User perception of multi-modal interfaces for mobile applications

Florian Metze[1], Roman Englert[2], Udo Bub[3], Ingmar Kliche[4], and Thomas Scheerbarth[4]

[1] Deutsche Telekom Laboratories, Technische Universität Berlin; Germany
[2] Deutsche Telekom Laboratories, Ben Gurion University; Beer Sheva, Israel
[3] Deutsche Telekom Laboratories; Berlin, Germany
[4] T-Systems Enterprise Services GmbH; Berlin, Germany
{florian.metze|roman.englert|udo.bub|ingmar.kliche|thomas.scheerbarth}@telekom.de

Abstract

This paper presents a comparative study on the usability of a service presented in telephone, PC-based web interface, and mobile/ multi-modal variants. The goal is not to analyze individual strengths and weaknesses of the different modalities, but to understand the user's perception of the SUMI criteria (efficiency, affect/ likability, helpfulness, control, learnability), and the overall impression of a service with respect to the access variant tested. As multi-modality is often framed as a technology to make usage more "intuitive", we were particularly interested in the differences between experienced and novice users. To this end, we conducted a study with 80 participants and conclude that, while multi-modality is accepted by experienced users, it seems to be asking too much from novice users, particularly with respect to learnability and efficiency.
Index Terms: multi-modality, perception, usability, evaluation, mobile applications

1. Introduction

Given the continuing trend towards miniaturization of electronic components, ever more processing power will be available on small, portable, durable, and "intelligent" devices. These devices can support increasingly complex applications and increasingly lend themselves to offer convenient mobile access to applications previously available through voice- or web-based services.

This situation is usually being dealt with in the "multi-modal" paradigm, arguing that interaction can be optimized by relying on devices offering several modalities, which the user can choose from according to his preferences or the environmental context. For a small device, the prevalent assumption is that (multi-)touch, stylus, gesture, or speech input is the most suitable modality for input, while small, but increasingly high-resolution screens are available for output, along with audio and, for some events, tactile (vibration) feedback.

Reality, however, has not seen many applications making full use of multi-modal hardware, in the sense that the application supports all modalities supported by the device, a large fraction of the user base actually uses them, and they experience a measurable (or perceived) benefit from doing so.

Multi-modality is usually presented as a technology to make life easier, and to make technology usable by a broader spectrum of the average population [1]. Still, users often do not use multi-modal features offered by a device or service, particularly once uni-modal dialog patterns have already been established [2]. Presumably, this is due to our still incomplete knowledge about the natural integration patterns that typify people's combined use of different input modes [3]. As a consequence, these patterns strongly depend on a user's previous experience. If companies are to exploit multi-modality commercially, their customers will usually already be using certain products and services, and their perception of a new multi-modal service will depend upon their previous knowledge. As companies are introducing more stringent look'n'feel rules for their products and services, the problem aggravates. This paper therefore looks at users' perception of the same service, if it is presented in two uni- and one multi-modal setting, and assesses the added value of multi-modality by comparing the SUMI scores [4, 5] of the different interfaces.

2. Multi-modality for Mobile Applications

Over the past two years, we developed various multi-modal applications, mainly for mobile devices [6, 7]. Applications include customer care automation, gesture-driven avatars, and pedestrian navigation systems. They were in most cases derived from existing services. The primary goal was to increase service usage by enabling more convenient and intuitive usage, for example by allowing one-hand operation of services in mobile settings.

2.1. Standards and Semantic Interaction Management

Having multiple channels available in parallel enables a user to provide input via, e.g., stylus and speech. The system has to combine potentially conflicting inputs and to generate consistent tasks for the interaction manager, which in turn generates appropriate system responses. In our systems, we use the Extensible MultiModal Annotation markup language (EMMA) [8] to annotate multi-modal features usign XML code.

Various standard bodies are working on multi-modal architectures and frameworks [9]. Our implementation aligns with the W3C's multi-modal architecture [10]. The GUI component was implemented using HTML, while the voice modality component uses browser plug-ins for both embedded and server based automatic speech recognition.

2.2. Related Work

Multi-modal interfaces support users interacting with applications by offering several modes like speech, keyboard, stylus, gestures, and/ or mimic. The question of how to fuse several input channels according to user preferences and context has been studied extensively in theory and practice [3], and the problem of generation of appropriate, symmetric output is also under intensive investigation [1, 11]. For interpretation, systems have to

Figure 1: The multi-modal interface to the "music greeting card". The user just entered a telephone number (step 6 of the application). He would have done so by pressing a key on the side of the device or by tapping the microphone button, and then speaking, or by using a soft keyboard.

generate hypotheses of the user intention and to select the most promising one, for example by skipping hypotheses which violate pre-defined integrity constraints. For generation, evaluation of appropriateness is a major concern.

Oviatt et al [2] conclude that users will interact multi-modally in order to manage their cognitive load efficiently. Users will generally interact multi-modally, when given the opportunity to do so. While the ratio of users' multi-modal interactions was 59.2% on a low complexity task, this number increased to 68.2% for a high complexity task. There is a dramatic increase in ratio (from 18.6% to 77.1%), when users had to establish a new dialog context.

While this describes conditions under which users will use multi-modality, it does not measure or describe user perception, i.e. the perceived quality of use as described e.g. in [12]. When working to boost usage or commercial success of multi-modal applications, both factors have to be taken into account.

When evaluating multi-modal interfaces, generally accepted criteria like effectiveness, efficiency, or satisfaction depend on individual usage preferences for modalities. A qualitative approach is represented by design-based usability engineering [13], where experts present design sketches for different modes to the user and ask for a judgement. The feedback is used to improve the interface design and functionality. To refine further, a quantitative approach can be applied to compare results over iterations, with other interfaces, and across modalities, irrespective of individual usage patterns. This approach is frequently applied and data for more than 2000 evaluations are provided by the SUMI test [4, 5].

3. Application and Prototype

To investigate how different user groups perceive multi-modal applications on current mobile devices when compared to uni-modal applications on "conventional" devices, we compared several interface variants for the same service, for which we expect a high added value of a multi-modal user interface, because of easier operation, for example while on the road. We developed a multi-modal user interface for the "electronic music greeting card", a commercial service already accessible through traditional voice-only ("telephony") and keyboard ("PC-based Internet") variants. This service allows to send a voice-mail message accompanied by a music song (chosen by the sender) to another telephone subscriber. The sender provides a telephone number, a time and date, leaves his own (voice) message, and chooses a song, which will then be combined into a "music greeting". In the deployed variants, the greeting can be composed by using an Interactive Voice Response (IVR) system from any standard phone, or the greeting can be composed inside a Web-browser on any personal computer (PC).

The multi-modal prototype converged the two variants of user interfaces into a single application which allowed to use voice input as well as tactile (stylus) input and audio-visual output, and runs on a personal digital assistant (PDA). For input, voice and tactile (stylus) modalities were available in sequential mode. Thus the user was able to navigate through the application by voice or stylus, to use voice shortcuts or to use voice or stylus to fill in information (e.g. date and time). The availability of voice input was signaled by a microphone button, and recordings could be activated using a push-to-talk button on the side of the device, which was also used to record voice messages.

The multi-modal application for mobile devices was implemented on a PDA (T-Mobile MDA Vario II, a Microsoft Windows Mobile 5 device with touch screen) and was intended to be used (mainly) when traveling. The goal during development of the multi-modal variant of the user interface was to increase accessibility of the service for mobile users and therefore to increase service usage. Obviously, a mobile browser is not the best platform for a multi-modal interface, but it is supported by a large fraction of deployed devices and we therefore opted for this approach instead of a "green-field" specific solution. We chose to use Windows Mobile, because we expect a certain consistency in look-and-feel between the PC- and PDA-based solutions to be an advantage for first-time users of PDA-based solutions.

The application has been designed using a wizard-like concept containing the following steps to be completed: (1) login, (2) select song category (e.g. fun, charts, top seller), (3) select song, (4) (optionally) get detail information per song (like artist, album or price), (5) record personal message (and listen to the recorded message), (6) enter target phone number, (7) (optionally) enter delivery date and time (otherwise immediate delivery), (8) listen to complete music greeting card (incl. accompanying music), (9) send message, and (10) (optionally) re-send message to another receiver.

The prototype application is fully functional and integrated with production platforms (e.g. for message delivery) using the technologies described in Section 2.

The user interface was optimized prior to the experiments described in this paper using expert evaluation, a first usability test with 10 subjects per target group, during which experts observed users, collected qualitative and quantitative feedback (including, but not limited to, the SUMI questionnaire), and noted usability deficiencies, followed by a second expert evaluation

of the re-engineered interface. The most frequent remaining usability problems were related to the activation of speech input (using either a soft button on the screen or a hard button on the side of the device) and the speed of system responses, particularly as no immediate visual feedback could be provided in our architecture when dictating telephone numbers.

4. Evaluation

In order to be able to test the users' acceptance of our application on different access devices, we evaluated the user perception using the general and standardized SUMI questionnaire.

4.1. SUMI Questionnaire

The SUMI (Software Usability Measurement Inventory) questionnaire [4, 5] is referenced in the ISO 9241 standard as a recognized tool for testing user satisfaction [12]. The SUMI questionnaire consists of 50 statements, with which users "agree", "cannot decide" or "disagree", after having tested the service under evaluation, and with reference to the overall interaction, not individual steps. Targeted at software products in general, the test is widely used for the usability evaluation of the client side of client-server applications or any other system requiring interaction with a user.

SUMI results are computed by SUMISCO (SUMI Scoring Package), a report generation tool which compares the system under evaluation to user ratings of about 2000 commercially available software packages. SUMI reports describe usability in several dimensions:

Efficiency: the degree to which users feel that the software assists them in their work and is related to the concept of transparency;

Affect (likability): the user's general emotional reaction to the software;

Helpfulness: the degree to which the software is self-explanatory, including specifics like the adequacy of help facilities and documentation;

Control: the extent to which the user feels in control of the software, as opposed to being controlled by the software, when carrying out the task; and

Learnability: the speed and facility with which the user feels that they have been able to master the system, or to learn how to use new features when necessary.

The SUMI scale is normalized so that the scores have a mean of 50 and a standard deviation of 10. 68% of software will therefore have a score between 40 and 60.

4.2. Test Preparation

Paid test subjects were recruited from two different ICT (information and communication technology) user groups using a screener. Both user groups were expected to be familiar with computers, the Internet, cell phones, and IVR systems, but "trust guided ICT users" (TG) had no experience with Personal Digital Assistants (PDAs), while "experienced ICT users" (EX) had used PDAs before. Overall, 41 users were female, and 39 were male. However, only 7 of the 40 EX users were female, and 6 of 40 TG users were male. All participants were between 18 and 75 years of age. The average age for the EX user group was 29, while for TG it was 52. These user groups were chosen to represent diverse attitudes towards and experience with technology in general, and mobile devices in particular. The test language for documents and interfaces was German, and tests were conducted in quiet office rooms.

The telephone and PC-based variants as described in Section 3 were accessed using a live service, while the multi-modal interface was accessed using the prototypes developed. Users received written instructions before the tests and those working with the multi-modal interface were shown how to operate the touch screen using the stylus and the speech input using an activation ("push-to-talk") button on the side of the device or a soft button on the screen. They did however not work with the interface themselves before the actual test. Still, all users successfully completed the tasks given to them. Two series of experiments were conducted: during the first series, every user tested one interface only, while in the second series users first completed the test with one interface before starting the test with another interface. A test consisted of the user completing a screener[1], reading the task description, then executing the task, and finally filling out the questionnaire. A supervisor was available to answer participants' questions and note usability deficiencies, but did not otherwise interfere with the experiment.

4.3. SUMI Test Results

In our first experiment, the three different interfaces were tested by 20 users each. Of these, 10 were "TG" users, and 10 were "EX" users. Nearly all individual SUMI scores for the "telephony" and "PC-based Internet" variants of the interface were in the 50 to 60 range, indicating good general usability. The only statistically significant observation[2] is that "TG" users find the telephone more helpful (score of 63 vs. 53), while "EX" users find the PC-based interface more efficient (58 vs. 50).

The multi-modal interface is consistently rated worse than the telephony- and PC-based solutions, although the differences are statistically insignificant. However, there is a significant difference in the perceived efficiency (50 vs. 31) and learnability (63 vs. 37) of the multi-modal interface for the "EX" and "TG" user groups. The trust-guided user group rates the multi-modal interface significantly worse than the experienced user group, which means the service is not intuitive to use for the trust-guided user group.

In order to be able to directly compare two interfaces, and to also investigate the influence of previous experience on the perception of the multi-modal interface, we performed a second experiment, in which trust-guided users tested the PC-based interface and the multi-modal interface in sequence. Here, we tested 10 users for each order of presentation. We decided to contrast the PC-based interface with the multi-modal interface, as these two appear more similar to each other, than to the telephony interface. Table 1 presents the result: for the PC-based interface, the scores do not depend on the order of presentation, while the multi-modal interface is rated significantly worse when it is being tested after the users had worked with the PC-based interface.

In detail, we find that the perceived efficiency is not affected by the order of use, as is learnability. Control is unaffected by the order for the PC interface, but users feel less confident with the multi-modal interface after having tested the PC-based interface first. However, Table 1 shows a significant difference in affect and helpfulness for the two services, depending on the order of presentation. Affect (likability) drops from 46 to 23

[1] This step was of course skipped for the test of the second interface in the second series.

[2] Here and in the following, "statistical significance" is reached at the 5% probability for error level.

for the multi-modal interface and increases from 51 to 58 for the PC interface, if it is the second interface to be tested, respectively. The same effect can be observed for helpfulness, i.e. even though the multi-modal interface is based on Windows Mobile and tries to emulate the PC-based experience as much as possible, including the possibility to act uni-modally, users learn to compare it to the PC-based interface and like the multi-modal interface significantly less. The perceived helpfulness of the PC-based interface increases, when contrasted with the multi-modal interface. "Control" is also reduced for the multi-modal interface, when tested after the PC-based interface.

Efficiency scores from Table 1 can also be compared to objective measurements of task completion time, which for our experiments confirm the subjective impressions given by our subjects (see Table 2): for experienced users, there is no significant difference between the times for the multi-modal interface and the telephony interface, while trust-guided users (and the combined user group) takes significantly longer to complete the task using the multi-modal interface, than with the two conventional interfaces.

Table 1: Median of the SUMI scores for the various interfaces and the trust-guided user group. TEL=telephony channel, WEB=PC-based Internet, MOD=multi-modal interface. The lower two blocks contain the results of the "order effect" experiment. *WEB-MOD* means the PC-based interface is tested before the multi-modal interface.

Interface *(Order)*	Efficiency	Affect	Helpfulness	Control	Learnability	Overall
TEL	59	57	63	61	52	62
WEB	50	59	57	58	56	59
MOD	31	46	34	46	37	37
WEB						
(WEB-MOD)	50	51	49	52	57	53
(MOD-WEB)	54	59	60	54	51	55
MOD						
(MOD-WEB)	31	46	34	46	37	37
(WEB-MOD)	29	23	22	29	39	27

Table 2: Median ranks for time needed to complete tasks in pairwise TEL vs. MOD and WEB vs. MOD comparison as measure of (objective) time needed to complete the task. p is computed using the Mann-Whitney-U test.

User group	TEL	MOD	p	WEB	MOD	p
ALL	14.95	26.05	0.002	13.63	27.38	0.000
TG	6.4	14.6	0.001	6.2	14.8	0.000
EX	7.9	13.10	0.052	6.5	14.5	0.002

5. Conclusion

Our main conclusion from these exeriments is that multi-modal interfaces can be more difficult to understand than interfaces that provide only one input or output channel, despite frequent assumptions of the contrary. Users can fail to perceive the advantages of multi-modality, particularly when uni-modal use patterns have already been established. These mainly influence their perception of likability, control, and helpfulness. The perception of efficiency and learnability seems unaffected. Familiarity with an interface therefore is an issue: once users know an interface, they find it difficult to be confronted with a different, multi-modal interface, which uses similar or the same metaphors. In our experience, this observation seriously obstructs efforts to enhance established, live services by adding multi-modality, as suitable multi-modal interaction patterns, which users would recognize across services, are currently being researched, but have not yet been deployed commercially.

Future work on the automatic, and consistent interpretation of multi-modal input will hopefully bring about multi-modal applications exhibiting a unique, and alluring look'n'feel, which also derives from existing use patterns. This development should allow users to operate multi-modal interfaces more intuitively and allow multi-modality to fare well in today's market research, which is generally based on conjoint analysis selection processes and frequently leads to new functionalities being preferred over improved access to existing functions, which results in today's feature-laden devices, with poor usability.

6. Acknowledgments

The authors would like to thank Dorothea Kugelmeier and Britta Hofmann from FhG FIT for conducting and evaluating the SUMI study.

7. References

[1] S. Oviatt, P. Cohen, L. Wu, L. Duncan, B. Suhm, J. Bers, T. Holzman, T. Winograd, J. Landay, J. Larson, and D. Ferro, "Designing the user interface for multimodal speech and pen-based gesture applications: State-of-the-art systems and future research directions," *Human-Computer Interaction*, vol. 15, no. 4, pp. 263–322, 2000.

[2] S. Oviatt, R. Coulston, and R. Lunsford, "When do we interact multimodally?: cognitive load and multimodal communication patterns," in *Proceedings of the 6th international conference on Multimodal interfaces*. College Park, PA; USA: ACM, 2004, pp. 129–136.

[3] L. Wu, S. Oviatt, and P. Cohen, "Multimodal integration - a statistical view," *IEEE Transactions on Multimedia*, vol. 1, no. 4, pp. 334–341, Dec. 1999.

[4] J. Kirakowski, *SUMI User Handbook*, 2nd ed. Human Factors Research Group, University College Cork, Ireland, 1998.

[5] H. Cavallin, M. Martin, and A. Heylighen, "How relative absolute can be: SUMI and the impact of the nature of the task in measuring perceived software usability," *Journal of AI & Society*, vol. 22, no. 2, pp. 227–235, 2007.

[6] R. Englert and G. Glass, "Architecture for multimodal mobile applications," in *20th International Symposium on Human Factors in Telecommunication (HFT 2006)*. Sophia Antipolis, France: ETSI, March 2006.

[7] O. Schreer, R. Englert, P. Eisert, and R. Tanger, "Real-time vision and speech driven avatars for multimedia applications," *International Journal of IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 352–361, 2008.

[8] "EMMA: Extensible MultiModal Annotation markup language," http://www.w3c.org/TR/emma, 2007.

[9] "IETF DMSP: Distributed Multimodal Synchronization Protocol," http://www.ietf.org/internet-drafts/draft-engelsma-dmsp-04.txt, 2007.

[10] "Multi-modal Architecture and Interfaces," http://www.w3c.org/TR/mmi-arch, 2006.

[11] W. Wahlster, "Smartkom: Symmetric multimodality in an adaptive and reusable dialogue shell," in *Proceedings of the Human Computer Interaction Status Conference*, R. Krahl and D. Günther, Eds. Berlin, Germany: DLR, 2003, pp. 47–62.

[12] International Organization for Standardization, "ISO 9241: Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11: Guidance on usability," 1998.

[13] R. Englert and G. Joost, "Design and usability for personalized user interfaces of telecommunication services," in *16th International Conference on Engineering Design (ICED07)*, Paris, France, 2007.