A Model of Shortcut Usage in Multimodal Human-Computer Interaction

Stefan Schaffer[1], Robert Schleicher[2], and Sebastian Möller[2]

[1] Research training group prometei, Franklinstr. 28-29, 10587 Berlin, Germany
Stefan.Schaffer@zmms.tu-berlin.de
[2] Deutsche Telekom Laboratories, TU Berlin, Ernst-Reuter-Platz 7, 10587 Berlin, Germany
{Sebastian.Moeller,Robert.Schleicher}@telekom.de

Abstract. Users of multimodal systems have to choose between different interaction strategies. Thereby the number of interaction steps to solve a task can vary across the available modalities. In this work we introduce such a task and present empirical data that shows that strategy selection of users is affected by modality specific shortcuts. The system under investigation offered touch screen and speech as input modalities. We introduce a first version of an ACT-R model that uses the architectures-inherent mechanisms production compilation and utility learning to identify modality-specific shortcuts. A simple task analysis is implemented in declarative memory. The model reasonably accurate matches the human data. In our further work we will try to get a better fit by extending the model with further influence factors of modality selection like speech recognition errors. Further the model will be refined concerning the cognitive processes of speech production and touch screen interaction.

Keywords: Multimodal HCI, User Modeling, Automated Usability Evaluation.

1 Introduction

The research interest in automated usability evaluation has been rapidly grown in the last years [1,2]. At the same time systems with novel interaction techniques like multimodal input become more and more popular. Thus the simulation of multimodal interaction is of explicit interest for the field of automated usability evaluation.

Simulation frameworks which can directly be used by interaction designers are still rare. Möller et al. [3] introduced the MeMo workbench for semi-automatic usability testing. MeMo incorporates a general user simulation framework, applicable to different kinds and classes of systems, such as spoken dialog systems and graphical user interfaces (GUIs). The workbench supports the design of system models with multiple input modalities, but a mechanism to select a specific modality comparable to the behavior of a real user is still missing. In CogTool [4] the designer is also able to build multimodal system models. But the modeler further has to select manually which modality will be used to solve a task.

Studies on user behavior in multimodal human-computer interaction revealed that multiple factors are involved when a user selects the input modality for the next

V.G. Duffy (Ed.): Digital Human Modeling, HCII 2011, LNCS 6777, pp. 337–346, 2011.

interaction step. Each user of a system has individual attributes. Deploying the MeMo workbench, Jameson et al. [5] distinguish the following user attributes: perceptual and motor capabilities (e.g. visual acuity); relevant prior knowledge and experience (e.g. amount of experience with systems similar to the one under consideration); and dynamic factors like the amount of attention that the user is devoting to the performance of the task. Cognitive theories assume that the cognitive workload perceived during the interaction also affects modality selection [6]. McCrasken and Aldrich [7] developed a theory of workload which implies that for a single interaction step, speech input can be more demanding than touch input: preparing and uttering a speech phrase is assumed to be more straining than the processes of initiating and performing touch screen interaction [8].

Further situational constraints like a noisy environment (e.g. traffic), so called "eyes busy, hands busy" settings (e.g. driving) and the system itself may affect modality selection. Regarding the system, efficiency in terms of the number of interaction steps can vary for different modalities. In our previous work we observed that users selected speech instead of touch screen input, if speech offered a shortcut in terms of interaction steps [9]. In addition, error-proneness of a system and the accuracy of input devices affect the effectiveness of interaction [10,11]. In several studies it was shown that users tend to prefer more accurate modalities [12,13]. Further Naumann et al. [14] revealed that the likelihood of task success influenced modality selection. All these factors taken together play a part in each individual modality selection what makes it hardly possible to model single effects.

The scope of this paper is to build a model for tasks offering shortcuts in terms of interaction steps. We present the Attentive Display study, where human data about modality usage was recorded. More specifically, the *Employee Search Task* offered a shortcut for speech input compared to touch screen input. In contrast, the *Floor Change Task* involved the same number of interaction steps in either modality. Both tasks have been modeled with the cognitive architecture ACT-R [15]. In the following, we will briefly summarize the collection of the empirical data which were then used for modeling.

2 Data Collection: Attentive Display Study

The Attentive Display is a wall-mounted information system for employee and room search tasks in a smart office environment. Different floors are displayed on an interactive map of an office building. The system is controlled via touch screen or a speech interface. All tasks can be done in either modality. The output is always given via GUI.

2.1 Task

In the Attentive Display study participants performed six different tasks [16]. Our modeling efforts so far focused on two specific tasks: the *Employee Search Task* and the *Floor Change Task*.

The *Floor Change Task* can be accomplished with one interaction step in both modalities. Using speech, a floor change can be performed by saying a command like

"Go to floor eighteen". After processing speech input the map of floor eighteen is displayed. Using touch screen a specific floor button has to be pressed to accomplish the task (cf. Fig.1).

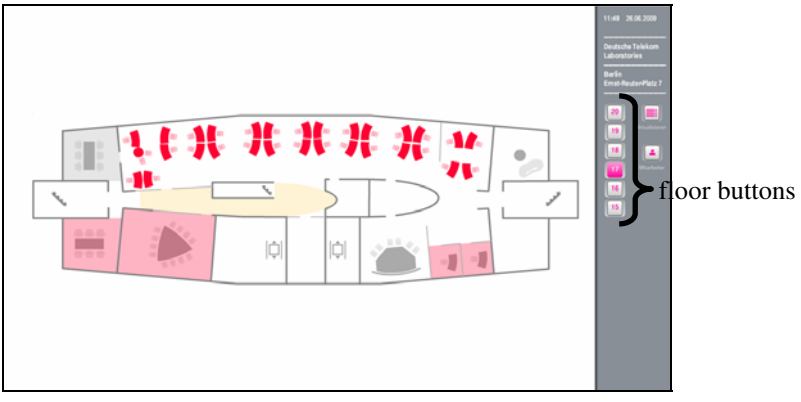

Fig. 1. The GUI of the Attentive Display. After pressing a floor button the map of a specific floor is displayed.

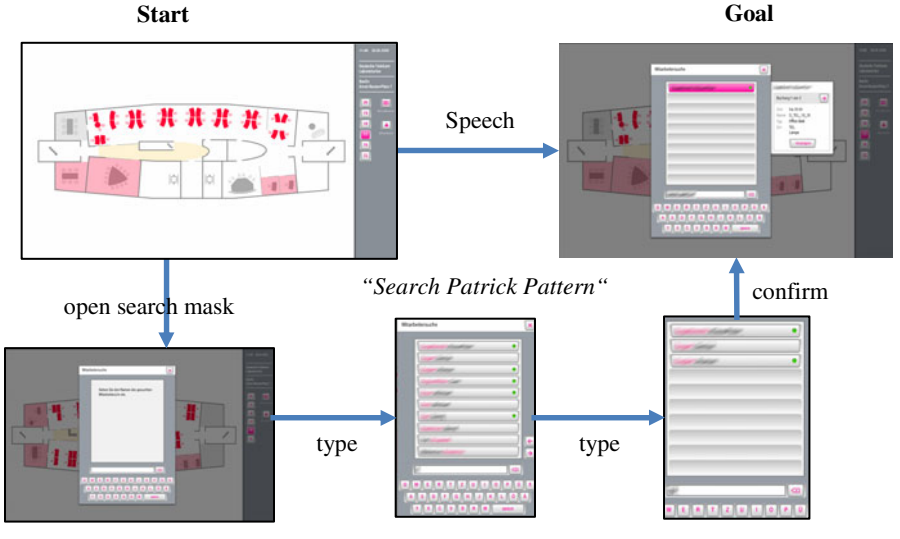

Fig. 2. The *Employee Search Task*. Speech input: a direct transition from the start state to the goal state is possible by saying a keyword in combination with an employee's name. Touch screen input: press the button to open the employee search mask; type the name; press the button with the desired name (the information with the employee's bookings pops up).

Fig. 2 shows an example for the *Employee Search Task*. To search an employee by means of the touch screen, a user has to open the search mask, type a name and confirm the entry. While typing a name, a list with all names matching the letters

entered so far is shown. The more letters are correctly entered, the shorter the list gets. The list consists of buttons with employee names as textual labels. If a button is pressed, information about the current workplace and room bookings of this employee is displayed. Via speech input direct search is possible by saying a keyword in combination with an employee's name: e.g. "Search Patrick Pattern". The system straightly displays the information of the employee.

2.2 Participants

Thirty-six German-speaking participants (17 male, 19 female) between the age of 21 and 39 (M=31.24, SD=3.45) took part in the study. A single experiment took approximately one hour. Participants received a remuneration of € 10.

2.3 Procedure

At first demographic data was gathered using a questionnaire. After that the system was explained and the usage of touch and speech was demonstrated. Next the participants obtained detailed information about the system in written form. Open questions were answered, if the information was not sufficient. The real test comprised 3 blocks: (1) speech, (2) touch, and (3) multimodal. To consider learning effects the first two blocks were permuted after each participant. In the multimodal run the preferred modality could be selected in each interaction step. As the multimodal block was always the last, the participants had prior experience with both modalities in this block.

Within each block the same six tasks were presented. Table 1 shows the interaction steps (IS) the participants could derive from the instructions for the *Employee Search Task* and the *Floor Change Task*.

Table 1. Interaction steps of the *Employee Search Task* and the *Floor Change Task*

Task	Modality	Interaction Step
Floor Change Task	Touch	1. Press the button of the desired floor
	Speech	1. Say "Go to floor <floor number>"
Employee Search Task	Touch	1. Press the button "Employees" 2. Type first letter of employee name 3. Type second letter of employee name[1] 4. Press the button labeled with the desired name
	Speech	1. Say "Search <first name><last name>"

2.4 Results

As the *Floor Change Task* (FCT) can be conducted with one interaction step in either modality, its profit P_{FCT} of speech usage amounts to zero:

[1] Here the minimum number of necessary interaction steps is derived from the instructions. It was also possible to type the whole name before pressing the button labelled with the desired name.

$$P_{FCT} = IS_{Touch} - IS_{Speech} = 1 - 1 = 0 . \quad (1)$$

Speech input was selected by 22.2% of the participants. A χ^2-test revealed significant differences between both modalities (χ^2 (1, N=36) = 11.11, p=.001).

The profit P_{EST} of speech usage for the *Employee Search Task* (EST) amounts to three:

$$P_{EST} = IS_{Touch} - IS_{Speech} = 4 - 1 = 3 . \quad (2)$$

Here, speech usage in 75% of the cases could be observed. Again a χ^2-test revealed significant differences between both modalities (χ^2 (1, N=36) = 9, p=.004).

2.5 Discussion

The study indicates that the profit of speech interaction affects modality selection. Speech input was strongly preferred in the *Employee Search Task*. The shortcut via speech was used. Users appeared to seek efficient interaction strategies.

Different results were observed for the *Floor Change Task.* An equal amount of IS for both modalities implied that usage frequency should be equal for both modalities. However, the data showed that users rather tended to use touch input in this condition. A possible explanation for this pattern will be given in the general discussion in section 5.

3 ACT-R Model

A simple task analysis was implemented in the declarative memory of the model. An example for the *Employee Search Task* is shown in Table 2. Each interaction step is represented by a chunk with a unique name (e.g. "VOICE_1"). In the *step* slot the current interaction step is encoded. The *action* slot specifies which modality will be used to perform system input. The *next* slot defines which chunk will subsequently be retrieved to go on within a task.

Table 2. Representation of interaction steps of the *Employee Search Task* in the declarative memory of the model. Each line in the table represents a chunk encoding an interaction step.

Declarative Memory
VOICE_1 step ONE action SPEECH next **FINISH**
TOUCH_1 step ONE action TOUCH next TWO
TOUCH_2 step TWO action TOUCH next THREE
TOUCH_3 step THREE action TOUCH next FOUR
TOUCH_4 step FOUR action TOUCH next **FINISH**
FINISH step **FINISH** action FINALIZE next ONE

In the current version the models procedural memory consists of production rules for retrieving instructions ("retrieve"), encoding instructions ("encode") and finishing the task ("finish"). Fig. 3 shows how the production rules are linked.

At the beginning of a new task the model always retrieves the chunk with the value "ONE" in the *step* slot. For touch screen usage this is the chunk "TOUCH_1". The value "TWO" is derived from the *next* slot. That means that the chunk to be subsequently retrieved should have the value "TWO" in its *step* slot. The information in the *action* slot is currently not affecting the model behavior. It will be used in upcoming versions to consider cognitive processes of touch and speech interaction and visual search. The models retrieval of interaction steps goes on, until the finish chunk is decoded. When the production rule "finish" fires, the model realizes that the task is processed to the end. For speech usage the chunk "VOICE_1" is retrieved at first. The *next* slot of this chunk directly guides to the "FINISH" chunk. When this chunk is decoded, the "finish" production fires again and task processing ends.

To model the *Floor Change Task* only slight changes have to be made. The chunks "TOUCH_2", "TOUCH_3" and "Touch_4" can be deleted and the *next* slot of the chunk "TOUCH_1" has to be changed to the value "FINISH". The speech interaction steps do not have to be changed in this version of the model.

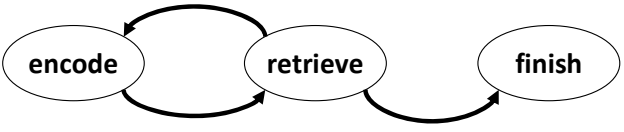

Fig. 2. The procedural memory of the model

The declarative instruction chunks are associated through the *step* and *next* slots. This representation features a practical flexibility, which can be used for simulating multimodal interaction with ACT-R. If two different chunks with the same *step* slot are added to declarative memory, different interaction strategies (chunks) can be retrieved (e.g. "TOUCH_1" and "VOICE_1"). Further different subsequent chunks can be defined in the *next* slots. Thereby it has to be taken into account that chunks with the same value in the *step* slot are usually chosen randomly. Hence the model does not reproduce human behavior. We solved this problem by making use of the ACT-R inherent mechanisms production compilation and utility learning. The production compilation mechanism combines two productions into one new rule (production) and substitutes retrievals from declarative memory directly into the new rule. Thus specialized productions for speech and touch interaction are created.

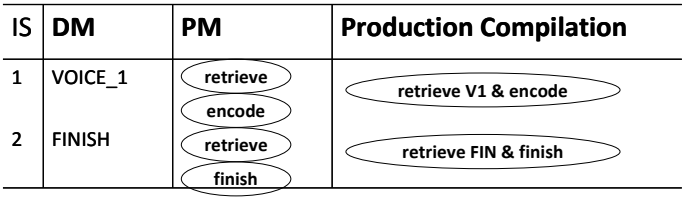

Fig. 3. The effect of production compilation for speech interaction. For each interaction step the "retrieve" production, the "encode" (or "finish") production and a chunk are compiled into one new production.

Utility learning rewards all productions, which are involved in reaching the goal. The total reward is a stated value and spreads over the involved productions. Consequently the reward per production is lower if more rules were involved. By means of these mechanisms we let ACT-R learn the utilities of new production rules. After several model runs the strategy involving less production rules has a higher utility. Hence a modality is used with a higher probability if it offers a shortcut.

4 Results

Fig. 4 and 5 show the usage of speech input of respectively the *Employee Search Task* and the *Floor Change Task*, with the data in dashed lines and the model results in solid lines. The presented model data represents the average speech usage after 1000 independent model runs. Within each model run the task was conducted 100 times.

Within the first approximately (~) 30 trials of the *Employee Search Task* random modality selection can be observed. From ~30 to ~40 trials the utility values of the newly learned speech productions (cf. Fig 3) increase. The increase ends at 61.84%. In the Attentive Display study speech usage of 75% emerged in the empirical Data. The model underestimates modality usage for the *Employee Search Task*. Misjudgment accounts for over 10%.

Within the first ~40 trials of the *Floor Change Task* random modality selection can be observed. After ~40 trials the utility learning ends up and speech usage sets at a level of 51.97%. In the Attentive Display study speech usage of 22.2% emerged in the empirical Data. The model highly overestimates modality usage for the *Floor Change Task*. Misjudgment accounts for nearly 30%.

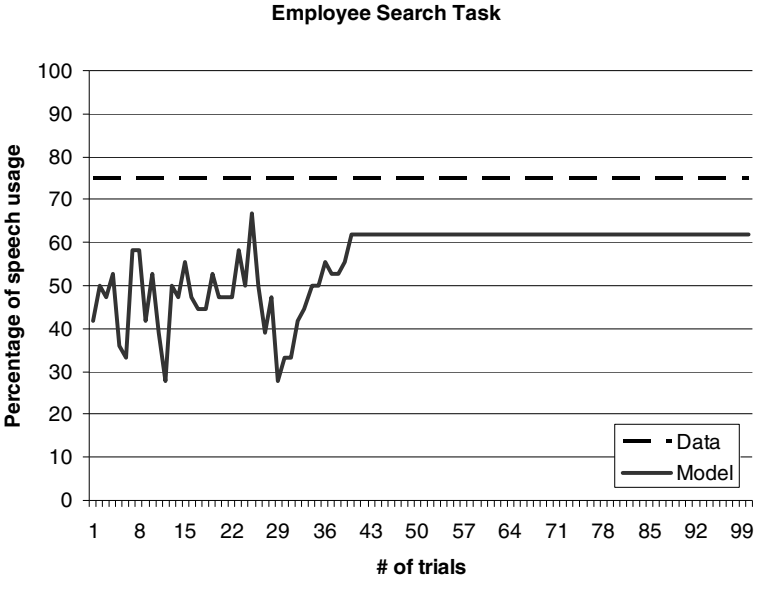

Fig. 4. Modality usage in the *Employee Search Task*

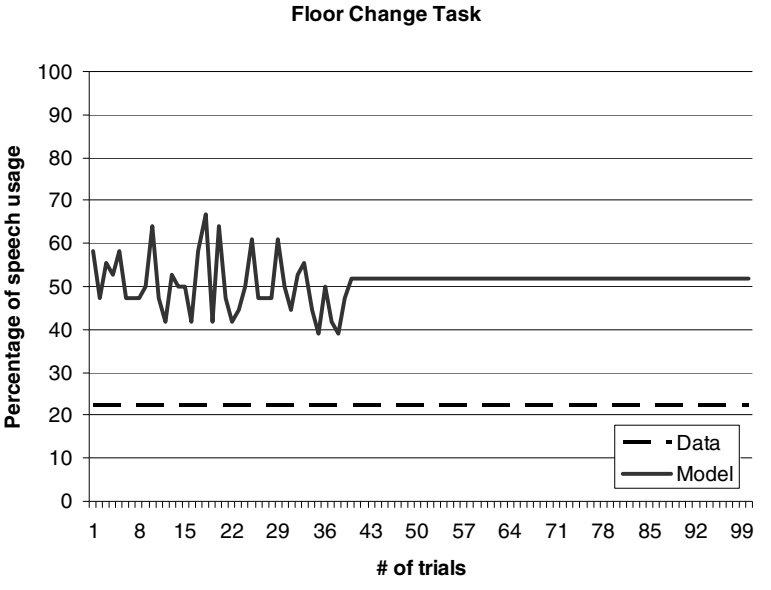

Fig. 5. Modality usage in the *Floor Change Task*

5 Conclusion

The results indicate that that the models' trend to approximate human data is reasonably accurate for the high profit task. The ACT-R inherent mechanisms production compilation and utility learning cause speech usage to increase up to a level of 61.84%. With a better parameter tuning the models' fit to human data should even be improvable for the *Employee Search Task*.

For the *Floor Change Task* the model fails. Respective to our assumptions the level of speech usage adjusted to ~50%. This is the level where modality learning should theoretically end up due to the used ACT-R learning mechanisms. However with a speech usage of only 22.2% human data showed a considerably different value. The results indicate that other factors of modality selection have a higher influence, if the number of interaction steps is equal for both modalities. Personal preferences or situational factors might be of high impact. Other efficiency- related factors might also be relevant. An easy to measure factor is the time that one interaction step takes. In many studies touch was more efficient in terms of time, if one interaction step was needed in both modalities to fulfill the task [17,18]. Up to now the model does not consider the effect of automatic speech recognition (ASR) errors. This type of system errors highly affects the efficiency of interaction. If an ASR error occurs the profit of speech usage might vanish, because the user needs interaction steps to recover from the error. Users might stick at touch input if they cannot benefit from speech usage. In the Attentive Display study 66.8% ASR errors occurred. This high error rate might explain sparse speech usage in the non-profit condition.

It also must be mentioned, that the cognitive processes of language production and touch screen input are difficult to compare. Cognitive theories imply that preparing and uttering a speech phrase might be more loading then performing touch input [7]. But up to now these differences are difficult to measure and no standardized method has established.

In order to make the interaction more realistic, we will further develop the ACT-R model and the Attentive Display simulation. A module for realistic reproduction of errors will be integrated to consider the influence of ASR errors on modality selection. Additionally the model will be refined concerning the cognitive processes of speech production and touch screen interaction

Acknowledgments. The research is funded by the German Research Foundation (DFG - 1013 'Prospective Design of Human-Technology Interaction') and supported by Deutsche Telekom Laboratories.

References

1. Möller, S.: Assessment and Evaluation of Speech-Based Interactive Systems: From Manual Annotation to Automatic Usability Evaluation. Speech Technology, 301–322 (2010)
2. Baker, S., Au, F., Dobbie, G., Warren, I.: Automated Usability Testing Using HUI Analyzer. In: 19th Australian Conference on Software Engineering, ASWEC 2008, pp. 579–588 (2008)
3. Möller, S., Englert, R., Engelbrecht, K., Hafner, V., Jameson, A.: MeMo: Towards Automatic Usability Evaluation of Spoken Dialogue Services by User Error Simulations. System, 1–4 (2006)
4. John, B.E., Salvucci, D.: Multipurpose prototypes for assessing user interfaces in pervasive computing systems. Pervasive Computing 4, 27–34 (2005)
5. Jameson, A., Mahr, A., Kruppa, M., Rieger, A., Schleicher, R.: Looking for Unexpected Consequences of Interface Design Decisions: The MeMo Workbench. In: 19th Australian Conference on Software Engineering, ASWEC 2008, pp. 579–588 (2008)
6. Wickens, C.D.: Multiple resources and performance prediction. Theoretical Issues in Ergonomics Science 3, 159–177 (2002)
7. McCrasken, J.H., Aldrich, T.B.: Analysis of selected LHX mission functions: Implications for operator workload and system automation goals (Technical Note ASI479-024-84). Fort Rucker, AL (1984)
8. Bierbaum, C.R., Szabo, S.M., Aldrich, T.B.: A comprehensive task analysis of the UH-60 mission with crew workload estimates and preliminary decision rules for developing a UH-60 workload prediction model (Technical Report ASI690-302-87[B], Vol I, II, III, IV). Fort Rucker, AL (1987)
9. Wechsung, I., Engelbrecht, K.-P., Naumann, A., Möller, S., Schaffer, S., Schleicher, R.: Investigating Modality Selection Strategies. In: Workshop on Spoken Language Technology (SLT). IEEE, Los Alamitos (2010)
10. Card, S.K., Mackinlay, J.D., Robertson, G.G.: The design space of input devices. In: Proc. SIGCHI 1990, pp. 117–124. ACM Press, New York (1990)
11. Chen, X., Tremaine, M.: Patterns of Multimodal Input Usage in Non-Visual Information Navigation. In: Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS 2006). IEEE, Los Alamitos (2006)

12. Bilici, V., Krahmer, E., Riele, S., Veldhuis, R.: Preferred Modalities in Dialogue Systems. In: Proc. ICSLP 2000, pp. 727–730 (2000)
13. Suhm, B., Myers, B., Waibel, A.: Model-based and empirical evaluation of multimodal interactive error correction. In: Proc. CHI 1999, pp. 584–591. ACM Press, New York (1999)
14. Naumann, A.B., Wechsung, I., Möller, S.: Factors Influencing Modality Choice in Multimodal Applications. In: Perception in Multimodal Dialogue, pp. 37–43 (2008)
15. Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., Qin, Y.: An integrated theory of the mind. Psychological Review 111, 1036–1060 (2004)
16. Schaffer, S., Seebode, J., Wechsung, I., Metze, F., Möller, S.: Benutzerstudien zur Bewertung multimodaler, interaktiver Anzeigetafeln in unterschiedlichen Entwicklungsstufen. In: Kain, S., Struve, D., Wandke, H. (eds.) Workshop-Proceedings der Tagung Mensch und Computer 2009, pp. 22–27. Logos Berlin, Berlin (2009)
17. Perakakis, M., Potamianos, A.: Multimodal system evaluation using modality efficiency and synergy metrics. In: Proc. ICMI 2008, pp. 9–16. ACM Press, New York (2008)
18. Wechsung, I., Naumann, A.B., Möller, S.: Multimodale Anwendungen: Einflüsse auf die Wahl der Modalität. Mensch & Computer, 437–440 (2008)