Modality Selection and Perceived Mental Effort in a mobile Application

Stefan Schaffer[1], *Benjamin Jöckel*[1], *Ina Wechsung*[2], *Robert Schleicher*[2], *Sebastian Möller*[2]

[1] Research Training Group prometei, TU Berlin, Germany
[2] Deutsche Telekom Laboratories, Quality &Usability Lab, TU Berlin, Germany
{stefan.schaffer, ina.wechsung, robert.schleicher, sebastian.moeller}@telekom.de

Abstract

This paper describes a study investigating the influence of efficiency and effectiveness on modality selection and perceived mental effort. Each participant had to perform several tasks with a smart phone application offering touch screen and Wizard-of-Oz speech recognition simulation as input modalities. The results show that efficiency and effectiveness have a strong influence on modality selection. Speech usage increases with increasing efficiency of speech input. A lower effectiveness of speech input raised the threshold for changing the modality selection strategy. For effectiveness mental effort differed significantly between the groups presented with low and high speech recognition errors.

Index Terms: modality selection, multimodal systems

1. Introduction

The aim of multimodal human-computer interaction is to provide multiple parallel or sequential communication channels for a user [1]. This should make a system more flexible, efficient and increase its robustness [2]. If multimodality is offered serially a user can select between different input modalities to interact with a system [3].

A classical metric for the efficiency of interaction is time. However speech input was the preferred modality in [4] though it was less efficient in terms of task-completion time. An explanation for these results might be that speech actually was more efficient in terms of interaction steps. Often a task can be solved with fewer steps in one specific modality [5]. Several studies with multimodal systems showed that more efficient, input modalities were preferably selected to solve a task [6,7]. It has to be noted that the most efficient modality in terms of necessary interaction steps is not necessarily most efficient in terms of task duration [5].

Additionally to efficiency, effectiveness was shown to be of influence for modality selection strategies [8]. The effectiveness of interaction depends on parameters like the error-proneness of a system and the accuracy of input devices [9,10]. In a direct comparison between speech and keyboard input Bilici et al. identified automatic speech recognition (ASR) errors as primary source for modality switches [8]. In a similar context Suhm et al. observed that users tend to interact with the more accurate modality after repeated task execution [11]. These findings imply that users switch the modality, if the recently used modality is clearly more error-prone than the alternative input modality.

One major impact factor of effectiveness is ASR quality. Many ASR systems still do not perform satisfactorily. Poor acoustic conditions, diverse speakers and large vocabularies can lead to high Word Error Rates (WER). The WER is the sum of transcription errors, (word substitutions, deletions, and insertions) divided by the number of reference words whereas lower scores indicate a better performance. In the Rich Transcription 2007 Meeting Recognition Evaluation the WER for single distant microphones for speech-to-text transcription of conference room meetings amounted to 46.7 % [12]. High error rates are, in this context often, caused by large vocabularies which are necessary to transcribe natural language. In contrary for command and control operations in mobile phones a WER as low as 4.1% can be achieved [13]. In a previous study we tested a multimodal prototype with touch screen and speech input where ASR performed with an error rate of 33.1% [14]. ASR was implemented with a command and control grammar and a single distant microphone. However one task was to search employees by names in a list of 150 entries. The high WER was mainly caused by a relative big vocabulary in the employee search task. The studies disclose that effectiveness of speech input is highly dependent on WER. Hence the influence of this factor should be considered when investigating modality selection.

Cognitive theories assuming multiple modalities for human information processing are available [15] and are practically used for human performance modeling [16]. In consistence with Wickens theory [15] of multiple resources McCrasken and Aldrich [17] developed a theory of workload which implies that for one interaction step for a certain task speech input can be more demanding than touch screen input: The cognitive processes of preparing a speech phrase are assumed to be more straining than the processes of initiating touch screen interaction [18]. Considering an interface with graphical output and touch and speech as input modalities [18] implies: if a task can be solved with one interaction step in both modalities, touch should be preferred as it involves lower workload.

In this study the influence of efficiency, measured by the number of interaction steps to solve a task, and the effectiveness, varied by two different levels of ASR errors, is investigated. Referring to the research presented above the following assumption is made: (H1) the preference of one modality over another will increase with its efficiency [5-7]; (H2) the preference of one modality over another will decrease with a increasing ASR error rate [8,9,11]. It is assumed that participants switch to speech interaction for tasks with more interaction steps. This could have a positive effect on workload: the summarized load of several touch interactions could be higher than the load of one speech interaction. On the contrary the workload of speech input increases if more interactions steps have to be performed due to system errors [17]. Regarding the perceived mental workload the following is hypothesized: (H3) the variation of efficiency has no effect on perceived mental workload, (H4) perceived mental workload increases with increasing ASR errors.

2. Method

2.1. Participants

Thirty-three, German-speaking subjects (mostly students) participated in this study. Three users were excluded from further analysis, because they did not follow the instructions. ASR errors were generated by means of an error simulation module which randomly causes error rates between 0 and 40 percent. Thus each participant had an individual error rate. The participants were clustered into two groups with equal error intervals. One test person had to be excluded due to a failure of the error simulation resulting in too many errors. 17 subjects were presented with the lower error rate between 0 and 20 percent ($M_{0\text{-}20}$=9.13, $SD_{0\text{-}20}$=6.0) and 12 with the higher error rate between 21 and 40 percent ($M_{21\text{-}40}$=30.23, $SD_{21\text{-}40}$=4.28). The error rates were set in accordance with typically observed ASR error rates (cf. section 1). The mean age of the remaining 11 female and 18 male participants is 25 years (SD_{age}=3.70).

2.2. Material

2.2.1. System

A smart phone-based restaurant booking system with touch and speech as input modalities was tested. The mobile device (G1 HTC Dream) was running the Android operating system. Automatic speech recognition (ASR) was implemented by operating state changes of the system by an unseen human being (Wizard of Oz setting). The wizard performed speech interaction steps by means of a specially-designed interface. The connection between the wizard interface (a Java application on a UNIX notebook) and the mobile device was established using TCP-IP and wireless LAN. Hence wizard and subjects could easily be placed in different laboratory rooms and subjects believed the system to be autonomous. The participants assumed that ASR worked with an open microphone. Thus it was not necessary to push a button to enter a speech command. The commands were transmitted to the wizards' headphone over a hidden microphone. The ASR error rate was automatically modified in a between subject design. The graphical and voice user interface language was German.

In the restaurant booking system database requests consisting of a name of a city, a culinary category, a desired time and the number of persons are made. All user entries are entered via lists. The system contains three different kinds of screens: start screen, list screens (city, category, time, persons) and end screen (cf. Figure 1). At the start screen either a distinct list is selected or, if all necessary information is entered, the request is sent to the server. The latter is only possible if all fields (city, culinary category, time, number of persons) are filled. At a particular list screen a desired item can be selected. At the end screen the request is confirmed and a new request can be started. Each list screen contains 6 layers each with 4 items. The transition between layers is performed with touch or speech input. An item is selected by touch or by saying the written text label. The items are ordered alphabetically or numerically. Thus the layer of a distinct item can be anticipated by a user: e.g. the item to select 12 persons is located at layer 3. The number of the layer plus 1 (pressing a list button on the start screen) equals the minimum number of touch screen interaction steps.

2.2.2. Benefit of Speech Input

To select any desired item in a list a user can say the items label even if the item is currently not displayed in the list. Hence speech commands are not limited to the offered items on the actual graphical user interface. By using speech input it is possible to choose the item for 12 persons, which is located on layer 3, although layer 1 is currently displayed. Thus, speech shortcuts allow for saving interaction steps. In this example via touch screen a user would need 3 interaction steps to select an item on layer four: (1) browse from layer 1 to layer 2, (2) browse from layer 2 to layer 3, (3) select the desired item on layer 3.

By using speech input (saying the label of the desired item) this subtask is solvable with only one interaction step. The benefit of speech interaction steps B_{speech} thereby can be calculated as the difference between necessary touch screen interaction steps IS_{touch} and speech interaction steps IS_{speech}.

$$B_{speech} = IS_{touch} - IS_{speech} \qquad (1)$$

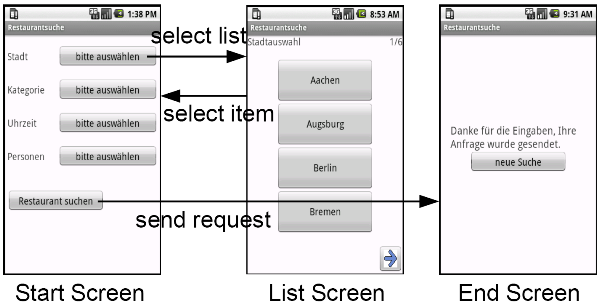
2.2.3. Task

The participants' task was to perform database requests with the restaurant booking system. Once an item has been selected the start screen is presented again. If a speech recognition error was induced, the start screen was presented with no result and the textual message "I did not understand". Then the search within the list screen had to be started again. Each task contained four subtasks with systematically randomized levels of benefit.

2.3. Procedure

A single experiment took approximately 45 minutes. Participants received a remuneration of € 10. At first demographic data was gathered using a questionnaire. Next the system was explained and the usage of touch and speech was demonstrated. After that the participants performed three training trails: touch usage only, speech usage only and multimodal with mixed modality usage. The real test comprised 15 trials. The tasks were presented in written and oral form (e.g. "Please look for a Chinese restaurant in Berlin at 8 pm for 12 persons"). The benefit of the speech modality was systematically varied between 0 and 5 interaction steps. A trial was finished, if all specified information was collected correctly and the request was sent to the server.

After each task the participants were asked to rate their perceived mental effort on the SEA-scale [19].



Danke für die Eingaben, Ihre Anfrage wurde gesendet.

Figure 1: *The three different screens of the restaurant booking system.*

3. Results

The aim of the interface design was to enable interaction as intuitive and easy to understand as possible. As the list items are ordered alphabetically or numerically the speech benefit of a certain item could easily be anticipated. To further minimize effects of learning and adaption the first two tasks were excluded in the further analysis.

In the study potential differences between the subtasks concerning the difficulty of anticipation were undesirable. To ensure that no differences between the subtasks occurred, a paired t-test was calculated for the perceived mental effort (SEA-scale ratings) and modality selection. Significant differences were not observed. Hence both task groups were similarly difficult.

3.1. Modality Selection

3.1.1. *The Influence of Efficiency*

Within the list screens the benefit of speech usage was calculated for each interaction step to investigate if speech shortcuts influence modality selection. The benefit equals zero if speech interaction compared with touch interaction offers no advantage. In this calculation the transition from the start screen to the list screen was not considered, because this state only offers interactions with a benefit of zero for speech and touch screen input. The analysis of variance with repeated measures showed an highly significant effect of benefit on speech usage (F(2.47,65.79)=74.222; $p_{1\text{-tailed}}$<.001; part.eta^2=.733). H1 could be confirmed. If the desired item was to be found in a deep-set list (indicating a high benefit of speech) the proportion of speech usage went up.

Post-hoc analyses using Bonferroni correction showed no significant differences between the levels of benefit three and four, three and five, and four and five. Between all other levels the difference was significant (p<.05).

3.1.2. *The Influence of Effectiveness*

To investigate the effect of effectiveness, modality selection was analyzed over all tasks in a between-subject design. Two different error levels were examined. An ANOVA revealed significant differences between the two error rates (F(1, 27)=6.94; p<.007; part.eta^2=.204). The more ASR errors arise the less speech is used. The threshold where participants tend to use more speech input then touch screen input shifts from 1 to 2 interaction steps (cf. Figure 2). H2 could be confirmed. The worse the ASR performance, the later the switch from touch screen to speech.

3.1.3. *Interaction between Efficiency and Effectiveness*

It was shown that both the number of interaction steps (efficiency) and ASR errors (effectiveness) have a significant impact on modality selection. But an significant interaction between both factors could not be observed (F(2.47,65.79)=74.222; p<.11; part.eta^2=.054).

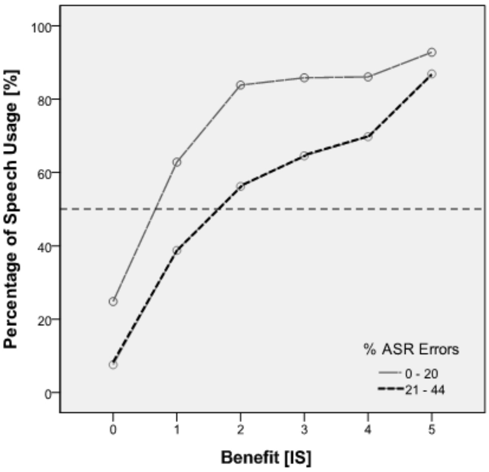


Figure 2: *The influence of efficiency and effectiveness on modality selection. The threshold for modality switches from touch usage to speech usage shifts with an increasing number of errors from benefit b=1 to b=2.*

3.2. Perceived Mental Effort

3.2.1. *The Influence of Efficiency*

For each level of benefit the participants conducted one task containing only subtasks with an equal amount of benefit. To investigate if efficiency has an effect on perceived mental effort differences between these six tasks were examined.

Respective to our hypothesis the analysis of variance with repeated measures showed no significant effect of benefit on perceived mental effort (F(3.44, 92.94)=1.000; $p_{1\text{-tailed}}$<.202; part.eta^2=.036). H3 could be confirmed. If the desired item had to be found in a deep-set list (indicating a high benefit of speech) the perceived mental effort did not increase.

3.2.2. *The Influence of Effectiveness*

To investigate the effect of effectiveness modality selection was analyzed over all tasks in a between-subject design. Two different error levels were examined. An ANOVA revealed significant differences between the two error rates (F(1,27)=13.50; p<.001; part.eta^2=.333). H4 could be confirmed. The more ASR errors arise the higher is the perceived mental effort (cf. Figure 3).

3.2.3. *Interaction between Efficiency and Effectiveness*

No interaction could be observed between effiency and effectiveness (F(3.44, 92.94)=.357; p<.405; part.eta^2=.013. The SEA ratings alternate about 60 in the high error condition which means that task solving was just under “rather demanding”. In the low error condition the ratings alternate about 20, which means that task solving was “hardly demanding”.

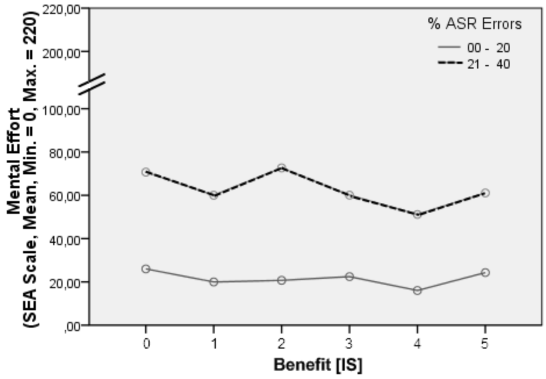


Figure 3: *The influence of efficiency and effectiveness on perceived mental effort. The SEA values are higher for the high error condition. For the variation of efficiency no significant changes can be observed.*

4. Discussion and Conclusion

The results confirm previous findings [5,6,8,11] indicating that modality selection is influenced by the efficiency of the modality. If desired items are to be found in a deep-set layer of a list, speech usage was preferred over touch input due to shortcuts only available by speech. Apart from that users tend to use touch if the benefit of speech equals zero. The preference of touch input in this case indicates that other influencing factors come into play if only one interaction step is necessary for both modalities. As speech still can be considered as a novel input modality the familiarity of touch interaction might be one important factor for modality selection if speech does not offer a shortcut. However the interaction behavior could change if speech input is established in more interfaces and thus gets more familiar.

Another explanation could be that touch input was more efficient in terms of time. The average duration of one interaction step was longer for speech input. Further cognitive workload of speech processing compared to touch usage is assumed to be higher [17]. The factor time and a further investigation of the cognitive demand of touch and speech will be objectives of our future studies.

It also was shown that speech as an input modality is less used if the probability of ASR errors arises. Users adapt their interaction behavior to the accuracy of a system when they decide which modality they use next. However [22] and [23] observed that users sometimes stay in the same modality for error correction. As the accuracy of future ASR modules might increase to human-like performance the influence of this factor might decrease or even fade away. But interaction designers of presently developed systems have to be aware of the reliability of their input modules.

If a system is affected by ASR errors on the one hand and contains speech shortcuts on the other hand, a user has to distinguish which modality is more capable. The threshold at which speech becomes more efficient shifts with rising error rate and the advantage of speech shortcuts even may completely vanish if the error rate is too high.

Efficiency did not affect the participants' perceived mental effort significantly. We assume that here the advantage of speech shortcuts comes into play. If more interaction steps are needed to solve the task, the increased effort can be compensated by using speech input and thus the mental effort remains almost constant.

On the other hand the perceived mental effort increased with decreasing effectiveness. The more ASR errors occurred the higher was the perceived mental effort. This may be due to error-prone interaction resulting in more interaction steps and in longer task durations. Further, as mentioned above, if the advantage of speech shortcuts vanishes due to errors, the negative effect of tasks containing more interaction steps emerges and this could cause higher SEA ratings.

The interdependencies between factors influencing modality selection are highly complex. Multiple different factors commonly play a role when a user decides what modality will be used next. The study revealed that users seem to be aware of the changing characteristics of certain factors and try to weigh up the best decision. Other studies on multimodal user behavior disclosed that modality selection is also influenced by factors like e.g. expertise [20], the likelihood of task success, or the task itself [21]. For a better understanding of the interplay between relevant factors further research has to be done. Simplified models of modality selection only perform on certain tasks in a laboratory environment and are rarely applicable on real life situations. Otherwise for the research discipline of computational and cognitive user modeling the findings of this work can be implemented in basic algorithms for modality selection and go for empirical reference values of user models.

5. Acknowledgements

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

6. References

[1] F. Chen, "Designing Human Interface in Speech Technology", Berlin: Springer, 2006.

[2] S. Oviatt, "Ten myths of multimodal interaction", Communications of the ACM, vol. 42, 1999.

[3] L. Nigay and J. Coutaz, "A Design Space For Multimodal Systems: Concurrent Processing and Data Fusion. Proceedings of the," INTERCHI', vol. 93, 172-178, 1993.

[4] A.I. Rudnicky, "Mode preference in a simple data-retrieval task", Proc. of the workshop on Human Language Technology, Stroudsburg, PA, USA, 364-369, 1993.

[5] I. Wechsung, K.-P. Engelbrecht, A. Naumann, S. Möller, S. Schaffer, and R. Schleicher, "Investigating Modality Selection Strategies", Workshop on Spoken Language Technology (SLT), IEEE, 2010.

[6] M. Perakakis and A. Potamianos, "Multimodal system evaluation using modality efficiency and synergy metrics", Proc. ICMI '08, ACM Press, 9-16, 2008.

[7] I. Wechsung, A.B. Naumann, and S. Möller, "Multimodale Anwendungen: Einflüsse auf die Wahl der Modalität", Mensch & Computer 2008, 437-440, 2008.

[8] V. Bilici, E. Krahmer, S. te Riele, and R. Veldhuis, "Preferred Modalities in Dialogue Systems", Proc. ICSLP2000, 727-730, 2000.

[9] S.K. Card, J.D. Mackinlay, and G.G. Robertson, "The design space of input devices", Proc. SIGCHI '90, ACM Press, 117-124, 1990.

[10] X. Chen and M. Tremaine, "Patterns of Multimodal Input Usage in Non-Visual Information Navigation", Proc. of the 39th Annual Hawaii International Conference on System Sciences (HICSS 06), IEEE, 123-123, 2006.

[11] B. Suhm, B. Myers, and A. Waibel, "Model-based and empirical evaluation of multimodal interactive error correction", Proc. CHI '99, ACM Press, 584-591, 1999.

[12] J. Fiscus, J. Ajot, and J. Garofolo, "The Rich Transcription 2007 Meeting Recognition Evaluation", Multimodal Technologies for Perception of Humans, R. Stiefelhagen, R. Bowers, and J. Fiscus [Ed], Berlin: Springer, 373-389, 2008.

[13] I. Varga, S. Aalburg, B. Andrassy, S. Astrov, J.G. Bauer, C. Beauegant, C. Geissler, and H. Höge, "ASR in mobile phones: An industrial approach", IEEE transactions on speech and audio processing, vol. 10, 562-569, 2002.

[14] S. Schaffer, J. Seebode, I. Wechsung, F. Metze, and S. Möller, "Benutzerstudien zur Bewertung multimodaler, interaktiver Anzeigetafeln in unterschiedlichen Entwicklungsstufen", Mensch und Computer 2009, S. Kain, D. Struve, and H. Wandke [Ed], Berlin: Logos Berlin, 22-27, 2009.

[15] C.D. Wickens, "Processing resources in attention", Varieties of attention, R. Parasuraman and D.R. Davies [Ed], New York: Academic Press, 63-102, 1984.

[16] J. Keller, "Human performance modeling for discrete-event simulation: workload", Proc. of the 34th conference on Winter simulation: exploring new frontiers, 157-162, 2002.

[17] J.H. McCrasken and T.B. Aldrich, "Analysis of selected LHX mission functions: Implications for operator workload and system automation goals (Technical Note ASI479-024-84)", Fort Rucker, AL: 1984.

[18] C.R. Bierbaum, S.M. Szabo, and T.B. Aldrich, "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.

[19] K. Eilers, F. Nachreiner, and K. Hänecke, "Entwicklung und Überprüfung einer Skala zur Erfassung subjektiv erlebter Antrengung [Development and evaluation of a scale to assess subjectively perceived effort]", Zeitschrift für Arbeitswissenschaft, 215–224, 1986.

[20] C.A. Kamm, D.J. Litman, and M.A. Walker, "From Novice To Expert: The Effect Of Tutorials On User Expertise With Spoken Dialogue Systems", Proc. ICSLP98, 1211-1214, 1998.

[21] A. Naumann and I. Wechsung, "Factors Influencing Modality Choice in Multimodal Applications", Perception in Multimodal Dialogue, 37-43, 2008.

[22] J. Sturm and L. Boves, "Effective error recovery strategies for multimodal form-filling applications", Speech Communication, vol. 45, 289-303, 2005.

[23] S. Oviatt and R. VanGent, "Error resolution during multimodal human-computer interaction", ICSLP-1996, 204-207, 1996.