Benefit, Design And Evaluation Of Multimodal Interaction

Stefan Schaffer
Norbert Reithinger
German Research Center for
Artificial Intelligence (DFKI)
Alt-Moabit 91c, 10559 Berlin,
Germany
stefan.schaffer@dfki.de
norbert.reithinger@dfki.de

Abstract
In this paper we define our position regarding current research questions in the field of designing speech and multimodal interactions for mobile and wearable applications. Our argumentation is organized in three areas reflecting our following basic research questions questions: (1) what are the benefits that can emerge through multimodality, (2) how can these benefits be considered in the system design, and (3) how does multimodality affect the evaluation of HCI. We conclude that multimodality should be considered where the user has a *benefit* of it, that these benefits should be specifically supported by the interface *design* process, and that targeted use of multimodality should have positive impact on the *evaluation* results of HCI.

Author Keywords
Multimodal systems; spoken dialogue systems; Speech science in end-user applications

ACM Classification Keywords
H.5.2. Information interfaces and presentation (e.g., HCI): User Interfaces

Introduction
The intelligent user interfaces group of DFKI (German research centre for artificial intelligence) realizes and investigates human computer interfaces integrating both novel interaction techniques and intelligent data

Selected DFKI-related works in the field of speech and multimodal interaction

2016 – STREETLIFE: mobile app-based route companion with a wearable (smart watch) extension supporting speech based routing requests and on track mobility information.

2015 – Wir im Kiez: multimodal social network Web and Android app with a conversational speech interface for elderly people with support needs in everyday situations.

2014 – LeVer: multimodal cognitive training Web and Android app for older users with and without mild cognitive impairment (MCI).

2013 – Voice2Social: social networks enriched with audio content for POIs generated based on recorded and annotated audio snippets.

processing. Many of our HCI systems are implemented using MMIR, a framework providing a lightweight multimodal dialog manager [5]. As part of our recent research we also use MMIR in combination with wearable devices. The sidebar on the left lists selected research projects of our group. The applications developed within these projects have in common that they are integrating speech and multimodal interfaces.

Besides the development of the MMIR technology one of our current research questions in the filed of multimodal interaction is, what modalities the users prefer in which context and under the effect of certain factors influencing modality selection. Our own research revealed that users tend to utilize specific modalities if their use confers a certain *benefit*, like e.g. shortcuts implemented via speech input or higher input performance of touch screen input compared to speech input [7,8]. During interface *design*, we specifically apply multimodality where we realize such benefits. As part of a user-centred design process we apply an adapted usability inspection method to find out whether, where, and what specific speech commands as well as corresponding system feedback can improve the interaction between user and system [10]. Referring to the *evaluation* of multimodal HCI our further hypothesis is that targeted allocation of multimodality also has a positive effect on the perceived user experience.

In this paper we describe our position concerning various research questions in the field of designing speech and multimodal interactions for mobile and wearable applications. We present our position with regard to benefit, design, and evaluation of multimodal HCI describing a selected track of our research.

Benefit of Multimodal Interaction

One of the breakthroughs users can benefit of is that automatic speech recognition (ASR) improved highly significant over the last years. ASR now works good for dictation tasks. However, dictation is a highly specific use case which does not require the extraction of semantics from the utterances. Some applications use speech input for form filling. However, filling each single slot by speech is often not more efficient then typing. The question arises:

What are important challenges in using speech as a "mainstream" modality?

While ASR made significant efforts within the last years, e.g. partly driven by the successful application of deep neural networks, the identification of the intended semantic for a further processing by the dialog manager is still a rather difficult process. ASR capabilities are easy to integrate into new user interfaces by making use of available programming APIs. On the technical side one of the next challenges is therefore to realise conversational speech interaction in many applications. This requires to simplify the usage of NLP methods for information extraction, dialog processing and presentation, so that developers can easily deploy speech interfaces.

Since the Internet is mobile nowadays and conversational speech is the most convenient interaction mode of complex applications that require more than simple gestures, this will enable even more services at the hand of the users. In that matter it is important to better understand the specific benefits that emerge for individual users. Information about these benefits can be revealed by observing the users' modality choice behaviour. Understanding the factors

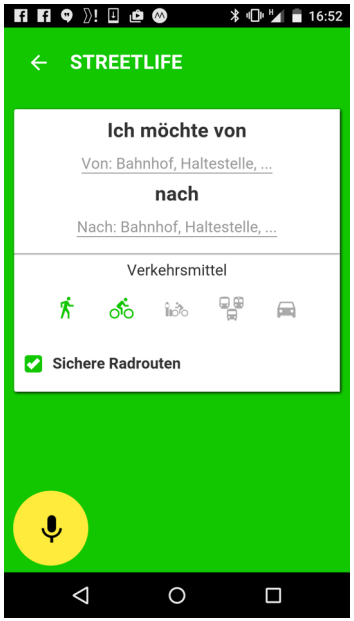

Figure 1: Conversational interface of the STREETLIFE app that was designed to indicate a valid utterance to the user. The GUI labels "Ich möchte von… nach…" (engl. "I want from… to…") reflect the utterance that can be optimally recognized by the grammar.

influencing users' modality choice will enable interface designers to adapt applications to the advantage of the user, and to inform the user about extra possibilities of interaction. Besides the factors modality efficiency and input performance that were extensively investigated by our own research [e.g. 7,8], other factors like mental effort [11], hedonic qualities, and personal preferences have to be better understood [6]. If the concrete influence of these factors is revealed the benefits of which users may take advantage from can be clearly explained to the user and also be used as a marketing strategy. The benefits of multimodal interaction reveal possible answers to the question:

What interaction opportunities are presented by wearable computing?
In our actual research project STREETLIFE, funded by the EU, we developed a smartwatch extension for a smartphone based mobility app [9]. Routing requests from the actual position to a freely selectable destination are enabled via speech input over the smartwatch. Further the user gets real time trip information over the graphical user interface of the wearable. First qualitative inspections of this multi-device interaction with smartwatch and smartphone indicate that users tend to prefer the smartwatch, if it is more efficient compared to the smartphone. With increasing usage of speech input the probability of experiencing ASR errors of course increases. The context of multimodal interaction leads to the question:

How does multimodal processing increase robustness over speech alone, and in what contexts?
With regard to multimodal processing we propose to distribute single information to single modalities. As an example, the STREETLIFE smartphone app integrates more flexible multimodal input than its smartwatch extension does. While the smartwatch version supports bicycle routing only, different traffic modes can be easily selected via touch screen input with the smartphone. In parallel the user can enter origin and destination via speech input (compare Fig. 1). Separating information over the modalities lowers the probability of ASR errors which increases the robustness of the speech interface. In this specific context the information can be distinguished in (1) location information (origin and destination) that has an intrinsic benefit regarding the efficiency of speech input, as speaking names of locations is more efficient then typing these name on a virtual keyboard of the smartphone, and (2) traffic mode information that can be selected by just one tap on the touch screen. For the latter an alternative speech alone input would not increase the efficiency of the user input, but rather decrease the robustness of the interaction. We therefore advocate that efficiency guided distribution of information input supports the overall robustness of the interface. In the near future when speech input gets even more common this paradigm could even evolve into a design pattern of multimodal interaction.

In which other applications has speech highest potential to help with?
We successfully employed speech input in various projects (see the sidebar for the most important). We shortly describe the two most recent ones:

In "**LeVer** - learning against forgetting", funded by the German Federal Ministry of Education and Research, we developed an interactive platform for cognitive training, for both mobile devices and web apps, using MMIR [5]. The platform is specifically aimed at older users with

and without mild cognitive impairment (MCI). The Geriatrics Center of Berlin's Charité hospital lead the development of cognitive exercises for the areas memory, attention, executive functions, language ability and information processing speed. The platform could be used individually, but provides also group training via AV communication. Additionally, target group-oriented information and communication facilities were developed to promote social interaction. A target group-specific interaction concept and interface design for such a complex system was essential. The interaction with the controls of the platform and the exercises should release more cognitive resources than they consume. Taking into account possible sensory, motor and cognitive limitations, we realized a user-centred design that allows for speech and standard tablet/computer interaction methods.

In the follow-up project **Wir im Kiez,** where a mobile interaction platform for seniors with "normal" cognitive capabilities was developed, we brought in our design experience, methodology, and tested style guide [12]. However, these special requirements were initially treated with low priority, and the interface was designed according to current web paradigms and favoured a pleasing-looking, standard web design over the special requirements of the target group. After iterative user feedback, the UI now uses speech and gesture modalities, which is still attractive but acknowledges the requirements of the users.

Design of Multimodal Interaction

What can the language technology community learn from CHI research?

We developed a user guided approach for gathering feedback about the suitability of speech input and expected system feedback, including speech synthesis [4]. In order to gather feedback about the suitability of speech input and proper system feedback, we apply an adapted walkthrough method. The methodology is inspired by the cognitive walkthrough method, which is used for identifying usability issues in interactive systems [13]. The aim of the adapted walkthrough method is to find out with respect to voice commands and system feedback, whether, where, and what specific extensions can improve the interaction between user and system [10]. For each interaction step of a specific task test users have to answer the following questions:

- Is speech input suitable for the task or single task steps?
- Would you expect speech or other auditory or tactile feedback?

In addition, the test users should specify which speech input they would prefer. They where asked to explain their answers and to state any kinds of expected system feedback.

How can the user-acceptance of language technologies be improved?

We believe that language technology will be better accepted by the user only if it is implemented in an easy to use and as intuitive as possible way. The evolution of ASR herewith has a positive effect. However, NLP still turns out to be a greater technological challenge. So far there are less guidelines for interaction designers of speech interfaces then for GUI designers. In order to provide necessary technological concepts addressing these issues we

argue for a framework based implementation of user interfaces. Many of our HCI systems are implemented using MMIR, a framework providing a lightweight multimodal dialog manager [5]. Using HTML5 as base technology MMIR provides an easy and common way to integrate a graphical user interface. The framework further offers capabilities for recognizing speech input and producing for speech output. The dialog manager combines input from the different modalities and generates the appropriate system output. Applications created with MMIR run as mobile apps as well as browser based desktop applications. As part of our recent research we also use MMIR in combination with wearable devices. In concordance with this technological evolution, we also worked on user-centred design and test methods for unimodal and multimodal systems, up to semi-automated usability tests [1] for speech dialogue systems and test systems for multichannel systems [2].

Evaluation of Multimodal Interaction

Can we bridge the divide between the evaluation methods used in HCI and the AI-like batch evaluations used in speech processing?

A possible approach for bridging evaluation methods in HCI and speech processing could be found in the field of automated usability evaluation (AUE). In order to identify and eliminate usability errors AUE simulations are performed with first prototypes before the implementation of the real system begins. The simulations are typically performed as batch evaluations, as compared to real users studies many runs of simulated user interactions can be performed. First approaches of such AUE simulations already exist. The MeMo Workbench e.g. enables the simulation of speech based and multimodal interaction considering typical ASR errors (substitutions, insertions, deletions, no-match) [3,6]. In SpeechEval we developed such a system for the automated test of unimodal speech Systems [1].

Can speech and multimodal increase usability and robustness of interfaces and improve user experience beyond input/output?

Results of our own studies revealed that the perceived usability is not always in line with the robustness of a multimodal interface: in [8] we compared versions of a system differing in the robustness of touch screen and speech input, and discovered that a system with perfectly working speech input and impaired touch input was rated better then a system with both modalities working perfectly. A reason for this outcome could be that participants highly valued well-performing speech input in the presence of touch screen errors because of speech shortcuts and accuracy.

Conclusion

Overall our work exemplifies that the benefits emerging from multimodality, the design of multimodal systems, and the evaluation of multimodal HCI are highly interrelated. We advocate that multimodality should be considered where the user has a benefit of it, that such benefits should be specifically supported by the interface design process, and that targeted use of multimodality should have positive impact on the evaluation results of HCI. Our present experiences with wearable applications imply that at least some design principles evolving from the understanding of the benefits of multimodal HCI may also be applicable for multi-device interaction.

Acknowledgements

This work is part of the STREETLIFE (Steering towards Green and Perceptive Mobility of the Future) project co-funded under the 7th RTD Framework Programme, FP7-SMARTCITIES-2013 – grant agreement 608991. For more information please go to http://www.streetlife-project.eu

References

1. Jana Götze, Tatjana Scheffler, Roland Roller, and Norbert Reithinger. User Simulation for the Evaluation of Bus Information Systems. 2010. In *Proceedings of 2010 IEEE Workshop on Spoken Language Technology. IEEE Workshop on Spoken Language Technology (SLT-2010)*, 454-459.
2. Christian Husodo Schulz. 2015. MultiRep: A Platform Enabling Seamless Mobile Interaction. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct (MobileHCI '15)*. 622-627.
3. Sebastian Möller, Roman Englert, Klaus-Peter Engelbrecht, Verena Vanessa Hafner, Anthony Jameson, Antti Oulasvirta, Alexander Raake, and Norbert Reithinger. 2006. Memo: towards automatic usability evaluation of spoken dialogue services by user error simulations. In *Proceedings of INTERSPEECH 2006*.
4. Norbert Reithinger, Aaron Russ, and Kinga Schumacher. 2015. User-centered Interaction Design of a Mobile Learning Platform for the Generation 60 +. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct (MobileHCI '15)*. 924-927.
5. Aaron Ruß. 2013. MMIR Framework: Multimodal Mobile Interaction and Rendering. In *Proceedings of the GI-Jahrestagung 2013*.2702–2713.
6. Stefan Schaffer. 2015. Modeling Modality Selection in Multimodal Human-Computer Interaction. Ph.D. Dissertation. TU Berlin, Berlin, Germany.
7. Stefan Schaffer, Benjamin Jöckel, Ina Wechsung, Robert Schleicher, and Sebastian Möller. 2011. Modality Selection and Perceived Mental Effort in a Mobile Application. In *Proceedings of INTERSPEECH 2011*. 2253-2256.
8. Stefan Schaffer, Michael Minge. 2012. Error-prone voice and graphical user interfaces in a mobile application. In *Proceedings of 10th ITG Conference on Speech Communication (ITG Speech '12)*.
9. Stefan Schaffer and Norbert Reithinger. 2014. Intermodal personalized Travel Assistance and Routing Interface. In *Proceedings of Mensch & Computer 2014*. 343-346.
10. Stefan Schaffer, Aaron Russ, and Norbert Reithinger. 2016. User Guided Speech Technology Integration for a Mobility Application. Mensch und Computer 2016. *Submitted*.
11. Stefan Schaffer, Robert Schleicher, and Sebastian Möller. 2011. Measuring cognitive load for different input modalities. In *Proceedings of the 9. Berliner Werkstatt Mensch-Maschine-Systeme*, 287-292.
12. Sven Schmeier, Aaron Ruß, and Norbert Reithinger. 2015. Wir im Kiez: Multimodal App for Mutual Help Among Elderly Neighbours. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)*. 379-380.
13. C. Wharton, J. Rieman, C. Lewis, and P. Polson. 1994. Usability inspection methods, J. Nielsen and R. L. Mack, Eds. New York, NY, USA: John Wiley & Sons, Inc., 1994, ch. The Cognitive Walkthrough Method: A Practitioner's Guide, pp. 105–140.