Predicting the quality of multimodal systems based on judgments of single modalities

Ina Wechsung[1], Klaus-Peter Engelbrecht[1], Anja B. Naumann[1], Stefan Schaffer[2], Julia Seebode[2], Florian Metze[3] and Sebastian Möller[1]

[1]Deutsche Telekom Laboratories, TU Berlin, Germany
[2]Research training group prometei, TU Berlin, Germany
[3]interACT center, Carnegie Mellon University, Pittsburgh, PA, USA

{ina.wechsung, klaus-peter.engelbrecht, anja.naumann, sebastian.moeller}@telekom.de
{julia.seebode, stefan.schaffer}@zmms.tu-berlin.de, fmetze@cs.cmu.edu

Abstract

This paper investigates the relationship between user ratings of multimodal systems and user ratings of its single modalities. Based on previous research showing precise predictions of ratings of multimodal systems based on ratings of single modality, it was hypothesized that the accuracy might have been caused by the participants' efforts to rate consistently. We address this issue with two new studies. In the first study, the multimodal system was presented before the single modality versions were known by the users. In the second study, the type of system was changed, and age effects were investigated. We apply linear regression and show that models get worse when the order is changed. In addition, models for younger users perform better than those for older users. We conclude that ratings can be impacted by the effort of users to judge consistently, as well as their ability to do so.

1. Introduction

The research interest in multimodal interfaces has been rapidly growing in the last years. These developments resulted in two different groups of multimodal interfaces [1]: Interfaces combining multiple input modalities (e.g. speech input in cars) and interfaces with multimodal output (e.g. warnings systems). Additionally to the two groups mentioned in [1], also systems having both multimodal input and multimodal output, are currently evolving. Although multimodal systems have been around for more than 25 years now [2] and besides the rapidly increasing technical developments in this area, evaluation methods and design guidelines are still rare and evaluation of multimodal systems is considered as problematic [2].

In a previous study we found that standardized usability questionnaires designed for unimodal systems are not appropriate for post interaction overall evaluation of multimodal systems. The only questionnaire yielding valid and reliable results was the AttrakDiff [3] questionnaire.

The theoretical base of the AttrakDiff questionnaire is Hassenzahl's model of user experience [3]. This model postulates that a product's attributes can be divided in hedonic and pragmatic attributes. Hedonic refer to the products ability to evoke pleasure and emphasize the psychological well-being of the user. Hedonic attributes can be differentiated into attributes providing stimulation, attributes communicating identity, and attributes provoking memory.

The AttrakDiff aims to measure two of these hedonic attributes: stimulation and identity. A product can provide stimulation when it presents new interaction styles, like for example new modalities. Thus multimodality should be measureable with the scale *Hedonic Quality-Stimulation* (HQ-S). The scale *Hedonic Quality-Identification* (HQ-I) measures a products ability to express the owners self. The scale *Pragmatic Quality* (PQ) covers a product's functionality and the access to the functionality. Additionally the perceived global quality is measured with the scale *Attractiveness* (ATT) The entire questionnaire comprises 28 items on a 7-point semantic differential.

We used this questionnaire to examine how judgments of unimodal system versions relate to the judgments of the multimodal version [5]. It was shown that accurate predictions of overall and global ratings of multimodal systems are possible on the basis of the respective ratings of the unimodal systems. Ratings of the multimodal systems matched the weighted sum of the ratings of the unimodal systems with the modality used more often having the stronger influence. However regarding specific measures (two of the AttrakDiff subscales) prediction performance was lower. Only for one subscale (*Hedonic Quality-Stimulation*) equally good prediction performance as for the overall judgments and the global scale was shown.

But these findings may have been a result of the test design and the tested system. The multimodal version was always the system tested last. Therefore it is possible that the participants tried to rate consistently, adding up their single-modality judgments in their minds. Consequently, the judgments of the multimodal version would not represent the actual quality of that system.

Moreover, when interacting with the multimodal system used in this study interference between the modalities was possible (e.g.: the speech recognizer was unintentionally switched on by talk not directed to the system). It is therefore plausible that without this interference different result may have been obtained.

Facing these limitations we conducted two follow-up studies to investigate how subjective ratings of unimodal system versions relate to the subjective ratings of multimodal system version.

2. Experiment 1

With this experiment we aimed to examine if the results obtained in [5] were a consequence of the test design and the participants effort to rate consistently. Therefore we changed the order of the system versions with the multimodal system presented first. Thus mentally adding up the single-modality judgments to the multimodal judgments was not possible. The

participants can only pre-estimate the quality of unimodal system version and their impact on the multimodal systems quality.

Furthermore addition is a less effortful mental operation than subtraction [6],[7]. This means that single modalities are more difficult to derive from multimodal ratings. Consequently less accurate predictions are expected.

2.1. Method

2.1.1. Participants and material

Eighteen German-speaking individuals (9 male, 9 female) between the age of 22 and 30 (M = 26.7) took part in the study. The tested system is a wall-mounted room management and information system. The system is basically the same as used in [5]. However, it was extended with face recognition to activate the speech recognizer. It is controllable via a graphical user interface with touch screen, speech input and a combination of both. The output is always given via the graphical user interface.

The AttrakDiff questionnaire [3] was employed to collect user ratings. The subscales were calculated in accordance with [3]. Furthermore an overall scale was calculated based on the mean of all 28 items. Modality usage was recorded with log files.

2.1.2. Procedure

The experiment consisted of three blocks. In each test block the participants were asked to perform six different tasks with the system. In the first block the participants were free to choose an input modality for solving the task. It was possible to switch or combine modalities after each task and also within a task. In the following blocks they were instructed to use a given modality. The tasks were the same for all blocks and the same as in [5]. In order to rate the previously tested system version the AttrakDiff questionnaire had to be filled out after each test block.

To analyze which modality was used by the participants in the first block, log-data of the test was annotated. For every interaction step the modality used to perform the step was logged. This way, the percentages of modality usage per task and across all tasks have been computed.

2.2. Results

Stepwise multiple linear regression analysis was conducted for each subscale and the overall scale using all cases. The questionnaire ratings obtained after the multimodal test block were used as response variable, the ratings obtained after the unimodal test blocks were used as predictors.

Prediction was not possible for two sub-scales as no significant predictor could be found by the stepwise inclusion algorithm (*Hedonic Quality-Identity, Pragmatic Quality*). The highest accuracy was observed for the scale *Hedonic Quality-Stimulation.* (cf. Table 1)

For all of the scales where prediction was possible beta-coefficients were not differing much between the modalities. Both modalities were used equally frequently (Touch: 51.7%; Speech: 48.3%). Thus it is plausible both modalities have a similar impact on the judgments of the multimodal system.

Table 1. *Results of multiple linear regression analyses for Experiment 1 using all cases. Only significant parameters (p<.05) are included in prediction*

Scale	Touch				Speech				R^2	RMSE	F (2,15)
	B	SE B	β	t (15)	B	SE B	β	t (15)			
Overall	.679	.218	.527	3.11	.553	.188	.497	2.93	.573	.546	10.08
ATT	.653	.235	.462	2.78	.545	.164	.552	3.32	.596	.725	11.05
HQ-S	.664	.118	.725	5.64	.485	.145	.430	3.35	.754	.414	22.97

To test for over-fitting effects leave one out cross-validation was conducted. Only for *Hedonic-Quality Stimulation* the model was stable (R^2=.533, RMSE=.276). Very large over-fitting effects were observed for *Attractiveness* (R^2=.061, RMSE=.724) and the overall scale (R^2=.044, RMSE=.415).

2.3. Discussion

Although the results indicate that the order of test conditions influences the accuracy and stability of the models, the same questionnaire scales as in [5] were suitable for prediction. Thus predicting judgments for multimodal systems based on judgements of single modalities might primarily be possible for overall and global (*Attractiveness* scale) judgements and for the subscale measuring stimulation. Additionally only data of 18 users was analyzed which besides the order effect may be an explanation for the poorer prediction performance.

3. Experiment 2

3.1. Method

This experiment was conducted to find out if the results in [5] and Experiment 1 are valid when using a system not having interaction (and thus possible interference) between the different modalities. In such cases participants can use the multimodal system like a unimodal one with using only the modality they like most. For systems like this, the modality not or rarely used should have a lower impact on the prediction of the multimodal ratings.

Furthermore, Experiment 1 showed that changing the order is of negative effect for the predictions: Accuracy and stability were lower than reported in [5]. If users are just summing up the unimodal judgments this should be influenced by memory capacity. Generally memory performance decreases with age [8]. Therefore we compared young and old users, expecting better models for the younger participants

3.1.1. Participants and material

Fifteen younger (<25 years/ M=29 y.) and fifteen older (>55 years/ M= 66 y.) users took part in the study.

The application tested was a multimodal mailbox system capable of handling speech-, e-mail- and fax-messages as well as call forwarding and notifications of mailbox messages. It was implemented on a smart-phone (HTC Touch Diamond) controllable via motion (tilt and twist), speech and touch screen. System output was graphical for all modalities. For motion control, additional tactile feedback and for speech control, additional auditive feedback was given.

3.1.2. Procedure

The participants had to execute 4 blocks of tasks with a total of 14 tasks (get messages, reply to them, forward, and sort messages as well as changing notification options).

First, participants were asked to solve all tasks with a given modality. Afterwards, participants rated the interaction via the AttrakDiff questionnaire. This was repeated for all three modalities, the sequence of the modalities was balanced between the participants. In the final block, participants were free to choose the modalities they used for solving the task. Here it was always possible to switch or to combine modalities after each task and also within a task. Again, the participants had to fill in the AttrakDiff after completing all tasks in this condition.

3.2. Results

Again stepwise multiple linear regression analysis was carried out for each scale and subscale (s. Table 2). Similar to the previous results best predictions were observed for overall judgments and the scale measuring *Hedonic-Quality Stimulation*. But in contrast to the previous experiments, prediction for the scale *Hedonic Quality-Identity* was relatively good and prediction for the global scale *Attractiveness* was relatively bad compared to the other scales. Again poorest results were obtained for the scale *Pragmatic Quality*. Overall, the models were much more accurate than in Experiment 1. For none of the predictions motion was included. Also speech was only included for one scale (*Hedonic Quality-Stimulation*).

Table 2. *Results of multiple linear regression analyses for Experiment 2 using all cases. Only significant parameters (p<.05) are included*

Scale	Touch				Speech				R^2	RMSE	F (df)
	B	SE B	β	t (df)	B	SE B	β	t (df)			
Overall	.693	.105	.844	7.20 (23)	-	-	-	-	.693	.475	51.85 (1,23)
ATT	.794	.140	.750	5.66 (25)	-	-	-	-	.562	.680	32.10 (1,25)
HQ-S	.598	.114	.606	5.24 (24)	.354	.105	.389	3.37	.862	.397	68.95 (2,24)
HQ-I	.749	.202	.830	7.44 (26)	-	-	-	-	.689	.452	55.36 (1,26)
PQ	.489	.128	.601	3.83 (24)	-	-	-	-	.361	.771	14.70 (2,24)

Regarding four out of the five scales only the judgments for touch were included in the predictions. Thus judgments for speech and motion could not explain a significant part of the variance in the judgment for multimodal system version. The actual modality usage is in accordance with these results: Touch was used most (68%) All other modalities were chosen much fewer (speech 19 %; motion 7%; combination 6%).This supports our assumption that for multimodal systems with no potential interference between modalities, the modality not used should have no impact on the multimodal judgments.

Leave-one-out cross-validation showed as in the previous studies the least stabile models for the scale *Pragmatic Quality*. Again prediction based on the scale *Hedonic Quality-Stimulation* was most stable (cf. Table 3).

Table 3. *Results of leave-one-out cross-validation for Experiment 2 by age group and overall cases.*

Scale	R^2			RMSE		
	Young	Old	All	Young	Old	All
Overall	.623	.403	.578	.225	.537	.317
ATT	.524	.065	.486	.451	.959	.445
HQ-S	.748	.615	.854	.151	.738	.170
HQ-I	.701	.308	.653	.170	.435	.214
PQ	.056	$-.232^2$=.054	.219	.651	1.541	.611

To test for memory capacity effects multiple linear regression was conducted for each age group separately. As hypothesized, predictions were less accurate for older participants (cf. Table 4)

3.3. Discussion

Results showed that the memory capacity has an influence on predictions. But since we did not measure the actual memory capacity of the older participants, another age effect might have been causing this result. Moreover it might have been that the older users were not only unable to memorize their previous judgments but were also unable to judge the interaction correctly, which would consequently decrease the prediction.

Furthermore our hypothesis that for systems with no interaction between the modalities the judgments of the most used modality match to a large extent the judgment for the overall system is supported.

4. General Discussion and Conclusion

The results observed in [5] are partially supported by the reported experiments. As in [5], prediction was best for overall and general judgments and judgments on the scale *Hedonic Quality-Stimulation.* The poor prediction performance for the scale *Hedonic Quality-Identity* in Experiment 1 might be explained by the underlying construct measured via this scale.

As mentioned before the scale *Hedonic-Quality Identity* measures a product's ability to express the owner's self. The room information and management system we tested in Experiment 1 was not developed (and is not available) for personal use. Moreover it was custom tailored for Deutsche Telekom Laboratories (T-Labs). None of the participants was employed at T-Labs. Thus this system was in none of its versions designed to promote self expression or identification by communicating personal values. Furthermore the test was very task-oriented and goal-oriented. All tasks referred to the system's functionality for daily business of T-Labs employees and had no actual relevance for the participants. Data confirmed that users rated neutral on this scale. The mean differed between 0.3 and 1 on a scale from -3 to 3. Thus this scale might have been an inappropriate measure especially for testing the room information system.

For the smart-phone used in Experiment 2 it can be assumed that it has, at least to some extent, the ability to provide identification. But again the test was goal-oriented with predefined tasks and stimuli given by the experimenter. The content of the messages and notifications they dealt with during the test had no personal meaning for them. Thus this scale might have been an inappropriate measure especially for testing the room information system.

Scale		Touch				Speech				Motion				R^2	RMSE	F (df)
		B	SE B	β	t (df)	B	SE B	β	t (df)	B	SE B	β	t (df)			
Overall	Young	.830	.139	.866	5.99 (12)	-	-	-	-	-	-	-	-	.749	.413	35.85 (1,12)
	Old	.803	.213	.800	3.78 (8)	-	-	-	-	-	-	-	-	.641	.573	14,26 (1,8)
ATT	Young	.903	.184	.806	4.91 (13)	-	-	-	-	-	-	-	-	.650	.614	24.14 (1,13)
	Old	.846	.260	.717	3.25 (11)	-	-	-	-	-	-	-	-	.514	.742	10.58 (1,11)
HQ-S	Young	.403	.104	.444	3.87 (12)	.447	.082	.630	5.49 (12)	-	-	-	-	.884	.280	45.93 (2,12)
	Old	1.059	.089	1.073	11.94 (7)	-	-	-		-.360	.138	-.235	-2.61 (7)	.957	.294	78.52 (2,7)
HQ-I	Young	.851	.131	.875	6.51 (13)	-	-	-	-	-	-	-	-	.765	.390	42.34 (1,13)
	Old	.709	.182	.777	3.91 (10)	-	-	-	-	-	-	-	-	.604	.533	15.25 (1,10)
PQ	Young	.539	.182	.649	2.96 (12)	-	-	-	-	-	-	-	-	.422	.747	8.74 (1,12)
	Old	.500	.209	.585	2.39 (11)	-	-	-	-	-	-	-	-	.343	.823	5.73 (1,11)

Table 4. *Results of multiple linear regression analyses for Experiment 2 by age group. (p<.05)*

However, for both experiments no such explanation can be found regarding the scale *Pragmatic Quality,* measuring to a large extent the concept of usability. Further research how single modalities influence ratings of multimodal systems should be conducted at this point.

The hypothesis that prediction performance is partially a consequence of the participants' effort to rate consistently was supported. Both order of the system version and memory capacity seemed to be of influence. The effect of order implicates that within test designs have to be considered as critical for evaluation studies. Based on the observation of the influence of age it can be concluded that, at least for old users, post-hoc questionnaires are not suitable for valid evaluation.

Generally it has to be remarked that for all experiments the sample size was very small. More accurate results could be obtained with a larger sample. Furthermore all results are based on only one questionnaire. As shown for the scale *Hedonic Quality-Identity* results are heavily dependent on the appropriateness of construct intended to measure.

5. References

[1] Sarter, N. B. Multimodal information presentation: Design guidance and research challenges. *International Journal of Industrial. 36* (5), 439-445, 2006

[2] Jokinen, K. User interaction in mobile navigation applications. In L. Meng, A. Zipf and S. Winter (Eds.), *Map-Based Mobile Services: Design, Interaction and Usability*, *Lect. Notes in Geoinformation and Cartography*, pp. 168-197. Springer, 2008

[3] Hassenzahl, M., Burmester, M. and 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*, Stuttgart: B.G. Teubner, 187-196, 2001

[4] Hassenzahl, M. The thing and I: understanding the relationship between user and product. In M. A. Blythe, K. Overbeeke, A. F. Monk, and P. C. Wright (Eds.), *Funology: From Usability To Enjoyment*, Kluwer Academic Publishers, Norwell, 31-42, 2004

[5] Wechsung, I., Engelbrecht, K.-P., Schaffer, S., Seebode, J., Metze, F. and Möller, S. Usability Evaluation of Multimodal Interfaces: Is the whole the sum of its parts?, accepted for: *13th International Conference on Human-Computer Interaction*, 2009

[6] Kamii, C., Lewis, B. A., and Kirkland, L. D. Fluency in subtraction compared with addition. *Journal of Mathematical Behavior, 20*, 33-42, 2001

[7] Dixon J. A., Deets, J. K., and Bangert, A. The representation of the arithmetic operations include functional relationships. *Memory & Cognition*, 29, 462-477, 2001

[8] Spencer W.D. and Raz N., Differential effects of aging on memory for content and context: a meta-analysis. *Psychology & Aging. 10* (4):527–539, 1995