10.1016/j.ijhcs.2014.11.004
1071-5819/© 2014 Elsevier Ltd. All rights reserved.

interaction time and cognitive resources. With these systems, any GUI and VUI input will be entirely possible for each achievable task. Further, users will have to select input modalities in each step during the dialog with the system.

The choice of input modalities is influenced by various factors. Several studies revealed that input performance (like ASR error rate) and modality efficiency (measured in the number of turns to solve a task) are significant moderators in multimodal systems integrating a GUI and VUI (Möller et al., 2011; Schaffer et al., 2011a; Wechsung et al., 2012). Other documented influence factors are cognitive demand (Wickens and Hollands, 2000), time (Bevan, 1995), hedonic quality (Hassenzahl et al., 2003), environment, and dynamic and static user attributes (Bohn et al., 2005; Ren et al., 2000). However, even if quantitative data is collected in most of the studies, the obtained results are often qualitative in nature. But even more harmful for the purpose of modeling is the fact that, in most cases, the gathered quantitative data is not comparable due to differing experimental setups, interfaces, logged data, or user groups. This lack of quantitative data still keeps HCI researchers from a better understanding of the interdependencies of the factors influencing modality choice and complicates the construction of applicable theoretical and computational models. The existence of these models is a requirement for facilitating simulation of multimodal HCI between user and system models, which further enables automatic usability evaluation.

1.2. Automatic usability evaluation

Ongoing research efforts deal with the evaluation of the quality of MDS (Möller et al., 2011; Perakakis and Potamianos, 2008; Turunen et al., 2010). Judgments on quality are usually gained from user perception. Using questionnaires, users are asked to rate quality features of the system. The SASSI questionnaire was developed to assess subjectively perceived quality of spoken dialog systems (Hone and Graham, 2000). SASSI and other questionnaires were adapted for MDS (Metze et al., 2009) and new questionnaires are currently being developed (Kühnel et al., 2010; Turunen et al., 2010). Several studies have been carried out in this field in recent years. However, studies with real test subjects are quite expensive with respect to typically limited resources of time and money. Here effort could be saved by automatic usability evaluation (AUE), which is therefore a current research topic.

AUE deploys computational (statistical or cognitive) modeling for task solving and decision making in a simulated user. Statistical models can integrate rules (Möller et al., 2006) and utility functions (Gray et al., 2006) for making rational decisions. Cognitive models usually build up on cognitive architectures (Anderson et al., 1997), taking into account human attention, working memory, and strategy selection during decision making. Utility functions are often utilized during simulation of human cognitive processes (Fu and Gray, 2006). In AUE decisions like “how to initialize the next input into the system” have to be made at each step by a simulated user.

Employing task knowledge, information about the user, and system specifications, the AUE Simulation realizes interaction between user and system models. Interaction is simulated by means of tools like the MeMo workbench (statistical modeling) or CogTool (cognitive modeling) and interaction parameters are extracted by gathering log data during the simulated human-machine dialog (Möller et al., 2006; John et al., 2004). A usability profile is generated at the end of the AUE process.

Thus far, no solutions for MDS have left the research phase. One issue is that it is unclear how the simulated user should behave if more than one input modality is offered. Concrete knowledge about users' modality choice strategies is missing. This knowledge is needed to build and validate both theoretical and computational models. As only a small quantity of such data has been collected so far, this step has only rarely been taken. AUE simulation approaches and recent developments in the multimodal domain are discussed in Section 1.3.2.

1.3. Related work

1.3.1. Modality choice

In this section, key findings regarding the factors considered by users when selecting input modalities are presented. Apart from the previously mentioned factors of input performance and modality efficiency, additional factors of mental effort, hedonic quality, subjective preferences, and user attributes, as well as situational context will be discussed.

In HCI, effectiveness is defined as “the accuracy and completeness of users' tasks while using a system” (ISO, 1998). In interactive systems, effectiveness is essentially affected by the input performance (error-proneness) of the interface (Card et al., 1990). Bilici et al. (2000) tested a multimodal setup, where participants switched from speech to manual input when ASR errors arose. Suhm et al. (1999) further found that users tend to select the less error-prone modality after repeated usage of a multimodal system. Effects of input performance are not well documented for the GUI. In a study of our own we observed that participants strongly preferred the VUI if GUI errors occurred (Schaffer and Minge, 2012). Input performance is typically quantified in terms of error rates. The word error rate (WER) is the sum of transcription errors (word substitutions, deletions, and insertions), divided by the number of reference words with lower scores indicating better performance (Möller et al., 2011). Actual ASR systems can widely differ in WER, reaching e.g., from 4.1% for command and control operations in mobile phones (Varga et al., 2002) to 46.7% for speech-to-text transcription of conference room meetings (Fiscus et al., 2008). For AUE of multimodal systems this variance means that a wide margin of error rates has to be considered in order to forecast modality choice. These large variances result from different underlying conditions such as acoustic conditions or the size of vocabulary and grammar. Metrics comparable to WER can also be calculated for other input modes like a GUI (Kühnel et al., 2010).

Efficiency is defined as the effort expended by the user in relation to the accuracy and completeness of goals achieved (ISO, 1998). High efficiency is reached when the user fulfills a task while expending as few resources as possible. For HCI, different metrics can be used to assess effort, such as task completion time, monetary costs, mental effort of the user, or the number of interaction steps required to solve the task (Bevan, 1995). Duration-related metrics are often used to assess the efficiency of multimodal systems. Perakakis and Potamianos (2008) showed that as VUI usage increases, if it is more efficient compared to a GUI, measured in overall time spent in a modality. On the other hand, users tend to use modalities that save additional or inconvenient interaction steps, even if interaction time increases (Rudnicky, 1993). Similar results were found by Wechsung et al. (2010) and in our own studies presented in Section 2.

Mental effort can be described as operators' attentional capacity in relation to environmental demands (Kahneman, 1973). The Multiple resource theory (Wickens and Hollands, 2000) introduces a model of attentional capacities to describe mental workload. High mental workload can be caused by multiple tasks accessing identical cognitive resources. Task fulfillment can be improved if perception and information processing is allocated to distinct resources. Workload interferences between tasks and allocation capabilities can be identified by means of the model. In a SIMS study, Schaffer et al. (2011a) showed that perceived mental effort increases with increasing task complexity for both a VUI and GUI. A follow-up study revealed that participants select the modality

that reduces task complexity if one of the modalities is more efficient in terms of mental effort (Schaffer et al., 2011b). In summary, participants were able to keep perceived mental effort constant by employing specific modalities.

If a product not only satisfies the task-related requirements, but also generates positive feelings in the user, it has hedonic quality (Jordan, 2002). In contrast to pragmatic quality, which focuses on efficient and effective goal achievement, hedonic quality asks for novel interaction techniques and communicates a desired identity through a professional, cool, modern, or different appearance (Hassenzahl et al., 2003). For multimodal HCI, users' modality choice can be affected by the perceived innovative energy or originality attributed to a modality, with innovative modalities being used more frequently.

Personal preferences thus can have a decisive influence on individual modality usage profiles. It was observed that some users did not switch modalities at all but used only either a GUI or VUI (Schaffer et al., 2011a). Further user attributes like physical impairment, attitude, character traits, education, expertise, and affinities can have an influence (Bohn et al., 2005). Especially when dealing with mobile devices, the situational context (Dey and Häkkilä, 2008) and aspects like lighting conditions, surrounding sounds, temperature, legal requirements, and social desirability also have to be considered to explain user behavior (Bernsen, 2008).

Overall, it must be observed that all considered factors are interrelated. Because personal preferences change dynamically, they can only be forecast if information about previous behavior is available. Furthermore, the influence of static user attributes and situational context on modality choice so far has not been examined sufficiently enough to extract computational models. Another constraint relevant to most of the reviewed studies is that not all findings can be generalized and thus cannot usually be compared directly. The conclusions drawn are always post-hoc explanations of user studies with differing material, setup, and participants that have an uncontrolled impact on the various factors influencing modality choice. The need for stable experimental conditions to gain deeper insights for specific factors has motivated our series of experiments. In consequence of a rational decision-making approach, factors related to efficiency and effectiveness, such as input performance and modality efficiency should be taken into account to forecast average user behavior.

1.3.2. Developments in multimodal AUE

An overview of usability evaluation automation (Ivory and Hearst, 2001) summarized 10 simulation approaches supporting automatic usability analysis: 9 cognitive architectures and 1 statistical modeling technique. Morrison (2003) tested the applicability of 19 computational human behavior representations for military simulations, including 7 of the cognitive architectures discussed by Ivory and Hearst (2001). Both papers conclude that the level of expertise required to successfully deploy cognitive modeling is an impediment. Therefore the applicability of cognitive architectures for usability evaluation is still limited due to considerable learning time and the effort needed to employ them. Nevertheless, efforts are being made to utilize cognitive models for AUE. Kieras et al. (1997) predicted human performance in telephone operator tasks with models constructed in the EPIC architecture for human information processing. A simulated operator had to act in multiple modalities: handling a telephone system and processing customer requests. However, the sequence of modalities was predetermined by the task and modalities could not be selected. Anderson et al. (1997) developed the cognitive architecture ACT-R, which is used to model various psychological aspects, including strategy selection. Schaffer and Reitter (2012) produced a model, learning about the efficiency of multiple modalities while interacting with a SIMS. However, the cognitive processes implemented are as yet unproven. Further developments have tried to utilize cognitive modeling for AUE tools by integrating simplified models for specific domains. Distract-R allows designers to prototype new in-vehicle interfaces and to evaluate the specifications by generating predictions of driver distraction (Salvucci, 2009). CogTool is a user interface prototyping tool that automatically evaluates the design with a predictive human performance model to assess total task time (John et al., 2004; Bellamy et al., 2011). Basic modeling of SIMS is possible with CogTool and Distract-R. However, single task steps as well as modality changes have to be specified manually by the modeler.

The statistical simulation method discussed by Ivory and Hearst (2001) is AMME (automatic mental model evaluator), developed by Rauterberg (1996). Petri nets (Reisig and Rozenberg, 1998) are employed to reconstruct and analyze the user's problem-solving process. In addition to a system description, AMME requires log file data in order to generate the Petri net. Although the Petri net is applicable for subsequent simulations, reusability for other systems cannot be assumed as interaction parameters and system description may not be reusable. Möller et al. (2006) presented the MeMo Workbench as a tool for semi-automated evaluation of interactive systems. The approach, supporting the design process as well as the evaluation of the design, is based on simulation with user models and system models and is suitable for VUIs or GUIs (Engelbrecht et al., 2008). MeMo user models build upon statistical simulation according to data-driven, pragmatic, or theory-driven solutions. So far the workbench does not support the simulation of multimodal interaction. An approach by Wechsung and Schleicher (2012) predicts "later" modality preference, based on interaction parameters and quality ratings of the component modalities. As the method builds on ratings of perceived quality (available only if a prototype is already available) the applicability for simulation is limited. Although the authors report that models fit well, they conclude that results have to be validated with larger samples.

Efforts within the computational modeling community show that a high value is assigned to AUE. As future dialog systems will increasingly rely on multimodal input, multiple modalities will also have to be considered during the simulation of interactions. The utility of single modalities thereby plays a central role for SIMSs. In engineering psychology decision makers often choose the option with the greatest expected value (Wickens and Hollands, 2000). Cost-benefit considerations are employed in cognitive decision-making, e.g., for the selection of interactive behavior (Gray et al., 2006) as well as in rational decision-making (Sheridan and Parasuraman, 2000). In conclusion, both approaches rely on a rational analysis perspective, maximizing expected utility. Predictions are based on the assumption that human beings act similar to naive statisticians (Kahneman, 2011). Referring to the last named approaches, we developed a utility-driven model for modality choice.

1.3.2.1. Aim of this paper. Our motivation is to examine the interdependencies of the above-outlined factors that determine modality choice, *modality efficiency* and *input performance*, for the domain of multimodal mobile HCI, and to build models that are able to predict modality usage by means of these factors. The selected task is from the domain of "list browsing," which is of importance for a number of systems, especially for smart phone applications where screen size is limited. Since not all information can be displayed at a time, users have to browse through lists to find desired items. We present three interaction experiments

with a SIMS, testing our following hypothesis regarding modality choice:

H1. If the input performance of a specific modality decreases, the usage of this modality decreases as well.

H2. If GUI efficiency decreases, VUI usage increases.

The eventual aim of this paper is to compile a model for the prediction of modality usage in SIMS that utilizes a utility function. We will judge the predictive quality of the models with goodness of fit measures (R^2 and RMSE), as well as data interpolation and extrapolation. To show the applicability for AUE simulation, a prototypical implementation of the prediction algorithm will be outlined.

1.3.2.2. Content. Three experiments investigating the impact of modality efficiency and input performance on modality choice are reviewed in Section 2. In Section 3 the construction of the model is explained, following with the analysis of predictive power in Section 4. In Section 5 a prototypical AUE simulation is outlined. Section 6 concludes and gives an overview on potential applications of such simulations.

2. Experimental data

In this section a series of three experiments will be summarized. The overall goal of all experiments was to gather comparable data on users' modality choice behavior. Therefore all participants conducted the same tasks with one particular system in controlled laboratory setups. Throughout the experiments six conditions of input performance were differentiated[1]. Modality efficiency was systematically varied in six levels within each experiment. The first experiment, comprising the first condition [G00, V00], aimed at the mere effect of modality efficiency. Perfect input performance for both modalities was simulated using Wizard of Oz speech recognition (Schaffer and Reitter, 2012). In the second experiment, targeting the input performance effect, ASR-error rates of 10% ([G00, V10]) and 30% ([G00, V30]) were simulated (Schaffer et al., 2011a). In the third experiment, ASR-errors of 20% ([sG00, V20]), as well as two further conditions were tested: [G20, V00] touch input producing errors at a rate of 20% while ASR errors amounted to 0%, and [G20, V20] with touch and speech input, both comprising error rates of 20% (Schaffer and Minge, 2012).

2.1. Material

2.1.1. System

For all experiments, a prototype smartphone-based app called the "Restaurant booking Application" (RBA) was used. The app ran on an Android device (G1 HTC) and any possible interaction could be made using both GUI and VUI modalities. Speech input was directly possible at each point in the interaction. There were four slots that had to be filled by the user to accomplish the task of booking a restaurant: city, cuisine, desired time, and number of people. For each slot, the user had to choose his request from a list. Each list contained 24 items and was split into 6 layers, each containing 4 items. To make a request, the user had to press a category button on the touch-screen or speak the written label and then step through the layers to find the designated item (see Fig. 1). After making a selection by touching an item or verbalizing it, the user was redirected to the Home Screen. When all slots were filled, the user was able to send his request, accomplishing the task, and the End Screen was shown.

To simulate ASR, a Wizard-of-Oz design was used. Therefore an open microphone was hidden in the room the participant was working in. In a different room, unseen by the participant was the "wizard," a human operator wearing headphones. The wizard listened to the participants' VUI input and used a graphical user interface (a Java application on a UNIX notebook), to execute the voice commands. Following the WIMP (window, icons, menu, pointer) concept, the wizard interface was designed in a way that responses could quickly be generated. The wizard tried to perform the input about 0.5–1.0 s after an utterance. The feedback was sent to the mobile device via TCP-IP and wireless LAN, simulating the ASR-system. The simulation of errors is separately described for each experiment.

2.1.2. Benefit of speech usage

To make a request using the GUI, the user had to click on a category and then manually switch through layers to find the desired item. In contrast to this, the VUI offered a shortcut. Using the VUI, the user could not only make exactly the same inputs as with the GUI but also choose an item from any list screen without the need to switch through layers by simply pronouncing the name of the desired item. For example, one was able to select "Hamburg" from the first layer, while it would have taken him four interaction steps to select it using the GUI. The benefit of VUI-usage B_{VUI} thereby can be calculated as the difference between necessary GUI-interaction steps IS_{GUI} and VUI-interaction steps IS_{VUI}.[2]

$$B_{VUI} = IS_{GUI} - IS_{VUI} \quad (1)$$

2.2. Task

In all experiments the task was to perform predefined database requests using the RBA. Each request comprised four list-browsing subtasks, namely the choice of a city, a culinary category, a time, and a number of persons. For example, participants had to request for a Chinese restaurant in Berlin at 8 p.m. for 12 persons. The exact steps and system states for this task example are described in Table 1. At first a list had to be selected at the "Home Screen". In doing so a user can choose touch screen or speech input at any time. The participants were instructed that the written labels in the GUI have to be used as speech commands. After the selection of a list the first layer was presented. The 24 items within a list are ordered alphanumerically on 6 layers, whereas each layer presented 4 list items in the GUI. Fig. 1 depicts the choice of the city "Berlin" as an example. The participants were instructed that touchscreen or speech input could be used to select an item. Using speech input all list items from all six layers of a list are already recognized in the first layer, while touch screen input only allows selecting items that are directly visible. Users preferring touch screen input therefore have to browse through the list by pressing the arrow in the downright corner, until the desired list item is displayed.

[1] The naming of the input performance conditions is based on corresponding values of error rates. [G00, V20] for example stands for the condition with a GUI error rate of 0% and a VUI error rate of 20%.

[2] For the RBS IS_{VUI} amounts to a constant 1, as speech input is always directly possible. However IS_{VUI} can amount to other values in different systems. Alternatively, using an OKS setup for the actual system (compare Section 1.1), IS_{VUI} could amount to a constant 2, due to a necessary ASR activation command. If no speech shortcuts were implemented in the actual system and a speech command would still be implemented for each possible GUI input IS_{VUI} would equal to IS_{GUI}. For other systems (like some navigation systems), where the VUI does not directly match the GUI, varying values of IS_{VUI} are possible. As our aim is to create a model that can in principle cover all these kinds of systems, IS_{VUI} is handled as variable at the model construction.

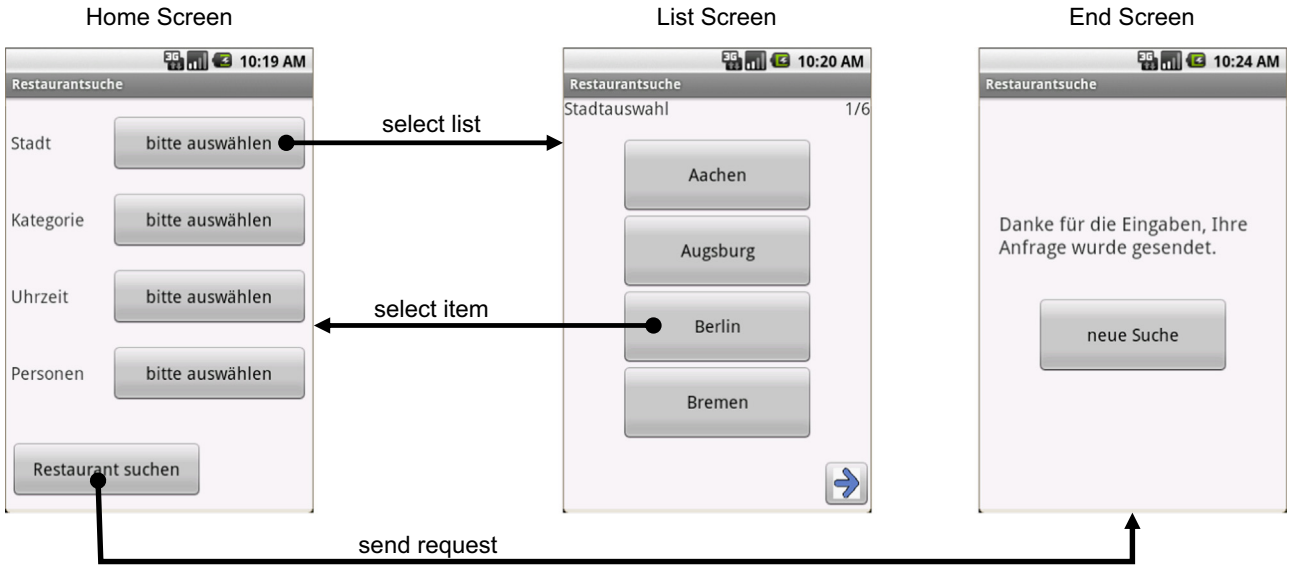

Fig. 1. Screens of the 'Restaurant Booking Application' (RBA).

Table 1
Task example for "Look for a Chinese restaurant in Berlin at 8 pm for 12 persons". From top to bottom the column *System state* indicates the states involved during task processing. Sub-goals are marked with SG. B_{VUI} is calculated as described in Section 2.1.2. The columns "GUI button" and "VUI utterance" explain the user input interaction. Interactions printed in bold lead to a sub-goal. If the VUI is utilized to say "twelve" at *Persons list 1*, the home state (*Home SG3*) is directly presented. For the calculation of B_{VUI}, the columns IS_{GUI} and IS_{VUI} indicate the corresponding values of interaction steps (the needed steps until the next sub-goals).

System state	B_{VUI}	GUI button	IS_{GUI}	VUI utterance	IS_{VUI}
Home	0	Press "select City"	1	Say "select city"	1
City list 1	0	**Press "Berlin"**	1	**Say "Berlin"**	1
Home (SG1)	0	Press "select Category"	1	Say "select category"	1
Category list 1	1	Press "right arrow"	2	**Say "Chinese"**	1
Category list 2	0	**Press "Chinese"**	1	**Say "Chinese"**	1
Home (SG2)	0	Press "select Persons"	1	Say "select persons"	1
Persons list 1	3	Press "right arrow"	4	**Say "twelve"**	1
Persons list 2	2	Press "right arrow"	3	**Say "twelve"**	1
Persons list 3	3	Press "right arrow"	2	**Say "twelve"**	1
Persons list 4	0	**Press "12 persons"**	1	**Say "twelve"**	1
Home (SG3)	0	Press "select time"	1	Say "select time"	1
Time list 1	2	Press "right arrow"	3	**Say "eight"**	1
Time list 2	1	Press "right arrow"	2	**Say "eight"**	1
Time list 3	0	**Press "8 pm"**	1	**Say "eight"**	1
Home (SG4)	0	Press "Search Restaurant"	1	Say "search restaurant"	1

2.3. Procedure

All experiments followed roughly the same procedure. Adaptions can be found in the individual subsections. After welcoming the participants, demographic data was gathered in a short questionnaire. Then, the participants were given an introduction, explaining the usage of the system with both the GUI and VUI. To control whether participants understood the instruction and were able to use the system with both modalities, three training trials had to be done: one using GUI only, one using VUI only, and one with a free choice of modality. After this training, the target trials started. Adaptions regarding the simulation of errors are described in the individual subsections. Participants were informed about the Wizard-of-Oz design only after all target trials were finished. Depending on the specific conditions of the experiments, input performance could be impaired. However the participants were not informed about the possible occurrence of errors beforehand. As the measure of modality efficiency, B_{VUI} was systematically varied between 6 levels (0–5 interaction steps) within the individual trials. Modality usage data was collected using log files. This experimental setup enabled to measure the effects of modality efficiency and input performance on modality usage by means of two independent variance analyses. During the experiment, also data about the perceived mental effort and product quality was gathered. This data will not be further examined here (see Schaffer and Reitter (2012), Schaffer et al. (2011a) and Schaffer and Minge (2012) for more information). All experiments took roughly 45 min to one hour and were remunerated with 10€.

2.4. Experiment 1: the influence of modality efficiency

2.4.1. Adapted procedure

In the first experiment (Schaffer and Reitter, 2012) 16 German-speaking participants (8 female) tested the RBA. The age ranged from 22 to 31 years ($M=26$, SD=2.95). The condition [G00, V00] was tested. Since the VUI and GUI were both working perfectly, neither producing any errors. For each of the 12 trials, an instructor was presenting the current task on a paper.

2.4.2. Results

The means and standard deviations of speech usage are reported in Table 2. A one-factorial repeated measures ANOVA showed a highly significant effect of B_{VUI} on the usage of speech ($F(2.27, 33.97) = 27.503; p_{1-tailed} < 0.001;\ \eta_p^2 = 0.647$). The results confirmed the hypothesis that modality efficiency in terms of benefit with reference to interaction steps is moderating users' modality choice. Speech usage increases with increasing efficiency of the VUI.

2.5. Experiment 2: the influence of ASR performance

2.5.1. Adapted procedure

In the second experiment, 33 German-speaking participants were tested (Schaffer et al., 2011a). Four participants had to be

Table 2
Experiment 1 means *M* and standard deviations SD of speech usage in dependence of B_{VUI}.

Condition	B_{VUI}	0	1	2	3	4	5
[G00, V00]	*M* [%]	0.31	0.73	0.83	0.90	0.94	0.96
	SD [%]	0.36	0.26	0.22	0.18	0.09	0.08

excluded as three did not follow the instructions and one experienced a severe malfunction of the system. The remaining sample consisted of 11 females and 18 males, with a mean age of 25 years (SD=3.7). Error rates were generated randomly, varying between 0 and 40 percent. Thus the error-rate was individual and varied between participants. The participants were clustered in two groups post-hoc. The first group (constituting the condition [G00, V10]), consisted of 17 subjects who experienced error rates of approximately 10% and (*er*=0–20%, *M*=9.13%, SD=6.0%). The second group (constituting the condition [G00, V30]) contained 12 subjects who experienced error rates of approximately 30% (*er*=21–40%, *M*=30.23%, SD=4.28%). ASR errors were randomly inserted and could occur in each step. In terms of the error probability corresponding to the individual error-rate, it was possible that errors were induced directly in succession. After an error the home screen was presented with no result and the textual message "I did not understand," and the search within the list screen had to be started again. The 15 trials per user took place in the same experimental setup as in the first experiment.

2.5.2. Results

The means and standard deviations of speech usage are reported in Table 3. Corresponding with the first experiment a one factorial repeated measures ANOVA showed a highly significant effect of B_{VUI} on the usage of speech ($F(2.47,65.79)=74.222$; $p_{1\text{-}tailed} < 0.001$; $\eta_p^2 = .733$). Further significant differences between the two error conditions could be observed ($F(1, 27)=6.94$; $p < 0.007$; $\eta_p^2 = 0.204$). The results confirmed the hypothesis that input performance moderates users' modality choice. Speech usage decreases with increasing ASR error rate. A significant interaction between both factors could not be observed.

2.6. Experiment 3: the influence of ASR, GUI and mixed performance

2.6.1. Adapted procedure

Finally, in the third experiment (Schaffer and Minge, 2012), 48 German-speaking subjects (24m, 24f) participated. The mean age was 24.2 years (SD=3.73). ASR errors were implemented as in experiment 2, with the difference that fixed error rates of 0% or 20% were simulated. GUI errors were integrated by blocking one in five input attempts on the touch screen for 1.4 s, the average time needed to recover after an ASR error. The blocking mechanism was implemented so that it counted across individual tasks. Therefore GUI errors could also occur in the first step when using GUI input as off the second task. Due to VUI input involvement, GUI errors were further distributed over the interaction steps. At the end of the experiment, each subject was asked about the perceived authenticity of errors. None of the participants noticed that errors were simulated, and none stated that they had recognized the errors as artificial. The participants were evenly distributed to the following four error conditions:

[G00, V00] GUI and VUI both working perfectly (as in the first experiment).
[G20, V00] 20% GUI error rate and VUI working perfectly.
[G00, V20] GUI working perfectly and 20% VUI error rate.
[G20, V20] 20% GUI error rate and 20% VUI error rate.

Table 3
Experiment 2 means *M* and standard deviations SD of speech usage in dependence of B_{VUI}.

Condition	B_{VUI}	0	1	2	3	4	5
[G00, V10]	*M* [%]	0.35	0.75	0.92	0.93	0.90	0.93
	SD [%]	0.32	0.30	0.12	0.16	0.17	0.19
[G00, V30]	*M* [%]	0.17	0.66	0.70	0.81	0.77	0.84
	SD [%]	0.18	0.26	0.25	0.11	0.22	0.18

Table 4
Experiment 3 means *M* and standard deviations SD of speech usage in dependence of B_{VUI}.

Condition	B_{VUI}	0	1	2	3	4	5
[G00, V00]	*M* [%]	0.40	0.55	0.68	0.74	0.79	0.69
	SD [%]	0.36	0.39	0.36	0.31	0.30	0.33
[G20, V00]	*M* [%]	0.75	0.89	0.88	0.87	0.86	0.89
	SD [%]	0.35	0.21	0.21	0.25	0.25	0.23
[G00, V20]	*M* [%]	0.37	0.72	0.71	0.80	0.73	0.86
	SD [%]	0.32	0.28	0.24	0.28	0.32	0.29
[G20, V20]	*M* [%]	0.59	0.86	0.93	0.92	0.91	0.88
	SD [%]	0.26	0.18	0.08	0.22	0.11	0.21

A total of 15 target trials were conducted in a closed acoustic booth. The instructor was sitting outside and the subject was alone during the trials. The tasks were automatically presented on a screen when the previous task was accomplished.

2.6.2. Results

The means and standard deviations of speech usage are reported in Table 4. Consistent with the first and the second experiment, a one factorial repeated measures ANOVA showed a highly significant effect of B_{VUI} on the usage of speech ($F(2.96, 130.32)=16.72$, $p < 0.001$; $\eta_p^2 = 0.275$). Further a highly significant effect of the error condition on speech usage could be observed ($F(3, 44)=4.75$, $p=0.006$, $\eta_p^2 = 0.25$). Post-hoc Scheffé tests showed significant differences between the conditions [G00, V00] and [G20, V00], as well as between [G00, V00] and [G20, V20]. The results confirmed that previously unconsidered GUI errors also affect modality choice. Speech usage increases if GUI errors occur. The study further revealed that touch errors are punished harder than ASR errors, as only marginal differences between the [G20, V00] and [G20, V20] conditions could be observed for high VUI efficiency. A significant interaction between benefit and error rate could not be observed.

2.7. Resulting database

The series of experiments was designed for minimizing inter-experimental differences. However comparability may be limited to some extent, as changes in laboratory rooms and task presentation method appeared. As baseline condition data comprising errorless interaction for both input modalities was gathered in experiments 1 and 3. No significant differences in modality choice could be found between the respective data sets ($F(1,26)=2.54$; $p < 0.123$; $\eta_p^2 = 0.089$). To avoid overrepresentation of the baseline condition only data from experiment 1 was used. Taken together data from six different error conditions comprising 82 participants in all was gathered and merged into one database.

The database was checked for outliers potentially causing significant degradation of a model's predictive power. Two criteria have been considered. The first pertains to the slope of the linear model computed from the modality usage profile of individual participants. With respect to the effect of the levels of interaction

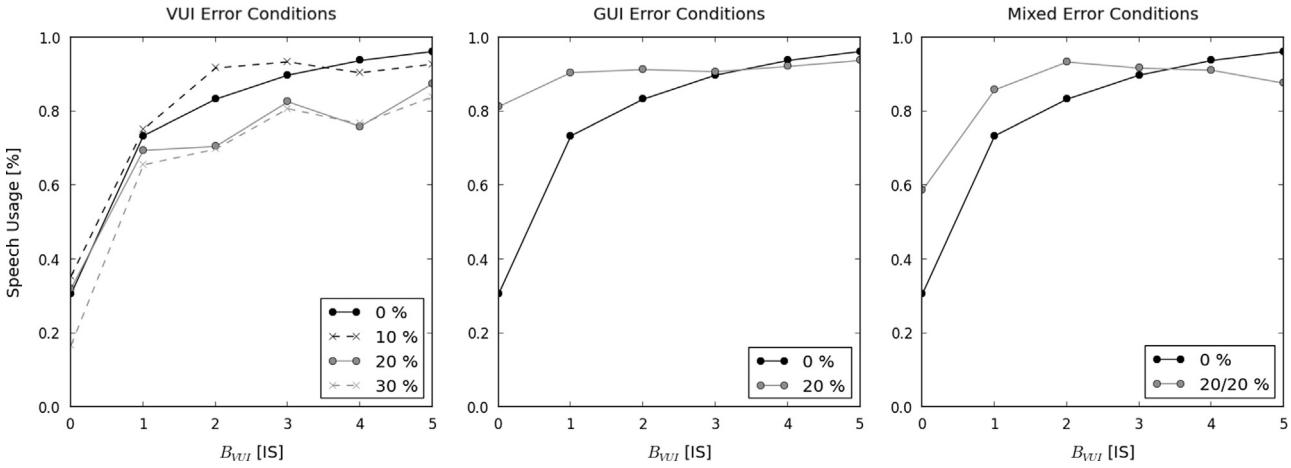

Fig. 2. Modality usage curves gathered from three experiments. The 0% baseline was gathered in experiment 1 (see also Section 2.4). The 10% and 30% VUI error conditions were gathered in experiment 2 (see also Section 2.5). The 20% VUI error condition as well as the 20% GUI error condition and the 20/20% mixed error condition were gathered in experiment 3 (see also Section 2.6).

steps, the slope should be positive for all participants of all experiments. If the slope is negative the participants' modality choice trend is contrary to average user behavior. Considering all experiments, the slope of 5 participants turned out to be negative. It was determined that participants with a slope smaller than -0.5 should not be used for model development, as their data will significantly reduce model performance. One participant from the [G00, V20] condition of experiment 3 did not match this requirement and was therefore omitted. Furthermore, the impairment of model performance is less intense, if negative slopes are small and the average amount of speech usage is relatively high. Thus, as a second criterion, it was determined that participants should be excluded if the slope was negative (between -0.5 and 0) and average speech usage was smaller than 50%. One participant from the [G20, V00] condition of experiment 3 was therefore omitted from the model development. Differences between Fig. 2 and the tables reporting the results of the experiments can be attributed to the omission of outliers.

For each of the 80 remaining participants the resulting database contains the error condition and the average speech usage for each level of benefit. Generated by averaging speech usage over all participants of a condition, the speech usage curves in Fig. 2 yield a summary of the data. All curves are arranged according to the error conditions. All graphs illustrate the percentage of speech usage (y-axis) as a function of B_{VUI} (x-axis). Despite certain inter-experimental differences the curves are mostly consistent with each other. Regarding all VUI error conditions, speech usage increases with increasing B_{VUI}. Furthermore speech usage decreases with increasing ASR error rate. Only minor differences can be observed between error curves of 0% and 10%, as well as between 20% and 30%. The GUI error conditions, consisting of a baseline and a 20% touch error curve, show mostly consistent behavior. Speech usage increases in the presence of GUI errors. Regarding the 20/20% curve in the mixed condition speech usage is decreased for $B_{VUI} = 0$, as the VUI cannot assert its benefit in efficiency. For higher levels of interaction steps, speech usage increases and the curve runs similar to the 20% GUI error condition curve. As both ASR and GUI errors are present in the mixed condition, this implies that touch errors are punished more severely then ASR errors.

3. Prediction model construction

The derivation of a computational model that enables the prediction of modality usage if more than one input modality is offered is described in this section. Section 3.1 briefly summarizes the motivation for a utility-driven model. In Section 3.2, the number of required interaction steps as expected by users to solve a task is defined as a criterion to assess the costs of using specific modalities. Afterwards, modality utility and modality usage probability are established in Sections 3.3 and 3.4.

Table 5
Factors relevant to the utility driven model.

Factor	Description	Domain
i	Interaction steps needed for solving a task or subtask using a specific modality	$i \in \mathbb{N}$
er	The average error rate of a modality	$er \in [0, 1]$

3.1. Utility-based modality choice

The empirical data presented in Section 2 implies that system users adapt modality usage to the estimated utility of modalities. VUI is usually preferred, if it is more efficient in terms of interaction steps. In contrast speech usage decreases with increasing ASR error rate. The perceived utility, guiding users' modality choice, is affected by modality efficiency and input performance. The factors come along with a cost-benefit tradeoff: the expected utilities of modalities are offset against each other. In our model, efficiency is operationalized as the number of needed interaction steps to solve the task. If a task can be solved with fewer interaction steps in a specific modality, a shortcut exists, increasing the probability of modality usage. Further, input performance is operationalized by system errors like ASR errors or touchscreen malfunction. These factors are outlined in Table 5. The factor i is determined either by overt clicks in the GUI or by utterances aimed at inputting information using the VUI. An example of interaction steps considered by the model is depicted in Table 1. The steps partly correspond to the steps of a process model, as defined by KLM or GOMS (John and Kieras, 1996). A close affinity between the utility model and process models is worthwhile, as thereby the integration of the modality choice mechanism into AUE tools like CogTool (John et al., 2004) is supported. Process models mainly predict the time it takes to complete a task and are not directly able to predict modality selection. In a GOMS way of speaking, GUI and VUI input can rather be understood as different methods and the modality prediction model could eventually be used as a selection rule.

3.2. Expected number of interaction steps

3.2.1. Probability criterion for a one-step task

As mentioned before, empirical evidence implies that system users weigh the expected number of needed interaction steps while concurrently considering alternative modalities. The user can easily anticipate the pure impact of interaction steps if the steps to solve the task are easily comprehensible. However, the user cannot be absolutely sure about forecasting the exact number of interaction steps, since more steps than expected may be necessary due to system errors. Therefore, the aim is to operationalize the expected number of interaction steps by means of both factors, specifically interaction steps and system errors. Thus, a criterion incorporating interaction steps and error rate is needed to capture the costs of using a certain modality.

At first, *one-step* tasks are considered, i.e., tasks solvable in only one interaction step. In the case of the RBA, selecting a specific list and browsing for an item within a list are defined as distinct subtasks. *One-step* tasks are subtasks where the required list item can be found on the first list screen. Thinking about possible system errors er, such a task can be conducted with 1 to n steps, as errors entail that system input has to be re-performed. The probability $p(n)$ for solving the task after n steps is

$$p(n) = (1-er)\cdot er^{n-1} \tag{2}$$

Fig. 3 illustrates this formula. System errors are indicated by Xes, task completion is indicated by a checkmark. If no errors occur $(n=1)$, $p(1)$ equals the accuracy $(1-er)$ of the input interface. If errors occur $(n>1)$, the error probability er has to be considered for each error-prone interaction. Thus, if n steps are needed, er takes effect $n-1$ times until the goal is reached after the nth step at which the accuracy $1-er$ has to be attributed.

According to the probability axioms, the total probability of all possible events $(1\le n\le\infty)$ must equal 1. In this situation the limited number of screens in the RBA does not act as an upper bound of the sum. The upper bound only states the number of entries that can be misinterpreted by the system directly after another, which is by definition not limited.

$$\sum_{n=1}^{\infty} p(n) = 1 \tag{3}$$

Utilizing the convergence of the geometric series (Cgs.), one can prove that this is the case for the aforementioned formula of $p(n)$

$$\sum_{n=1}^{\infty} p(n) = \sum_{n=1}^{\infty} (1-er)\cdot er^{n-1} = (1-er)\cdot \sum_{n=1}^{\infty} er^{n-1} \overset{\text{Cgs.}}{=} (1-er)\cdot\frac{1}{1-er} = 1 \tag{4}$$

However, a limitation must be mentioned: in the case of occurring ASR errors, the system reaction can be diverse. ASR errors can lead to other system states from where it is not possible to proceed directly with the actual task. In these cases, more than one step may be necessary to start a new trial (depending on the system implementation and the error type). The proposed model in simplified terms assumes that a new trial is possible directly after each system error.

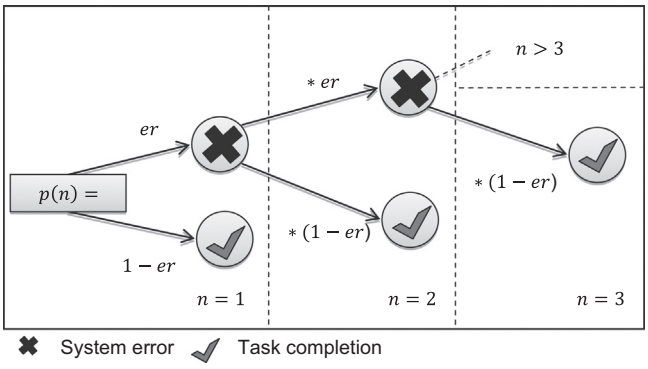

Fig. 3. The probability to solve a *one-step* task with n steps.

3.2.2. Expected number of interaction steps for a one-step task

For a task actually solvable with one step, a higher number of steps can be expected depending on the probability of system errors. If the number of actual interaction steps corresponds to a stochastic variable S_1, for a one-step task, the expected number of steps can be calculated by means of the expected value $E(S_1)$

$$E(S_1) = \sum_{n=1}^{\infty} n\cdot p(n) = \sum_{n=1}^{\infty} n\cdot(1-er)\cdot er^{n-1} \overset{\text{Cgs.}}{=} \frac{1}{1-er} \tag{5}$$

The last simplification of the above expression again results from the characteristics of the geometric series (Cgs.).

3.2.3. Expected number of interaction steps for an i-steps task

In an *i-steps* task, the probability of system errors has to be considered in each step. Thus, the expected number of interaction steps S can easily be deduced from the one-step task equation. A task with i steps involves one-step i-times, resulting in the expected value

$$E(S) = E(S_1)\cdot i = \frac{i}{1-er} \tag{6}$$

3.3. Modality utility

3.3.1. Objective utility

If two input modalities are offered alternatively, the number of interaction steps to solve the task can differ from each other for each modality. In other words, counting in interaction steps, the utility of two modalities can differ. Thereby, the higher utility is attributed to the modality incorporating less interaction steps to solve the task.

An objective representation of a modalities' utility U_o can be generated from the inverse of the expected value of interaction steps. U_o is then a function of interaction steps i and error rate er.

$$U_o(i,er) = \frac{1}{E(S)} = \frac{1-er}{i} \tag{7}$$

Assuming an ideal case with $er=0$, the utility for a task with one interaction step equals 1

$$U_o(1,0) = \frac{1-0}{1} = 1 \tag{8}$$

The calculation example illustrates the maximum objective utility. With $i\in\mathbb{N}$ and $er\in[0,1]$, one can assume that $U_o\in[0,1]$. Utility decreases if either interaction steps i or the error rate are increasing.

3.3.2. Modality specific weighting of interaction steps

The equation can further be extended by a weight that integrates modality-specific effects of interaction steps. The additional factor should thereby incorporate modality choice moderators like the average time per interaction step of a modality, cognitive load caused by using the modality or personal preferences.

$$U_t(i,er) = \frac{1-er}{i\cdot t} \tag{9}$$

As we assume that the weight correlates with the average task time, we choose the identifier t and determine its domain as $t\in\mathbb{R}^+$. The value of t will be derived by parameter fitting.

3.3.3. Modality specific weighting of system errors

Errors caused by different input components of a system are differently perceived and rated by the users. Empiricism shows that touch screen errors have higher influence on modality choice than ASR errors (Schaffer and Minge, 2012). According to the integration of modality-specific effects of interaction steps, the effects of system errors can be considered by extending the objective utility equation with a weighting factor w (with $w \in \mathbb{R}^+$). The value of w will be derived by parameter fitting.

$$U_w(i, er) = \frac{1 - er \cdot w}{i} \tag{10}$$

3.3.4. Perceived utility

To obtain a utility function affected by modality specific effects regarding interaction steps and input performance, both factors have to be considered jointly. Based on probability theory, the utilities U_t and U_w can be seen as statistically independent (Chow and Teicher, 2003). Thus the joint utility equals the product of single utilities. Perceived utility is therefore calculated by multiplying the weighted utilities U_t and U_w. The resulting function for perceived Utility U is

$$U(i, er) = U_t \cdot U_w = \frac{1 - er}{i \cdot t} \cdot \frac{1 - er \cdot w}{i} = \frac{1 - er - er \cdot w + er^2 \cdot w}{i^2 \cdot t} \tag{11}$$

We assume that free variables differ between systems and user groups. Thus, values have to be calculated for each system version and user group, to obtain valid predictions for modality choice.

3.4. Modality usage probability

To calculate the usage probability of a modality m_1, the utility U_{m1} is divided by the sum of utilities of the involved modalities. In this way, modality usage probability P_{m_1} for the usage probability of modality m_1 is constructed. Regarding two input modalities, P_{m_1} integrates all considered variables and is thus a function of all error rates and interaction steps of all modalities.

$$P_{m1}(i_{m1}, er_{m1}, i_{m2}, er_{m2}) = \frac{U_{m1}}{U_{m1} + U_{m2}} = \frac{i_{m2}^2 \cdot (1 - er_{m1} - er_{m1} \cdot w_{m1} + er_{m1}^{\,2} \cdot w_{m1})}{i_{m2}^2 \cdot (1 - er_{m1} - er_{m1} \cdot w_{m1} + er_{m1}^{\,2} \cdot w_{m_1}) + c \cdot i_{m1}^2 \cdot (1 - er_{m2} - er_{m2} \cdot w_{m_2} + er_{m2}^{\,2} \cdot w_{m_2})} \tag{12}$$

The free variable c integrates the factors t of both modalities ($c = t_{m_1}/t_{m_2}$). To enhance readability, we substitute arithmetic expressions in numerator and denominator[3] and get our final form of the modality probability after rearranging as

$$P_{m1}(i_{m1}, er_{m1}, i_{m2}, er_{m2}) = \frac{1}{1 + c \cdot (b/a)} \tag{13}$$

To summarize, we derived a model to forecast the probability of modality usage, based on the perceived utility of currently available input modalities. The perceived utility is determined by the expected number of interaction steps, which can be calculated as a function of input performance and modality efficiency. By incorporating performance- and efficiency-specific weights, the model is adaptable to new systems and different user groups. The model computes higher modality usage probability values if the usage of a modality leads to fewer interaction steps and maintains a lower error rate.

[3] With $a = i_{m2}^2 \cdot (1 - er_{m1} - er_{m1} \cdot w_{m1} + er_{m1}^{\,2} \cdot w_{m1})$ and $b = i_{m1}^2 \cdot (1 - er_{m2} - er_{m2} \cdot w_{m2} + er_{m2}^{\,2} \cdot w_{m2})$

4. Analysis of predictive power

In this section the predictive power of the model described in Section 3 is evaluated. Implementation, model parameters, target values, and the evaluation procedure are described in Section 4.1. Section 4.2 depicts the results of the performance analysis of specialized models (built from ASR, touch, or mixed error conditions) as well as an integrative model (built from all conditions). The performance of the integrative model on specialized data is further evaluated in section 4.3. The results are discussed in section 5.

4.1. Model settings

4.1.1. Implementation and parameter fitting

The modality probability model is implemented in Python. The free parameters w_{m_1}, w_{m_2}, and c are fitted by means of the Sequential Least Squares Programming (SLSQP) solver of SciPy 0.11. In our case, SLSQP optimization minimizes the error between predicted values of speech usage and corresponding empirical target values.

4.1.2. Input parameters and model data sources

The database described in Section 2.7 was used for parameter fitting. Table 6 outlines the data of 6 the different conditions[4]. Column n indicates the number of participants for each condition. According to the modality probability equation derived in Section 3, for each modality the model's input parameters are the simulated error rates (er_{VUI} and er_{GUI}) and the number of interaction steps (i_{VUI} and i_{GUI}) form the current state to the next (sub-) goal. Due to speech shortcuts i_{VUI} is constant, whereas i_{GUI} varies from 1 to 6 to reach the (sub-) goals. The parameter i_{GUI} was varied in all experiments, and is available for each experimental condition. Therefore for each combination of i_{VUI} and i_{GUI} the probability of speech usage p_{VUI} is also differentiated in six levels for each participant.

Four different versions of the model were trained, each comprising different experimental conditions as data sources. The last four columns of Table 6 specify precisely what data were used for each model. The first three models are a specialized speech error-driven (SED) model, a touch error-driven (TED) model, and a mixed error-driven (MED) model. The fourth model integrates all conditions (INT model). The SED model is trained using data sets comprising speech recognition errors. The TED model as well as MED model training data comprises data sets with touch screen errors respectively mixed errors conditions.

4.1.3. Target variables

The prediction models estimate the probability of users' modality choice. An individual modality-usage profile is available for each individual test subject. However, the individual data may introduce a significant amount of noise in the prediction, as modality-usage behavior strongly varies between participants. Furthermore, averaging over all participants of a condition may lead to a loss of information.

The experiments described in Section 2 were designed for collecting modality usage data for six different levels of benefit of speech (with B_{VUI} amounting from 0 to 5) and six experimental conditions (differing in the error setting). We therefore decided to use the following two target variables:

[4] The different conditions are gained from the three experiments described in Section 2. For model development the data from all three experiments was combined. The experiments were designed for minimizing inter-experimental differences (comp. Section 2.7).

Table 6
Factors relevant to the models.

Cond.	n	er_{VUI}	er_{GUI}	i_{VUI}	i_{GUI}	p_{VUI}	SED	TED	MED	INT
[G00, V00]	1–16	0	0	1	1–6	$p_{n,1}-p_{n,6}$	X	X	X	X
[G00, V10]	1–16	0.1	0	1	1–6	$p_{n,1}-p_{n,6}$	X	–	–	X
[G00, V20]	1–12	0.2	0	1	1–6	$p_{n,1}-p_{n,6}$	X	–	–	X
[G00, V30]	1–13	0.3	0	1	1–6	$p_{n,1}-p_{n,6}$	X	–	–	X
[G20, V00]	1–11	0	0.2	1	1–6	$p_{n,1}-p_{n,6}$	–	X	–	X
[G20, V20]	1–12	0.2	0.2	1	1–6	$p_{n,1}-p_{n,6}$	–	–	X	X

- Individual subject predictions are obtained by averaging modality usage over the same levels of benefit of each participant, resulting in 80 individual modality usage profiles.
- Averaged predictions are obtained by averaging modality usage over the same levels of benefit and over all participants of a condition, resulting in averaged modality usage profiles for the 6 conditions.

4.1.4. Performance evaluation

Performance has been analyzed for the distinguished target variables and two training cases:

- All cases (ALL): Within-data performance was analyzed using identical training and test sets (interpolation). The analysis was conducted for modality usage predictions of individual subjects as well as for averaged data. All cases are used as training data.
- Leave-one-out (L1O): Out-of-data performance was analyzed by performing leave-one-out cross-validation (extrapolation). For modality usage predictions of an individual subject, data from one user is omitted in the training. The respective data are taken for testing a model trained on the remaining $n-1$ users. For modality usage predictions of averaged data, data from one condition is omitted in the training, and the respective data is taken for testing a model trained on the remaining conditions.

The performance of the obtained models has been evaluated by means of the amount of covered variance R^2 and by the root mean squared error *RMSE*. Prediction power of the integrative model on specialized data (speech error, touch error or mixed error data) is evaluated by performance comparison with specialized models. The lower the degradation of the INT model performance for the specific data, the higher is the robustness that can be attributed to it.

4.2. Performance analysis results

The performance measures for all models are given in Table 7. For the within-data (ALL), R^2 amounts from 0.40 to 0.44 for individual subject target values and 0.90 to 0.95 for averaged target values. Best results are obtained for the MED model. According to our assumption, averaged data provides better performance than the prediction of individual subjects' modality usage. Individual modality-usage profiles can vary strongly, resulting in lower R^2. The *RMSE* amounts from 0.21 to 0.24 for individual subject target data and 0.04 to 0.06 for averaged data. Higher variance of individual subject data causes higher error values. Lowest error values are obtained for the MED model.

When testing the models on unseen test data (L1O), the performance decreases in all cases. This result shows that the models are better in interpolating the training data than in extrapolating to unseen test data. Regarding individual subject target data, the lowest degradation is observable for the SED model, whereas for averaged data the INT model is on a par with the SED model. The considerable degradation of the TED model for averaged target values can be attributed to the small amount of available training data. For the cross-validation, averaging over two conditions results in only one test and training set. The performance decreases due to considerable differences between these two conditions. Although the lack of training data also has an effect on the MED model performance, the decrease in R^2 is considerably smaller. The lowest *RMSE* can be reported for the INT model for averaged target data and for the MED model for individual subject data.

Table 7
Performance on training (ALL) and independent test data (L1O) with the speech error driven (SED), touch error driven (TED), mixed error driven (MED) and integrative (INT) models. Values in bold indicate best performance.

Configuration		Training	Performance	
Model	Target		R^2	*RMSE*
SED	Individual subject	ALL	0.433	0.243
SED	Averaged	ALL	0.910	0.066
TED	Individual subject	ALL	0.400	0.220
TED	Averaged	ALL	0.936	0.044
MED	Individual subject	ALL	**0.435**	**0.211**
MED	Averaged	ALL	**0.948**	**0.042**
INT	Individual subject	ALL	0.407	0.232
INT	Averaged	ALL	0.899	0.063
SED	Individual subject	L1O	**0.414**	0.245
SED	Averaged	L1O	**0.883**	0.075
TED	Individual subject	L1O	0.349	0.229
TED	Averaged	L1O	0.054	0.167
MED	Individual subject	L1O	0.402	**0.218**
MED	Averaged	L1O	0.522	0.126
INT	Individual subject	L1O	0.388	0.235
INT	Averaged	L1O	0.865	**0.073**

4.3. Specialized data prediction power of the integrative model

The integrative model is applicable to varying error conditions, as it is trained on all available data. However its capability in predicting special condition data may be limited. If only ASR and no GUI errors arise, the predictive power of the integrative model can be reduced, as the GUI error and mixed error conditions can have a contradictory impact on the fitting of free parameters used by the SED model. The performance of the specialized SED model may in this case be higher. However, with regard to conditions where more data is necessary for consistent model development, the inclusion of similar conditions might improve prediction performance. Therefore the predictive power of the integrative model on specialized error conditions is of interest. In Table 8 the performance values of the integrative and specialized models on specialized test data is compared.

Considering within data (ALL) comparisons, only marginal performance losses can be observed for the integrative model. The differences between averaged and individual subject target data are in line with the results from Section 4.2. The performance comparison on unseen data (L1O) reveals consistently increased predictive power of the integrative model. Significant gain can be observed for average MED and TED data. The inclusion of all available data here takes effect. If not enough specialized data is available, notable performance improvements are possible by employing the integrative model.

4.4. Discussion

Four models for predicting modality choice in differing error conditions were analyzed with respect to their performance in both describing known within-data (interpolation) and predicting unknown test data (extrapolation). For extrapolation leave-one-out

Table 8
Performance comparison of the integrative and specialized models on training (ALL) and independent test data (L1O). Values in bold indicate best performance.

Configuration		Training	Specialized models		Integrative model	
Test data	Target		R^2	*RMSE*	R^2	*RMSE*
SED	Individual subject	ALL	0.433	0.243	0.432	0.241
SED	Averaged	ALL	0.910	0.066	0.909	0.066
TED	Individual subject	ALL	0.400	0.220	0.391	0.222
TED	Averaged	ALL	0.936	0.044	0.911	0.051
MED	Individual subject	ALL	**0.435**	**0.211**	0.426	0.213
MED	Averaged	ALL	**0.948**	**0.042**	0.929	0.049
SED	Individual subject	L1O	0.414	0.245	**0.416**	0.245
SED	Averaged	L1O	0.883	0.075	0.883	0.075
TED	Individual subject	L1O	0.349	0.229	0.366	0.226
TED	Averaged	L1O	0.054	0.167	0.841	0.069
MED	Individual subject	L1O	0.402	0.218	0.407	**0.217**
MED	Averaged	L1O	0.522	0.126	**0.850**	**0.071**

cross-validation was performed. Further an integrative model was compared to three specialized models, regarding the predictive power on specialized error conditions.

Concerning the training on all available data (interpolation), all models show considerable fit on averaged target data. Significant performance degradation can be observed for all models when describing individual subject behavior. The effect can be attributed to differences in individual modality-choice profiles. Individual subject profiles sometimes show contradictory behavior compared to averaged data, resulting in decreased goodness of fit.

For extrapolation the SED model performs best with respect to both average and individual subject behavior. If only ASR errors are present in the data, average user behavior is generally predicted correctly. The extrapolation performance of the integrated model is similarly good. Considerable performance losses when compared to interpolation can be observed for the MED and the TED model. These losses can be attributed to the lack of available training conditions. Data for more differing GUI error conditions as well as mixed error conditions are needed to build models that can enable improved predictive power for unseen data.

The examination of the predictive power of the integrative model on specialized error conditions revealed only marginal performance losses for within data. Testing the integrative model on unseen data revealed consistently increased performance for all conditions. For average MED and TED data an enormous increase of predictive power was observed, strengthening the reliability of the integrative model. All in all, the integrative model seems to be a useful means to predict modality choice for various error conditions. The remarkable extrapolation strength for unseen user behavior, in particular, constitutes a substantial value and supports the model's transferability capabilities and validity.

It can be expected that the integrative model would beneficially support simulation tools for multimodal HCI since interactive behavior should be reproduced more realistically than by randomly selecting modalities. Existing impairments in predictive power and prediction errors can be caused by factors influencing modality choice that are not covered by the model so far. In the field of AUE such factors can also bias the usability predictions generated from simulated interactions. It is possible that unconsidered factors have an effect on specific tasks, systems, or user groups. To gain knowledge about which usability evaluations should be treated with caution, it will be necessary to identify reliability gaps in the proposed model as well as the factors causing these gaps. Detailed comparisons between real user and simulated task processing could reveal such unconsidered factors. The following section demonstrates how the model can successfully be applied. Furthermore, an existing limitation of the model will be shown.

5. Application example

Several HCI studies revealed that multimodality affects the quality judgments of system users (Metze et al., 2009; Wechsung et al., 2009). The employment of specific modalities can lead to different experiences during the interaction. As a consequence, modalities should be selected as accurately as possible for automatic usability evaluation (AUE), estimating the quality of multimodal systems. The application example aims at demonstrating benefits and limitations of the developed modality choice mechanism for averaged user behavior simulation as applied in AUE.

A system model of the Restaurant Booking Application (RBA) introduced in Section 2 is implemented as a finite state machine, a concept often used as a computational representation of interface designs in the field of AUE (Möller et al., 2006). Looking for a restaurant using the RBA, the main task is to select predefined list items. The list selections comprise the following subtasks: (1) city, (2) cuisine, (3) time and (4) people. Within a list, subtasks are directly processable using the VUI. As each list item is directly accessible by speech input, the system reveals VUI shortcuts. In the case of the GUI, all items were ordered alphanumerically. In the numerical subtask (4) the systems' list for selecting the number of people runs from 1 to 24, and in subtask (3) the list to select the time runs from 12 (via 24 and 1) to 11. In the alphabetical subtasks the number of items per letter was balanced (with two items per letter).

The overall empirical data implied that the participants were able to easily derive information about the efficiency of touch screen input from the GUI. The number of steps to solve a subtask could be estimated from the anticipated sequence of system states required to pass. Therefore the user model was provided with the following knowledge: the sequence of states to fulfill a subtask and, for each state, the number of interaction steps to finish the subtask by means of VUI and GUI. For the simulation three tasks from the real experiment were selected in such a way that all experimentally considered levels of speech shortcuts (B_{VUI}, compare Section 2.1.2) were covered. The simulated tasks and the respective sub-goals can be seen in Fig. 4. Referring to Section 3.4, the probability of VUI input P_{VUI} was predicted using $i_{m1} = i_{VUI} = 1$ (constant number of VUI interaction steps) and varying number of GUI interaction steps $i_{m2} = i_{GUI}$ (ranging from 1 to 6). The user model was equipped with the integrative modality choice mechanism trained with all available data (compare Section 4) and knowledge about the optimal path to solve a task[5]. The analysis of the empirical data revealed that the participants only switched modalities on the home screen and on the first list screen. Modality choice was therefore performed only at these screens during the simulation. Incorporating errors considerably increases the implementation effort of the simulation, as different follow-up states have to be considered. Decreased performance of input modalities was therefore not considered in this application example ($er_{VUI} = er_{GUI} = 0$).

Modality-usage data gained from the simulation was compared to human data gathered in Experiment 1, which is based on the corresponding input parameters i_{VUI}, i_{GUI}, er_{VUI} and er_{GUI} (compare Section 2.4). Overall a considerable fit of the model to empirical data can be reported ($R^2 = 0.868$, $RMSE = 0.065$). For both human and model data, Fig. 4 depicts the relation between the percentage of speech usage and B_{VUI}, whereas the latter was varied between the sub-goals[6]. The standard deviation of the model data shows the same relation as human data: the higher the percentage of speech usage, the narrower the standard deviation. The predicted

[5] Due to the simplicity of the task almost no user errors occurred during the experiments.

[6] Note that the distribution of B_{VUI} was well balanced over all of the 15 experimental tasks, although it is not absolutely balanced in task 10.

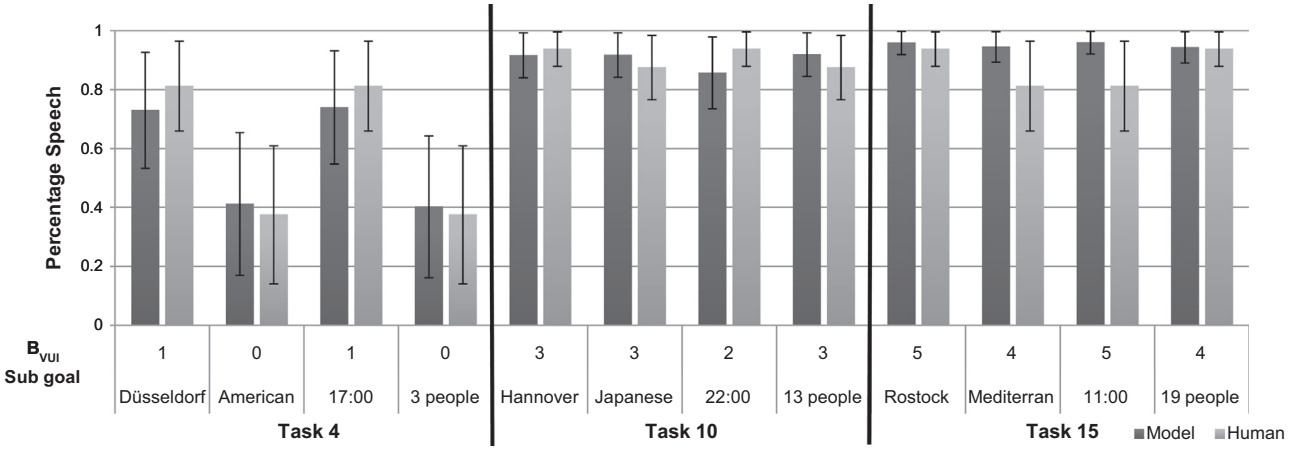

Fig. 4. The relation between B_{VUI} and the percentage of speech usage. Each task contains four subtasks. Each sub-goal corresponds to a specific amount of B_{VUI}.

percentage of speech usage is mostly in line with human data. Therefore, for most subtasks, decisions made on basis of the modality choice model are of use. In the application example, a system designer might draw the same conclusions from the model as from the human data, namely that the VUI is preferred by the user as soon as it becomes more efficient, and that the GUI is selected more often if both modalities are equally efficient. This information could be used to change the interface design. The user could activate speech input if no speech shortcut exists. If speech shortcuts exist, however, speech input could be automatically activated and an appropriate notification could be integrated in the GUI.

However, human and model data are not totally in line. For the numerical subtasks ("time" and "people") noticeable deviation can be observed for two cases: compared to model data, lower speech usage would be expected for human data in the sub-goal "22:00"of task 10 and higher speech usage in the sub-goal "11:00" of task 15. The model at this point strictly adjusts modality usage to B_{VUI}. However the human decision seems to be influenced by additional information. Considering the information the user gains from the task, for both cases human speech usage data is associated with the numerical values of the subtasks and not with B_{VUI}. This refers to an opposing effect of task and system design: in between tasks higher numerical values of sub-goals are not necessarily associated with higher B_{VUI}, as lists within the RBA start with different values ("12:00" for "time" and "1" for "people"). For alphabetical subtasks, a similar tendency could be observed. The observed speech usage points to the fact that the participants' decision for a modality is affected by the alphanumerical information obtained from the set task. However it has to be noted that the described effects of task design were not systematically varied in the studies. Further research will be needed to investigate the significance of the observations.

The modality choice model is so far not able to cover the described task effects. As the model was build using overall data from tasks, including subtasks with diverse combinations of B_{VUI}, task effects should generally average out. An overall comparison between alphabetical and numerical tasks revealed no significant effects regarding modality usage. A system designer using the model to forecast modality choice behavior should confirm possible mismatches of task and system design and carefully select the tasks to simulate.

6. Final conclusion and future work

Future developments in HCI will enable sequential independent multimodal systems (SIMS), thereby enabling free choice of input modalities. Users' modality choice is moderated by various factors. Our motivation was to examine the factors of input performance and modality efficiency and to build a model enabling the prediction of modality usage. The usefulness of the model was demonstrated in an application example in the field of simulation-based AUE.

The empirical results reported in this paper reveal that users of multimodal systems adapt modality usage to estimated modality efficiency as well as to input performance of modalities. On the one hand, VUI input is increasingly preferred if speech gets more efficient in terms of interaction steps. On the other hand, the usage probability of a modality decreases if its input performance is limited (e.g., due to ASR errors or touch screen malfunction). Previous research, mostly in line with our findings, revealed rather discrete insights into the continuum of parameters influencing modality choice (Bilici et al., 2000; Wechsung et al., 2012). With our series of experiments we were able to describe the relationships of multiple factor levels and give a coherent idea about essential moderators of modality choice. The empirical results turned out to be consistent across experiments. For an interface designer the resulting modality usage curves give useful indications about which input modalities might be preferred by a future user of a system, incorporating modality-specific shortcuts as well as error-prone input interfaces. For system errors note that the perceived error rate can be very different from the real error rate and can further change over time. These factors were not considered in the studies outlined in this paper and should therefore be part of future research.

Inspired by the observed user behavior a utility theory-driven model was established, forecasting an average users' modality choice behavior with considerable predictive power. A model comparison revealed that an integrative model that incorporates data about all available input performance conditions is qualified for beneficial estimations of modality usage. Particularly high prediction performances on unseen data and on conditions resting on sparse data indicate reliability of the integrative model. If individual subject data is predicted, substantial variances in individual modality usage profiles lead to decreased accuracy. Individual users appear to have different interaction strategies than those demonstrated by the model. The investigation of modality-usage patterns could expose user groups that differ in interaction strategies. By deploying specialized models for these groups individual differences between users could be taken into account. Further, hedonic quality, context, and other factors like the ones described in Section 1.3.1 provider further areas for the extension of the model.

An application example showed that the model is able to simulate plausible interaction between an average user model

and a system model. Predicted average VUI usage is mostly in line with human data. As the system model was realized as a state machine, the modality-choice mechanism can beneficially extend automatic usability evaluation tools like the MeMo workbench (Möller, et al., 2006). For a multimodal user interface designer AUE simulations provide working knowledge about the modality usage to be expected. As the model acceptably approximates the effects of modality efficiency and input performance, other factors of interests can be brought into the designers' attention. By means of simulation, variants of multimodal interfaces can easily be compared. Optimizing the multimodal interface design in very early stages of system development will save time and monetary costs, as design errors and usability issues can be addressed before a real prototype is available without user testing. Further interaction log data enables usability predictions for future multimodal systems. Relevant information for the typical system design question "(where/how) should VUI be integrated?" can be gathered. So far the model lacks a rigorous explanation of how the benefit of speech usage is derived from the task information and the interface. Effects of subtasks on each other and the interplay between task and system should be part of future research.

The wide field of multimodal systems offers several possibilities for improving the model. First, an expansion to other modalities will be needed, since, for example, non-contact gesture or gaze interaction and other input methods will increasingly emerge in the future. From a technical point of view the model is able to deal with arbitrary modalities, as long as it can be assumed that perceived utility determines modality usage. Looking beyond the simplified list-browsing task, novel interaction techniques like flick gestures allowing quickly scrolling through lists have to be taken into account. Furthermore, the number of input modalities should be adjustable within the model. However, much more data will be needed to fit these special conditions. Second, the combination of modalities, demonstrated by Bolt's long established "put-that-there" paradigm (Bolt, 1980), should be considered. Combining modalities will disclose a vast number of new conditions for the model that are feasible if the necessary data for parameter fitting is available. A difficulty will be that interaction steps are not easy to calculate for modern user interfaces integrating advanced interaction techniques. Third, in the field of automatic usability evaluation, interaction with new system models and unknown users is typically simulated. The model's extrapolation performance to other systems has so far not been tested. The Restaurant Booking Application (RBA) was a prototype use case and portability to other systems and tasks has to be demonstrated to disclose the validity of the model. Note as well that the user interface of the RBA is not the most efficient one. More efficient interfaces for both modalities could be provided by using different and optimized GUI components and a more natural speech interface. For improving the efficiency of interaction, multimodal systems often integrate mechanisms for multimodal error correction or context-specific ASR grammars, concepts so far not covered by the model. Furthermore, it should be possible to facilitate the model for other tasks allowing speech shortcuts, such as "keyword typing" or "speaking for searching" inside a database. The question arises if a data driven approach will scale in the future. One possibility to overcome this issue could be to investigate whether the utility-driven approach can be supported by insights that refer to cognitive modeling and other moderators of modality choice.

Acknowledgments

The research is funded by the German Research Foundation (DFG) as part of the Research Training Group "Prospective Design of Human-Technology Interaction" (No. 1013) and supported by the Telekom Innovation Laboratories. We would like to thank all colleagues at T-Labs for their valuable discussions, Matthias Rath for his impact on the described work as well as Benjamin Jöckel and Michael Minge for their help during the experiments.

References

Anderson, J.R., Matessa, M., Lebiere, C., 1997. Act-r: a theory of higher level cognition and its relation to visual attention. Hum.–Comput. Interact. 12, 439–462.

Apple, 2011. Siri, [Online]. Available : ⟨http://www.apple.com/ios/siri/⟩ (25.01.13).

Bellamy, R., John, B., Kogan, S., 2011. Deploying cogtool: integrating quantitative usability assessment into real-world software development. In: Proceedings of the 33rd International Conference on IEEE Software Engineering (ICSE), pp. 691–700.

Bernsen, N.O., 2008. Multimodality theory. In: Tzovaras, D. (Ed.), Multimodal User Interfaces. Springer, Berlin Heidelberg, pp. 5–29.

Bevan, N., 1995. Usability is quality of use. Adv. Hum. Factors Ergon. 20, 349–354.

Bilici, V., Krahmer, E., te Riele, S., Veldhuis, R.N., 2000. Preferred modalities in dialog systems. In: Proceedings of the Interspeech, pp. 727–730.

Bohn, J., Coroama, V., Langheinrich, M., Mattern, F., Rohs, M., 2005. Social, economic, and ethical implications of ambient intelligence and ubiquitous computing. Ambient Intell., 5–29.

Bolt R.A., 1980. "Put-that-there": voice and gesture at the graphics interface, In: Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '80), ACM, pp. 262–270.

Card, S., Mackinlay, J., Robertson, G., 1990. The design space of input devices. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Empowering People, ACM, pp. 117–124.

Chow, Y.S., Teicher, H., 2003. Probability Theory: Independence, Interchangeability, Martingales. Springer, New York.

Dey, A.K., Häkkilä, J., 2008. Context-awareness and mobile devices. User Interface Design and Evaluation for Mobile Technology 1, 205–217.

Engelbrecht, K.P., Kruppa, M., Mller, S., Quade, M., 2008. Memo workbench for semi-automated usability testing. In: Proceedings of the 9th Annual Conference of the International Speech Communication Association (Interspeech '08), pp. 1662–1665.

Fiscus, J., Ajot, J., Garofolo, J., 2008. The rich transcription 2007 meeting recognition evaluation. Multimodal Technologies for Perception of Humans, 373–389.

Franz, A., Henzinger, M., Brin, S., Milch, B., 2006. Voice interface for a search engine. US Patent 7,027,987.

Fu, W.T., Gray, W.D., 2006. Suboptimal tradeoffs in information seeking. Cogn. Psychol. 52, 195–242.

Gray, W., Sims, C., Fu, W., Schoelles, M., 2006. The soft constraints hypothesis: a rational analysis approach to resource allocation for interactive behavior. Psychol. Rev. 113, 461.

Hassenzahl, M., Burmester, M., Koller, F., 2003. AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In Mensch & Computer 2003, 187–196.

Hone, K.S., Graham, R., 2000. Towards a tool for the subjective assessment of speech system interfaces (SASSI). Nat. Lang. Eng. 6, 287–305.

ISO, W., 1998. 9241-11. Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs), The International Organization for Standardization.

Ivory, M.Y., Hearst, M.A., 2001. The state of the art in automating usability evaluation of user interfaces. ACM Trans. Comput.–Hum. Interact. 33, 470–516.

John, B.E., Prevas, K., Salvucci, D.D., Koedinger, K., 2004. Predictive human performance modeling made easy. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 455–462.

John, B.E., Kieras, D.E., 1996. Using GOMS for user interface design and evaluation: which technique? ACM Trans. Comput.–Hum. Interact. (TOCHI) 3, 287–319.

Jordan, P.W., 2002. Designing Pleasurable Products: An Introduction to the New Human Factors. Taylor & Francis, London.

Kahneman, D., 1973. Attention and Effort. Prentice-Hall, Englewood Cliffs, NJ.

Kahneman, D., 2011. Thinking, Fast and Slow, Farrar, Straus and Giroux.

Kieras, D., Wood, S., Meyer, D., 1997. Predictive engineering models based on the EPIC architecture for a multimodal high-performance human–computer interaction task. ACM Trans. Comput.–Hum. Interact. 4, 230–275.

Kühnel, C., Westermann, T., Weiss, B., Möller, S., 2010. Evaluating multimodal systems: a comparison of established questionnaires and interaction parameters. In: Proceedings of the 6th Nordic Conference on Human–Computer Interaction: Extending Boundaries, ACM, pp. 286–294.

Metze, F., Wechsung, I., Schaffer, S., Seebode, J. and Möller, S., 2009. Reliable evaluation of multimodal dialog systems. In: Proceedings of the 13th International Conference on Human–Computer Interaction (HCII '09), pp. 75–83.

Möller, S., Kühnel, C., Weiss, B., 2011. Parameters describing the interaction with multimodal dialog systems. ITU-T Contribution.

Möller, S., Englert, R., Engelbrecht, K., Hafner, V., Jameson, A., Oulasvirta, A., Raake, A., Reithinger, N., 2006. MeMo: towards automatic usability evaluation of spoken dialog services by user error simulations. In: Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech '06), pp. 1786–1789.

Morrison, J.E., 2003. A review of computer-based human behavior representations and their relation to military simulations. Technical Report DTIC Document.

Nigay, L. and Coutaz, J., 1993. A design space for multimodal systems: concurrent processing and data fusion. In: Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems, pp. 172–178.

Perakakis, M., Potamianos, A., 2008. Multimodal system evaluation using modality efficiency and synergy metrics. In: Proceedings of the 10th International Conference on Multimodal Interfaces, ACM, pp. 9–16.

Rauterberg, M., 1996. A Petri net based analyzing and modeling tool kit for logfiles in human computer interaction. In: Proceedings of the Cognitive Systems Engineering in Process Control, pp. 268–275.

Reisig, W., Rozenberg, G., 1998. Lectures on Petri Nets I: Basic Models: Advances in Petri Nets. Springer.

Ren, X., Zhang, G., Dai, G., 2000. An experimental study of input modes for multimodal human–computer interaction. Adv. Multimodal Interfaces—ICMI 2000, 49–56.

Rudnicky, A.I., 1993. Mode preference in a simple data-retrieval task, Association for Computational Linguistics. In: Proceedings of the Workshop on Human Language Technology, Association for Computational Linguistics, pp. 364–369.

Salvucci, D.D., 2009. Rapid prototyping and evaluation of in-vehicle interfaces. ACM Trans. Comput.–Hum. Interact. 16, 9–33.

Schaffer, S., Jöckel, B., Wechsung, I., Schleicher, R., Möller, S., 2011a. Modality selection and perceived mental effort in a mobile application. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association, Interspeech 2011, pp. 2253–2256.

Schaffer, S., Minge, M., 2012. Error-prone voice and graphical user interfaces in a mobile application. In: Proceedings of the ITG Symposium on Speech Communication 10, pp. 135–139.

Schaffer, S., Reitter, D., 2012. Modeling efficiency-guided modality choice in voice and graphical user interfaces. In: ICCM 2012 Proceedings, p. 253.

Schaffer, S., Schleicher, R., Möller, S., 2011b. Measuring cognitive load for different input modalities, 9. Berliner Werkstatt Mensch-Maschine-Systeme, pp. 287–292.

Schleicher, R., Wechsung, I., 2012. Modelling modality choice using task parameters and perceived quality. In: Proceedings of ITG Symposium on Speech Communication, 10, pp. 1–4.

Sheridan, T.B., Parasuraman, R., 2000. Human versus automation in responding to failures: an expected-value analysis. Hum. Factors: J. Hum. Factors Ergon. Soc. 42, 403–407.

Suhm, B., Myers, B., Waibel, A., 1999. Model-based and empirical evaluation of multimodal interactive error correction. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: the CHI is the Limit, ACM, pp. 584–591.

Sweller, J., 1988. Cognitive load during problem solving: effects on learning. Cogn. Sci. 12, 257–285.

Turunen, M., Hakulinen, J., Heimonen, T., 2010. Assessment of spoken and multimodal applications: lessons learned from laboratory and field studies. In: Proceedings of the Interspeech, pp. 1333–1336.

Varga, I., Aalburg, S., Andrassy, B., Astrov, S., Bauer, J.G., Beaugeant, C., Gei, Hoge, H., 2002. ASR in mobile phones-an industrial approach. IEEE Trans. Speech Audio Process. 10, 562–569.

Wechsung, I., Engelbrecht, K.-P., Möller, S., 2012. Using quality ratings to predict modality choice in multimodal systems. In: Proceedings of the Interspeech 2012: the 13th Annual Conference of the International Speech Communication Association, ISCA. pp. 1–5.

Wechsung, I., Engelbrecht, K.P., Naumann, A., Möller, S., Schaffer, S., Schleicher, R., 2010. Investigating modality selection strategies, I. Spoken Language Technology Workshop (SLT), 31–36.

Wechsung, I., Engelbrecht, K.-P., Schaffer, S., Seebode, J., Metze, F., Möller, S., 2009. Usability evaluation of multimodal interfaces: is the whole the sum of its parts? In: Proceedings of the 13th International Conference on Human–Computer Interaction, pp. 113–119.

Wickens, C.D., Hollands, J.G., 2000. Engineering Psychology and Human Performance. Prentice-Hall, Upper Saddle River, EUA.