Modeling Efficiency-guided Modality Choice in Voice and Graphical User Interfaces

Stefan Schaffer (sschaffer@zmms.tu-berlin.de)
Berlin Institute of Technology
Research Training Group prometei, Franklinstr. 28-29
10827 Berlin, Germany

David Reitter (reitter@cmu.edu)
Carnegie Mellon University
Department of Psychology, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Keywords: Multimodal interaction; input modality choice; strategy selection.

Introduction

In multimodal human computer interaction users can often select between specific input modalities. Modality choice is influenced by various factors including user attributes, system attributes, the task and the environment (e.g. Lemmelä et al., 2008). Here, we describe on-going research into cognitive models of input modality selection.

The efficiency to solve a task with a multimodal user interface can vary widely due to modality-specific shortcuts. For instance, comparing touch-screen and speech input, items in lists such as names in a directory can be more efficiently found via speech. The number of list items in a GUI is limited due to screen size and legibility. Using a touch-screen, users have to browse the list for the searched item. With speech, each item can be directly accessed, as the limitations of the GUI do not necessarily affect the voice interface. Thus, subjects find a list item in fewer steps by asking for it verbally. The benefit of speech, B_S, is defined as the difference in interaction steps between touch-screen (IS_T) and speech inputs (IS_S): $B_S = IS_T - IS_S$.

Our aim is to develop models of modality selection to support existing tools for model-based usability evaluation such as MeMo (Möller et al., 2006) or CogTool (John et al., 2004). In a classical usability experiment, a participant is instructed to solve a task with different user interface variants. Taatgen et al. (2006) presented a model where unimodal task knowledge was coded into instructional chunks of the declarative memory of ACT-R (Anderson et al., 2004). We extended Taatgen's concept for multimodal interaction and investigate to which extent our model is able to reproduce the modality selection behavior of real test participants.

Experiment

The Restaurant Booking System (RBS)

A smart phone-based RBS with touch and speech as input modalities was tested (for details, see Schaffer, 2011a). Automatic speech recognition (ASR) was simulated via a Wizard-of-Oz design: an unseen human operator changed the system state. This way, issues related to ASR errors could be avoided.

In the RBS database, requests consisting of a name of a city, a culinary category, a desired time and the number of people are made. All user entries are entered via different lists. Each list 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. However, all list items can also be accessed directly by using speech input. The items are ordered alphabetically or numerically. The benefit of speech input calculates to 0 steps ($B_S = 1 - 1$) for items located at the first layer of a list and increases to 5 steps ($B_S = 6 - 1$) at the last layer of a list.

Task

The participants' task was to perform database requests with the RBS. The benefit of the speech modality was systematically varied between 0 and 5 interaction steps.

Participants

Sixteen German-speaking participants (8 female, 8 male) between the age of 22 and 31 (M=26, SD=2.95) took part in the study. A single experiment took approximately one hour. Participants received a remuneration of €10.

Procedure

The system was explained and the usage of touch and speech demonstrated. Then, participants performed three training trials: touch usage only, speech usage only and multimodal with mixed modality usage. In the target phase, 12 trials with mixed, participant-chosen modality usage followed. The tasks were presented in written form (e.g., "Please find a Chinese restaurant in Berlin at 8 pm for 12 people").

Cognitive Model

Instructional steps are represented in declarative memory as chunks containing pre-condition, post-condition, action and modality. Pre- and post-conditions are used to chain the instructions. A key aspect of the model is that for each modality, instructional chunks with the same precondition occur in declarative memory. An earlier study revealed that for the RBS, speech is perceived to be more demanding than touch input (Schaffer 2011b). Therefore we use an action slot within each instruction to describe the interaction more precisely. One GUI interaction consists of two instructions distinguished by the statement in the action slot (search and

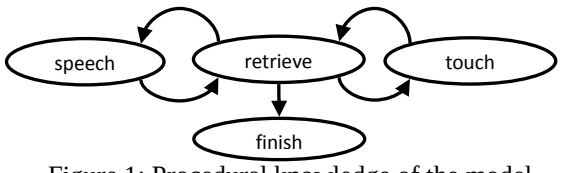

Figure 1: Procedural knowledge of the model.

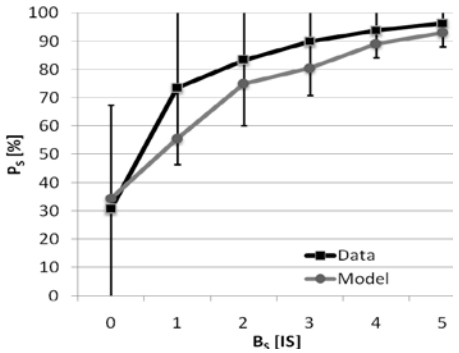

Figure 2: Percentage of speech usage P_S for each level of benefit B_S for human and model data.

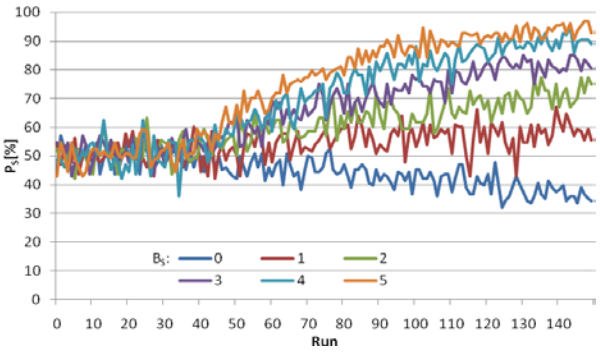

Figure 3: Development of speech usage P_S during 150 model runs for different levels of speech benefit B_S.

press). Speech input consists of three instructions (action slots: search, think and speak).

The general operation of the model is summarized in Figure 1. Instructions are being retrieved from declarative memory. Chunks with the same precondition (but differing modality) are chosen randomly. Retrievals are processed by modality specific production rules. By way of the production compilation mechanism, new production rules with integrated chunks are learned. After each finalization of the task a reward is propagated to the involved productions. Thus, the model adapts to modality success via a reinforcement-learning mechanism.

Results

Figure 2 shows the percentage of speech usage P_S in the human data (black) and the model data (grey). An analysis of variance with repeated measures showed an highly significant effect of B_S on P_S in human data ($F(2.27,33.97)=27.503$; $p_{1\text{-tailed}}<.001$; $part.eta^2=.647$).

Modality usage of the model is comparable to human behaviour. The model performs fairly well at $B_S=0$. For $B_S=1$, 2 and 3 the model fit worsens, whereas for $B_S=3$ and 4 model performance improves again.

Each participant (16) of the experiment executed eight subtasks for each level of B_S. Thus the model data was calculated from the average of 128 particular model iterations. Each iteration included 150 runs. Figure 3 shows the learning behavior for each level of B_S (colored lines).

Conclusion

Taken in context with our aim to design tools for model-based usability evaluation, the model provides a useful basis for a modality selection mechanism. Future work will extend the model to enable interaction with system prototypes and produce actual speech output. As is seen sometimes in reinforcement learning, adaptation seems slower than what is seen empirically. Once a better-fitting model is defined, further evaluation may demonstrate the learning behavior over repeated presentations and time, giving essential cues to the nature of the learning effect as a form of routinization or declarative memorization. One common effect of routinization is that early choices and experiences determine fixed, long-term strategy choices as routinized knowledge is less adaptive (an effect of primacy: first impressions matter). Showing such effects would critically examine the use of adaptive speech recognition technology in end-user applications, specifically if these systems start out with high error rates.

Acknowledgements

The research is funded by the German Research Foundation (DFG - 1013 'Prospective Design of Human-Technology Interaction') and supported by Deutsche Telekom Laboratories. We also thank Michael Minge for his help during the experiment.

References

Anderson, J.R., Bothell, D., Byrne, M., Douglass, D., Lebiere, C. & Qin, Y. (2004). An integrated theory of mind. *Psychological Review*, *111* (4), 1036-1060.

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). New York, NY, USA: ACM.

Lemmelä, S., Vetek, A., Mäkelä, K., & Trendalov, D. (2008). Designing and evaluating multimodal interaction for mobile contexts. In *Proceedings of the 10th Int'l Conf. on Multimodal Interfaces* (pp. 265-272). New York, NY, USA: ACM.

Möller, S., Englert, R., Engelbrecht, K., Hafner, V., Jameson, A., Oulasvirta, A. (2006). Memo: Towards automatic usability evaluation of spoken dialogue services by user error simulations. In *Proc. 9th International Conference on Spoken Language Processing* (p. 1786-1789). Pittsburgh, PA, USA.

Schaffer, S., Jöckel, B., Wechsung, I., Schleicher, R. & Möller, S. (2011a). Modality Selection and Perceived Mental Effort in a Mobile Application. In *Proc. 12th Annual. Conf. International Speech Communication Assoc.*. (p. 2253-2256). Florence, Italy.

Schaffer, S., Schleicher, R., & Möller, S. (2011b). Measuring cognitive load for different input modalities. In *9. Berliner Werkstatt Mensch-Maschine-Systeme* (p. 287-292). Berlin, Germany: VDI Verlag.

Taatgen, N. A., Huss, D., & Anderson, J. R. (2006). How cognitive models can inform the design of instructions. In *Proc. 7th Int'l Conf. on Cognitive Modeling* (p. 304-309). Trieste, Italy: Edizioni Goliardiche.