Usability Evaluation of Multimodal Interfaces: Is the Whole the Sum of Its Parts?

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

[1] Deutsche Telekom Laboratories, TU Berlin
Ernst-Reuter-Platz 7, 10587, Berlin
[2] InterACT center, Carnegie Mellon University,
Pittsburgh, PA
Ina.wechsung@telekom.de

Abstract. Usability evaluation of multimodal systems is a complex issue. Multimodal systems provide multiple channels to communicate with the system. Thus, the single modalities as well as their combination have to be taken into account. This paper aims to investigate how ratings of single modalities relate to the ratings of their combination. Therefore a usability evaluation study was conducted testing an information system in two unimodal versions and one multimodal version. Multiple linear regression showed that for overall and global judgments ratings of the single modalities are very good predictors for the ratings of the multimodal system. For separate usability aspects (e.g. hedonic qualities) the prediction was less accurate.

1 Introduction

Since human communication is multimodal in nature multimodal systems are expected to provide adaptive, cooperative and flexible interaction [1]. By providing multiple communication channels such systems are assumed to support human information processing by using different cognitive resources [2, 3].

But making a system multimodal by just adding a further modality to a unimodal system might not necessarily lead to improvement [4]. A higher cognitive load due to more degrees of freedom may be the result [5]. Furthermore, the different modalities may interfere with each other [5]: When presenting identical information via two modalities (e.g. reading and listening to the same text simultaneously) a synchronization problem can occur [6]. Moreover, if different modalities refer to the same cognitive resources task performance may decrease [3].

Apparently, usability evaluation of multimodal systems is a complex issue. The single modalities as well as their combination have to be taken into account. Established procedures usually cover only specific modalities [e.g. 7,8] and evaluating multimodal systems by combining weighted judgements of single modalities is difficult [9].

In the current study an information system is evaluated in two unimodal versions and one multimodal version. The aim is to investigate how user ratings of the single modalities relate to the rating of the multimodal system.

J.A. Jacko (Ed.): Human-Computer Interaction, Part II, HCII 2009, LNCS 5611, pp. 113–119, 2009.

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 the study.

The system tested is a wall-mounted information and room management system controllable via a graphical user interface (GUI) with touch input, via speech input and via a combination of both. The output is always given via GUI.

2.2 Procedure

The users performed six different tasks with the system. To collect user ratings the AttrakDiff questionnaire [10] was used.

Each test session took approximately one hour. Each participant performed the tasks with each system version. Participants were instructed to perform the tasks with a given modality. After that, they were asked to fill out the AttrakDiff in order to rate the previously tested version of the system. This was repeated for every modality. In order to balance fatigue and learning effects the order of the systems was randomized. After that, the tasks were presented again and the participants could freely choose the interaction modality. Again the AttrakDiff had to be filled out to rate the multimodal system.

The 4 AttrakDiff sub-scales comprising 7 items each (*pragmatic quality, hedonic quality-stimulation, hedonic quality-identity, attractiveness*) were calculated according to [10]. Furthermore an overall scale was calculated based on the mean of all 28 items. All questionnaire items which were negatively poled were recoded so that higher values indicate better ratings.

To analyze which modality the participants preferred when using the multimodal system version, the modality chosen first to perform the task was annotated. This way, the frequencies of modality usage were assessed.

3 Results

3.1 Rating for Different System Versions

The results show differences between the three versions of the system for all AttrakDiff scales. For the scale *pragmatic qualities* the touch-based version was rated best and the voice control version worst ($F(2,66) = 93.79$, $p = .000$, $eta^2 = .740$). For both hedonic scales the multimodal version was rated best. Regarding *hedonic qualities-stimulation* ($F(2,68) = 12.84$, $p = .000$, $eta^2 = .274$) the speech version received the lowest ratings. For *hedonic qualities-identity* the touch-based version was rated worst ($F(1.65, 55.99) = 15.35$, $p = .000$, $eta^2 = .311$)[1].

The *attractiveness scale*, the AttrakDiff scale covering pragmatic as well as hedonic qualities, showed the lowest ratings for the speech-based version ($F(1.51,$

[1] Greenhouse-Geisser-correction was applied to control for violation of the sphericity assumption.

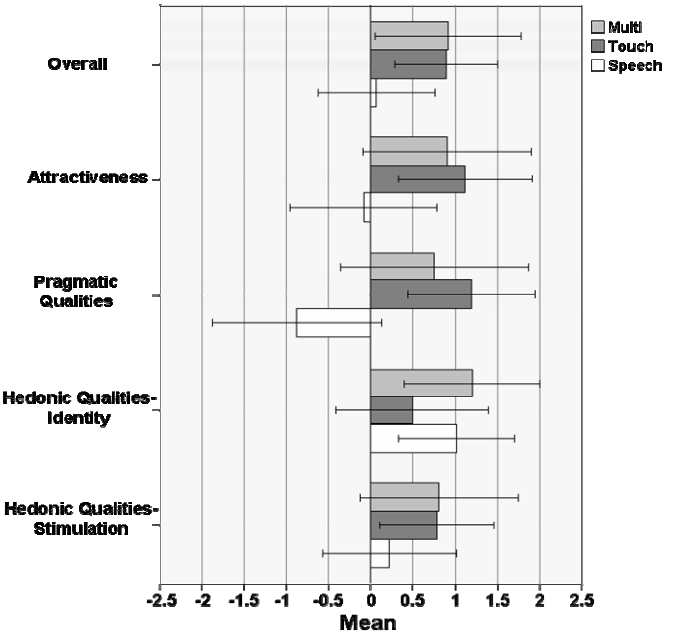

Fig. 1. Ratings on AttrakDiff overall scale and AttrakDiff subscales for all system versions. Error bars display one standard deviation.

51.22)= 47.53, p=.000, eta²=.583)[1] and highest ratings for the touch-based version. Regarding the overall scale, the scale based on the mean of all items, the speech-based version was rated worse than the touch-based and multimodal systems versions. The touch-based version and the multimodal version were rated equally good.

Differences between male and female user were not observable.

3.2 Relationship between Uni-and Multimodal Judgments

To investigate if and how the ratings of the unimodal system versions relate to ratings for the multimodal system version stepwise multiple linear regression analysis was conducted for each sub-scale and the overall scale. The judgments assessed after the interaction with the unimodal systems version were used as predictor variables, the judgments collected after interacting with the multimodal system version were used as the response variable.

The results show that for the attractiveness scale and the overall scale the judgments of the unimodal system are very good predictors of the judgments of the multimodal version. For both regression analyses the beta–coefficients were higher for the judgments of the touch-controlled version of the system. This is in line with the modality usage for the multimodal system: Touch-input was used more frequently. Thus the overall and global judgments of the multimodal system should be more influenced by the interaction with the touch-input.

Table 1. Results of multiple linear regression analysis using all data (*p<.01)

Scale	Touch				Speech				R^2	RMSE	F (df)
	B	SE B	β	t (df)	B	SE B	β	t (df)			
Overall	.805	.112	.566	0.21* (32)	.680	.098	.546	6.91* (32)	.829	.370	74.94* (2,31)
Attractiveness	.845	.094	.684	9.02* (32)	.478	.087	.419	5.52* (32)	.837	.411	81.99* (2,32)
Pragmatic Qualities	.797	.174	.537	4.57* (31)	.468	.130	.421	3.59* (31)	.628	.703	26.19* (2,31)
Hedonic Qualities Stimulation	.689	.134	.521	5.13* (32)	.633	.119	.536	5.31* (32)	.693	.508	36.05* (2,32)
Hedonic Qualities Identity	.282	.106	.331	2.66* (32)	.661	.144	.572	4.60* (32)	.612	.527	25.24* (2,32)

Regarding the hedonic qualities scales and the pragmatic qualities scale between 61 and 69 percent of the variance could be explained by using the ratings of the unimodal systems as predictors of the ratings for the multimodal system. The beta–coefficients of speech were higher than those of touch for both hedonic scales, therefore the rating of speech had a larger impact on the multimodal system judgment than the judgment on touch

A 10 fold cross validation was conducted to test for overfitting effects. For the *attractiveness scale* and the overall scale R^2 is still around .8 indicating a good fit. For the other scales the overfitting effects were larger, resulting in the worst accuracy for *hedonic qualities-identity*. Except for the pragmatic scale the models with beta-coefficients in line with the actual usage were more stable. The detailed results are given in Table 2 and visualized in Figure 2.

Table 2. Results of multiple linear regression analysis using 10 fold cross validation (*p<.01)

	Overall	Attractiveness	Pragmatic Qualities	Hedonic Qualities Stimulation	Hedonic Qualities Identity
R^2	.799	.805	.539	.607	.391
RMSE	.384	.431	.754	.572	.615

For the models with lower R^2, we also analyzed the ad-hoc assumption that the best- (*pragmatic qualities*, both hedonic scales) or worst-rated (*hedonic qualities-identity*) modality might determine the judgment of the multimodal system. However taking the maximum or minimum judgment as a predictor into a regression function did not produce a higher accuracy (s. Table3).

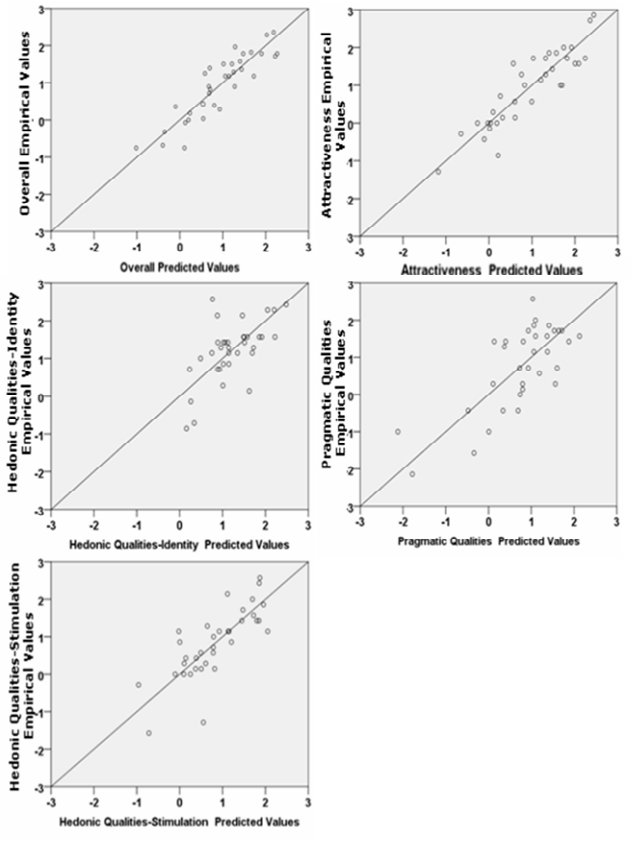

Fig. 2. Scatter plots for predicted values (after 10 fold cross validation) and empirical values

Table 3. Results of multiple linear regression analysis using the maximum or minimum judgement as predictor

Scale	**Max (Touch, Speech)**				**R^2**	**RMSE**	**F (df)**
	B	**SE B**	**β**	**t(df)**			
Pragmatic Qualities	1.002	.190	.688	5.37* (32)	.474	.823	28.81* (1,32)
Hedonic Qualities Stimulation	.944	.189	.657	5.01* (33)	.432	.705	25.09* (1,33)
Hedonic Qualities Identity	1.006	.143	.774	7.01* (33)	.598	.509	49.19* (1,33)
Scale	**Min (Touch, Speech)**				**R^2**	**RMSE**	**F (df)**
	B	**SE B**	**β**	**t (df)**			
Hedonic Qualities Identity	.560	.117	.640	4.79* (33)	.410	.617	22.95* (1,33)

4 Discussion

The current paper investigates how subjective judgments of unimodal system versions relate to subjective judgments of the multimodal version of the same system. It was shown that for overall and global measures (attractiveness scale) the judgments of the unimodal versions are good predictors for judgments of the multimodal version. Additionally the results indicate that the modality used more frequent in multimodal interaction has a higher influence on the judgment of multimodal version than the less frequent used modality. For more specific measures the prediction performance is lower.

Furthermore, in accordance with [4] it could be observed that adding a modality to a unimodal system does not automatically lead to better quality judgments. For the present study this means, that regarding overall and global judgments the whole is actually the sum of its parts. Ratings for the multimodal system are the sum of the ratings of the unimodal systems. However for scales measuring more specific constructs this assumption is not valid: Stabile predictions of the ratings for the multimodal systems based on the ratings of the unimodal systems were not possible. Hence further research is needed for multimodal measures of specific usability aspects.

Moreover this study is based on the results of one questionnaire only. Further research is needed to investigate if similar results would be obtained if performance measures were used. Additionally the findings are currently limited to the tested system and test design. For the multimodal system version interference between the modalities was possible (e.g.: the speech recognizer was occasionally unintentionally switched on by off-talk). Moreover 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.

So, in a follow-up study the order needs to be changed with the multimodal system version tested first.

References

1. Chen, F.: Designing Human Interface in Speech Technology. Springer, New York (2005)
2. Baddeley, A.D., Hitch, G.J.: Working memory. In: Bower, G. (ed.) The psychology of learning and motivation, pp. 47–89. Academic Press, New York (1974)
3. Wickens, C.D.: Multiple resources and performance prediction. Theoretical Issues in Ergonomics Science 3, 159–177 (2002)
4. Oviatt, S.L.: Ten myths of multimodal interaction. Communications of the ACM 42(11), 576–583 (1999)
5. Schomaker, L., Nijtmans, J., Camurri, A., Lavagetto, F., Morasso, P., Benoît, C., Guiard-Marigny, T., Le Goff, B., Robert-Ribes, J., Adjoudani, A., Defée, I., Münch, S., Hartung, K., Blauert, J.: A Taxonomy of Multimodal Interaction in the Human Information Processing System. A Report of the ESPRIT Project 8579 MIAMI. NICI, Nijmegen (1995)
6. Schnotz, W., Bannert, M., Seufert, T.: Towards an integrative view of text and picture comprehension: Visualization effects on the construction of mental models. In: Otero, J., Graesser, A., Leon, J.A. (eds.) The Psychology of Science Text Comprehension., pp. 385–416. Erlbaum, Mahwah (2002)

7. Kirakowski, J.: The software usability measurement inventory: background and usage. In: Jordan, P. (ed.) Usability Evaluation in Industry, pp. 169–177. Taylor & Francis, London (1996)
8. Hone, K.S., Graham, R.: Towards a tool for the Subjective Assessment of Speech System Interfaces (SASSI). Nat. Lang. Eng. 6(3-4), 287–303 (2000)
9. Beringer, N., Kartal, U., Louka, K., Schiel, F., Türk, U.: PROMISE: A Procedure for Multimodal Interactive System Evaluation. In: Procs of the Workshop Multimodal Resources and Multimodal Systems Evaluation, Las Palmas, Gran Canaria, Spain, pp. 77–80 (2002)
10. Hassenzahl, M., Burmester, M., Koller, F.: AttrakDiff: Ein Fragebogen zur Messung wahrge-nommener hedonischer und pragmatischer Qualität. In: Ziegler, J., Szwillus, G. (eds.) Mensch & Computer 2003, Interaktion in Bewegung, pp. 187–196. B.G. Teubner, Leipzig (2003)