Reliable Evaluation of Multimodal Dialogue Systems

Florian Metze[1], Ina Wechsung[2], Stefan Schaffer[2], Julia Seebode[2], and Sebastian Möller[2]

[1] School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213
[2] Deutsche Telekom Laboratories, TU Berlin, Ernst-Reuter-Platz 7, Berlin, Germany
fmetze@cs.cmu.edu

Abstract. Usability evaluation is an indispensable issue during the development of new interfaces and interaction paradigms [1]. Although a wide range of reliable usability evaluation methods exists for graphical user interfaces, mature methods are rarely available for speech-based interfaces [2]. When it comes to multimodal interfaces, no standardized approach has so far been established. In previous studies [3], it was shown that usability questionnaires initially developed for unimodal systems may lead to unreliable results when applied to multimodal systems. In the current study, we therefore used several data sources (direct and indirect measurements) to evaluate two unimodal versions and one multimodal version of an information system. We investigated, to which extent the different data showed concordance for the three system versions. The aim was to examine, if, and under which conditions, common and widely used methods originally developed for graphical user interfaces are also appropriate for speech-based and multimodal intelligent interfaces.

Keywords: usability evaluation methods, multimodal interfaces.

1 Introduction

Ever since the appearance of the "put-that-there" paradigm [4], multimodal user interfaces have been a subject of intensive scientific study in the human-computer interaction (HCI) community, and systems have been developed with a wide range of research foci [e.g. 5,6] and for a variety of applications [7,8]. Successful development and deployment of a multimodal interface requires proven performance of the underlying component modalities. Evaluation of a multimodal user interface therefore generally includes technical evaluation of individual components, to gather so-called *indirect* data [9]. This is usually relatively easy to do and reliable, in the sense that results can be reproduced, and, to some extent, predicted.

On a more abstract level, systems are usually evaluated with respect to their usability. HCI literature provides a wide range of methods to measure usability. Most of them were developed to evaluate unimodal graphical user interfaces (GUIs). Parameters used for usability measures include *direct* data, collected directly from users, often through questionnaires, and indirect data like for example log-files containing task duration or performance data. Since all these parameters are measuring at least roughly the same concept, namely usability, they would be expected to show high correlations between them.

J.A. Jacko (Ed.): Human-Computer Interaction, Part II, HCII 2009, LNCS 5611, pp. 75–83, 2009.

Multimodal interfaces between man and machines however cannot adequately be evaluated by using performance metrics originally developed for individual modalities only (e.g. word error rate or bandwidth), as humans cannot directly perceive these system properties, but unfortunately usability evaluation of multimodal interfaces has shown deficiencies in the past. In particular, it is unclear if methods well established for unimodal systems can provide valid and reliable results for multimodal systems and if preference for a certain modality can be correlated with measurable variables.

In earlier work, we observed that standardized questionnaires are not suitable for usability evaluation of multimodal systems, as there was only little agreement between the results measured with different questionnaires [3,10]. Concerning unimodal systems, a meta-analysis conducted by Nielsen and Levy [11] showed that performance and predicted preference are indeed correlated. Similar results were reported by Sauro and Kindlund [12], who found positive correlation between satisfaction (direct data) and time, errors and task completion (indirect data).

However, several studies reported opposing findings: Krämer and Nitschke [13] showed that user ratings of multimodal interfaces are not affected by increased intuitivity and efficiency. Möller [14] could not find correlation between task duration and user judgements when evaluating speech dialogue systems. Also, Frøkjær and colleagues [15] could not find correlation between user ratings and efficiency. Results from a meta-analysis by Hornbæk and Lai Chong-Law [16] showed that the user's experience of the interaction (direct data) and indirect data differ considerably from each other or show even negative correlations.

In view of the studies mentioned above, it seems necessary to use both kinds of data in usability evaluation, in order to obtain reliable results. Developing methods for usability evaluation of multimodal systems can only be done by validating within data types (e.g. validation across questionnaires), but also between data types (e.g. comparing indirect and direct results). The purpose of this paper is therefore to analyze, which questionnaire relates most to objective data and to investigate if the results are consistent with our earlier findings.

2 Method

2.1 Participants and Material

Thirty-six German-speaking individuals (17 male, 19 female) between the age of 21 and 39 (M=31.24) took part in this study of a multimodal system, which was originally available as a touch-only system, and then extended with speech input. Half of the participants were familiar with the touch-based system and thus considered as experts. The other eighteen participants were novices. They had no prior experience with the system.

The system tested is a wall mounted information and room management system operable via a graphical user interface with touch screen, speech control and a combination of both. The output is always given via GUI. The users performed six different tasks with the system (cf. Table 1).

Table 1. Description of tasks and required interaction steps

Task	Description	Minimum # of interaction steps required	
		Speech	Touch
T1	Show main screen	1	1
T2	Show 18. floor	1	1
T3	Search for employee	3	6
T4	Search for room	2	1 to ∞*
T5	Show event screen	1	1
T6	Show room for event	1	1

* If the room was accidentally booked at the time of the test, the task was solvable with one click. Otherwise a systematic search overall rooms was necessary.

To collect user ratings, the AttrakDiff questionnaire [17] was used in its original form. In addition the **S**ubjective **A**ssessment of **S**peech **S**ystems **I**nventory (SASSI) questionnaire [18] was used in a modified form, as shown in Table 2.

2.2 Procedure

Each individual test session took approximately one hour. Each participant performed the tasks in three blocks, one for every system version. At the beginning of each block, participants were instructed to perform the tasks with a given modality.

After each block, they were asked to fill out the AttrakDiff questionnaire [17] and a modified SASSI [18] questionnaire (cf. Table 2), in order to rate this version of the system. In order to balance fatigue and learning effects, the sequence of the unimodal systems was randomized in the first two blocks. In the third block,

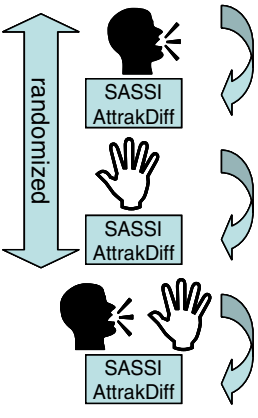

Fig. 1. Test sequence of touch, speech, and multimodal tests

Table 2. Modified and excluded SASSI items

Speech (original)	I sometimes wondered if I was using the right word
Touch (modified)	I sometimes wondered if I was using the right button
Multimodal (modified)	I sometimes wondered if I was carrying out the right action
Speech (original)	I always knew what to say to the system
Touch (modified)	I always knew which button to use
Multimodal (modified)	I always knew which action to carry out
Speech (original)	It is clear how to speak with the system
Touch (modified)	It is clear how to interact with the system
Multimodal (modified)	It is clear how to interact with the system
Excluded	The system is pleasant
	The system is friendly

participants could freely choose the interaction modality, and test the multi-modal system, before again filling out the SASSI and the AttrakDiff to rate the multimodal system (Figure 1).

The subscales for each questionnaire were calculated according to the instructions in the relevant handbook [17,18]. Furthermore for the SASSI an overall scale was calculated based on the mean of all items.

All negatively-poled items were re-coded, to ensure that higher values always indicate better ratings.

Furthermore, video, audio and log data were recorded during the sessions. As a measure of efficiency task duration was assessed via log-files and was, for each system version, summed over all tasks.

To analyze which modality the participants preferred, log-data of the multimodal test block was annotated. For every task, the modality used first to perform the task was selected for further analysis. This way, the percentages of modality preference per task have been computed.

3 Results

3.1 Direct Data - Questionnaires

Ratings on SASSI: The SASSI showed on all scales differences between all three system versions (cf. Table 3). The speech-based system was rated worst on all scales. The touch-based system was rated best on all scales except the speed scale. On the speed scale the highest ratings were observed for the multimodal system. No differences were found between expert and novice users.

Table 3. Ratings on SASSI subscales

Scale	System	Mean	SD	F (df)	p (part. eta²)
Accuracy	Touch	2.99	0.46	128.86 (2,68)	.000 (.791)
	Speech	1.18	0.53		
	Multimodal	2.20	0.79		
Likeability	Touch	3.16	0.40	139.22 (2,68)	.000 (.804)
	Speech	1.78	0.58		
	Multimodal	2.95	0.65		
Cognitive Demand	Touch	2.87	0.60	116.68 (2,70)	.000 (.769)
	Speech	1.25	0.57		
	Multimodal	2.52	0.78		
Annoyance	Touch	2.72	0.46	79.16 (2,70)	.000 (.693)
	Speech	1.73	0.50		
	Multimodal	2.58	0.68		
Habitability	Touch	2.64	0.85	62.70 (2,68)	.000 (.648)
	Speech	1.22	0.57		
	Multimodal	2.44	0.78		
Speed	Touch	1.90	0.33	5.34 (2,70)	.007 (.132)
	Speech	1.81	0.47		
	Multimodal	2.10	0.35		
Global Scale	Touch	2.70	0.36	180.25 (2,70)	.000 (.837)
	Speech	1.49	0.37		
	Multimodal	2.45	0.55		

Ratings on AttrakDiff: The AttrakDiff questionnaire revealed differences between all three systems versions: On the scale measuring pragmatic qualities the touch-based system got the highest ratings, the speech-based system got the lowest. On the scale hedonic quality-stimulation the multimodal and the speech-based system received higher ratings than the touch-based system. Regarding hedonic qualities-identity the touch-based and the multimodal system were rated better than the speech-based. On the attractiveness scale the touch-based was rated better than the multimodal and the speech-based system. The detailed results are given in Table 4. Again, novices and experts showed no differences.

3.2 Indirect Data – Performance Data

Task Duration: Participants needed least time to solve the tasks when using the multimodal system *($F(2,68)=185.02$, $p=.000$, part. $eta^2=.845$).* Means and standard deviation for each system version are given in Fig. 2. No differences were found between experienced and inexperienced users.

Table 4. Ratings on AttrakDiff subscales

Scale	System	Mean	SD	F (df)	p (part. eta²)
Pragmatic Quality	Touch	1.19	0.75	93.79 (2,66)	.000 (.740)
	Speech	-0.87	1.01		
	Multimodal	0.76	1.12		
Hedonic Quality - Stimulation	Touch	0.54	0.93	15.38 (2,68)	.000 (.311)
	Speech	1.03	0.68		
	Multimodal	1.21	0.79		
Hedonic Quality - Identity	Touch	0.82	0.78	12.84 (2,68)	.000 (.274)
	Speech	0.24	0.70		
	Multimodal	0.83	0.92		
Attractiveness	Touch	1.15	0.80	47.53 (2,68)	.000 (.583)
	Speech	- 0.06	0.87		
	Multimodal	0.92	0.99		

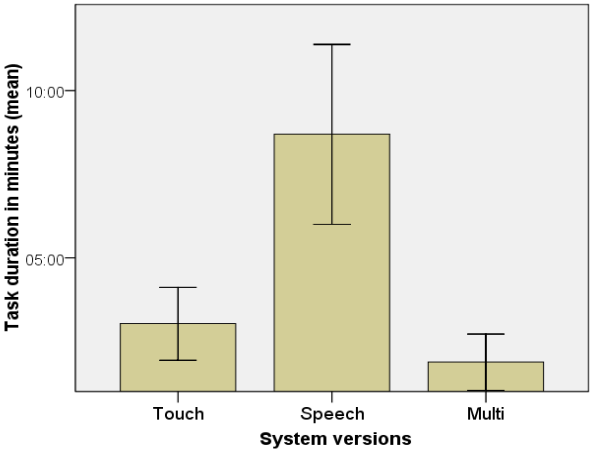

Fig. 2. Overall task duration for all three modalities

Modality Preference: Over all tasks, speech as input modality (52.3%) was slightly more preferred than touch (47.7%). A detailed analysis showed that modality preference was strongly determined by task characteristics: Users tended to prefer the modality most efficient (in terms of required interaction steps) for the specific task. The tasks T3 and T4 could be solved with less interaction steps when using speech. For both tasks speech was preferred more frequently than touch. For all other tasks either no differences were shown, or touch was preferred (cf. Table 5).

Table 5. Modality usage (percentages) by tasks

Task	Speech	Touch	χ^2	p	N
T1	50.0	50.0	.000	1.000	36
T2	22.2	77.8	11.110	.001	36
T3	75.0	25.0	9.000	.004	36
T4	83.3	16.7	16.000	.000	36
T5	57.1	42.9	.714	.500	35
T6	26.5	73.5	5.770	.024	34

3.3 Comparison of Direct and Indirect Data

Correlations between Scales Measuring Efficiency and Task Duration: Over all systems the scales measuring constructs related to efficiency were correlated with task duration. Significant negative correlations were observed between efficiency related scales and task duration. The highest correlation was shown for the AttrakDiff scale "pragmatic qualities". However significant correlations between the questionnaire ratings and task duration could only be found over all systems but not for the single system versions (cf. Table 6).

Table 6. Correlation (Pearson's r) between scales measuring efficiency and task duration (*p<.01; **p<.05)

Scale	Task Duration
SASSI Speed (N=107)	-.217*
AttrakDiff Pragmatic Qualities (N=105)	-.520**
AttrakDiff Attractiveness (N=104)	-.344**

Comparison between Modality Usage and Questionnaire Rating: The speech-based system was rated worst on all questionnaire scales except for one of the AttrakDiff scales ("hedonic quality – stimulation"). Furthermore task duration was longest with the speech-based system. Nevertheless, when interacting with the multimodal version speech was used more than touch or equally frequently used as touch to solve four of the six tasks.

4 Discussion

In the present study questionnaire ratings (direct data) partly matched the task duration (indirect data). As in a previous study [3], the AttrakDiff scale pragmatic qualities showed the highest agreement with task duration data. Thus this scale measures

the construct it was developed for. A similar conclusion can be drawn from the SASSI results: Ratings on the speed scale matched the task duration data. However, correlations were only observable over all system versions but not for every single version.

Furthermore the actual usage behaviour was hardly affected by the perceived quality (questionnaire ratings). Modality preferences and usage behaviour were strongly influenced by task characteristics. The most efficient modality (in terms of fewest necessary interaction steps) was chosen first. If the number of necessary interaction steps was the same for both modalities, either the touch system was preferred or no differences were observable. Since touch is more common it can be assumed that participants were more familiar with this modality. So possibly the usage of speech as input modality would increase as a function of practice.

In summary, the questionnaires showed correlations with the task duration measures but are not useful as a predictor of actual modality usage. Thus the current results are consistent with our earlier findings: Questionnaire initially developed for unimodal systems are not necessarily the best choice for the evaluation of multimodal systems. As in our previous studies [3,10] the current results point to the AttrakDiff as the most suitable questionnaire and thus as a proper basis for the development of new methods .

References

1. Sturm, J.: On the usability of multimodal interaction for mobile access to information services. PhD thesis, Radboud University Nijmegen, Nijmegen, The Netherlands (2005)
2. Larsen, L.B.: Assessment of spoken dialogue system usability - what are we really measuring? In: Eurospeech 2003, pp. 1945–1948 (2003)
3. Naumann, A., Wechsung, I.: Developing Usability Methods for Multimodal Systems: The Use of Subjective and Objective Measures. In: Proceedings of the International Workshop on Meaningful Measures: Valid Useful User Experience Measurement (VUUM), pp. 8–12 (2008)
4. Bolt, R.A.: "Put-that-there": Voice and gesture at the graphics interface. In: Proceedings of the 7th Annual Conference on Computer Graphics and interactive Techniques (1980)
5. Cohen, P.R., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L., Clow, J.: QuickSet: multimodal interaction for distributed applications. In: Proceedings of the Fifth ACM international Conference on Multimedia (1997)
6. Martin, J., Buisine, S., Pitel, G., Bernsen, N.O.: Fusion of children's speech and 2D gestures when conversing with 3D characters. Signal Process 86, 12 (2006)
7. Perzanowski, D., Schultz, A.C., Adams, W., Marsh, E., Bugajska, M.: Building a Multimodal Human-Robot Interface. IEEE Intelligent Systems 16(1), 16–21 (2001)
8. Thalmann, D.: The virtual human as a multimodal interface. In: Proceedings of the Working Conference on Advanced Visual interfaces (2000)
9. Möller, S., Engelbrecht, K.-P., Kühnel, C., Wechsung, I., Weiss, B.: Evaluation of Multimodal Interfaces for Ambient Intelligence. In: Aghajan, H., López-Cózar Delgado, R., Augusto, J.C. (eds.) Human-Centric Interfaces for Ambient Intelligence. Elsevier, Amsterdam (2009)
10. Wechsung, I., Naumann, A.B.: Evaluation Methods for Multimodal Systems: A Comparison of Standardized Usability Questionnaires. In: André, E., Dybkjær, L., Minker, W., Neumann, H., Pieraccini, R., Weber, M. (eds.) PIT 2008. LNCS (LNAI), vol. 5078, pp. 276–284. Springer, Heidelberg (2008)

11. Nielsen, J., Levy, J.: Measuring usability: Preference vs. performance. Communications of the ACM 37, 4 (1994)
12. Sauro, J., Kindlund, E.: A method to standardize usability metrics into a single score. In: Proc. CHI 2005. ACM Press, New York (2005)
13. Krämer, N.C., Nitschke, J.: Ausgabemodalitäten im Vergleich: Verändern sie das Eingabeverhalten der Benutzer? [Output modalities in comparison: Do they change user's input behaviour?]. In: Marzi, R., Karavezyris, V., Erbe, H.-H., Timpe, K.-P. (Hrsg.) Bedienen und Verstehen. 4. Berliner Werkstatt Mensch-Maschine-Systeme, VDI-Verlag, Düsseldorf (2002)
14. Möller, S.: Messung und Vorhersage der Effizienz bei der Interaktion mit Sprachdialogdiensten [Measuring and predicting efficiency for the interaction with speech dialogue systems]. In: Langer, S., Scholl, W. (eds.) Fortschritte der Akustik - DAGA 2006. DEGA, Berlin (2006)
15. Frøkjær, E., Hertzum, M., Hornbæk, K.: Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? In: Proc. CHI 2000. ACM Press, New York (2000)
16. Hornbæk, K., Law, E.L.: Meta-analysis of correlations among usability measures. In: Proc. CHI 2007. ACM Press, New York (2007)
17. Hassenzahl, M., Burmester, M., Koller, F.: AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität [A questionnaire for measuring perceived hedonic and pragmatic quality]. In: Ziegler, J., Szwillus, G. (eds.) Mensch & Computer 2003. Interaktion in Bewegung. B.G. Teubner, Stuttgart (2003)
18. Hone, K., Graham, R.: Subjective assessment of speech-system interface usability. In: Proceedings of Eurospeech 2001, vol. 3 (2001)