Quality aspects of multimodal dialog systems: identity, stimulation and success

Christine Kühnel[1], *Benjamin Weiss, Matthias Schulz and Sebastian Möller*

Quality & Usability Lab,
Deutsche Telekom Laboratories,
TU Berlin, Germany
[1]christine.kuehnel@telekom.de

Abstract

So far, not much is known on the relationship of quality aspects of multimodal dialog systems. This paper aims at closing this gap by analyzing the influence of input and output modalities on the systems' usability. The underlying study has been carried out with a smart-home system offering speech, gesture and touch as well as the combination of these three for input and a speech-to-text system, a TV screen and a smartphone screen for output. The results indicate that the usability of a multimodal system is composed of hedonic and pragmatic aspects. The hedonic aspects are influenced by the identity transported by the output channels and the stimulation of the input modalities. A measure for task success was sufficient to describe the pragmatic aspect.

Index Terms: multimodal dialog systems, evaluation

1. Introduction

New user interfaces combining different modalities for interaction are on the rise – owing to the ascribed advantages discussed in numerous publications (e. g. [1, 2]). And with the technical advances and market growth in the field, evaluation and usability of uni-modal and multimodal dialog systems are becoming crucial issues [3].

Established design methods exist that can be (partially) transfered to the context of multimodal systems. But so far, evaluations, as a part or finish of the design process, have been mostly individual undertakings [4]. Furthermore, "work on multimodal usability remains poorly understood" [5].

This situation arises partly from a missing agreement on the aspects a systems' quality or usability is composed off and how these could be measured systematically. In [4] a taxonomy of quality aspects of multimodal human-machine interaction has been proposed in order to better understand and differentiate between these general constructs currently used when speaking about assessment or evaluation.

It is the aim of this paper to contribute to the ongoing research on the evaluation of multimodal dialog systems. To this end the concept of usability is discussed in the context of established definitions. Based on [4] influences of input and output on usability are analyzed. A multimodal dialog system offering a combination of speech, touch and gestural input is chosen for the underlying evaluation study. These input modalities can nowadays – with the wide-spread use of smart-phones offering spoken and touch input as well as the recent release of, for example, Microsoft's Kinect – be considered state-of-the-art. For output a text-to-speech system is used, as well as a TV screen and the touch screen of a smartphone.

2. Usability of multimodal dialog systems

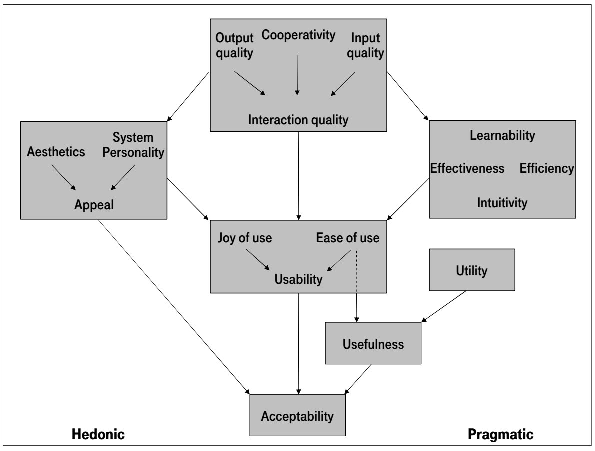

Figure 1: Taxonomy of quality aspects of multimodal human-machine interaction according to Möller et al. [4].

In the field of human-computer interaction focus has been for a long time on a systems' usability: the "extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" [6]. But, "the state of the art in spoken multimodal and mobile systems usability and evaluation remains uncharted to a large extent" according to Dybkjær et al. [3]. Möller and colleagues tried to address this problem in [4]. In Figure 1 an excerpt from their taxonomy of multimodal quality aspects is shown. The view taken on usability is a very broad one. It follows roughly the definition stated above by incorporating effectiveness, efficiency, as well as learnability and intuitivity via ease-of-use. Satisfaction is not named explicitly but is addressed by the joy-of-use component of usability – as is "User eXperience" (UX), which is not listed either. UX has been defined as "a person's perceptions and responses that result from the use or anticipated use of a product, system or service" [7]. The wrapping-up of UX in joy-of-use is analogous to the interpretation of UX as "an elaboration of the satisfaction component of usability" discussed in [8]. Joy-of-use specifically addresses those aspects of a user interface that appeal to a person's desire of pleasure – aspects that are fun, original, interesting, engaging, and cool. Thus, the joy-of-use component of usability is similar to hedonic quality as described in [9]. At the same time, ease-of-use could be related to pragmatic quality.

Therefore, the AttrakDiff [10], with its subscales 'attractiveness', 'pragmatic quality', 'hedonic quality-identity' and 'hedonic quality-stimulation', would be the questionnaire measuring a concept closest to the understanding of usability described here. 'Attractiveness' has been applied before to measure usability of multimodal systems [11, 12]. As ease-of-use is influenced by efficiency and effectiveness, interaction parameters, such as task success and dialog duration, could be expected to be suitable metrics for 'pragmatic quality' [11, 12].

The importance of aesthetics has been emphasized by Tractinsky [13]. It appears that hedonic quality has an impact on perceived usability in graphical user interfaces (GUI). In [14] Hassenzahl and colleagues suggested that the appeal of a product might be influenced by ergonomic as well as hedonic aspects, the latter addressing human needs for novelty or change and social power (status). They describe hedonic aspects as being induced, for example, by visual design, sound design, novel interaction techniques, or novel functionality.

Another approach to understanding multimodal quality was presented by Wechsung and colleagues [15] who analyzed the influence of unimodal quality on multimodal system quality. They found that a simple linear combination of unimodal judgments weighted with the actual (sequential) usage of the single modalities in the multimodal system is suitable to predict multimodal quality to a high degree.

Apart from the influence of hedonic and pragmatic aspects on usability, the taxonomy suggests that input, output and interaction quality relate to usability. In [16, 17] it could be shown that the overall quality of a dialog system is strongly related to the quality of its output component: a text-to-speech and an audio-visual system.

Our approach further analyzes the quality aspects and their interrelationship as stated in the taxonomy. The system used and the method applied is described below.

3. Methodology

For the experimental study a smart-home system was used offering sequential use of voice, smartphone-based input (touch) and three-dimensional free-hand gesture control (gesture). This system is set up inside a fully functional living room. Possible interactions include the control of lamps and blinds, the TV, an IP-radio, an electronic program guide (EPG), video recorder, and hi-fi system. Furthermore, the system offers an archive for music and supports the generation of playlists. The TV program, available radio stations, lists of recorded movies, an overview of the users' music (sorted by album, artist, etc.) or the playlist are displayed on the TV screen. Those lists are also displayed on the smartphone to allow touch input for the selection of list entries and the execution of corresponding actions, such as recording a movie or deleting a song from the playlist.

A German male voice was chosen for the TTS system. Thus, three different output channels are employed (TTS, GUI on the touch screen and lists on the TV screen), in some cases in parallel offering complementary or redundant information. In order to keep input accuracy comparably high for all modalities, speech recognition was replaced by a transcribing wizard. Participants were told that there was a speech recognizer in place, and a lapel microphone was used to further strengthen this impression.

A simple graphical user interface was developed and implemented on an Apple iPhone 3GS which communicated via wireless LAN with the smart-home system. To control the lamps, blinds, radio and TV the corresponding button on the main screen had to be pressed with a fingertip. This opens a list of options available for the respective device. Further buttons open a music archive, a music playlist, or an overview of recorded movies. List navigation was possible via scrolling (slide finger across screen) and selecting (touching an entry).

Table 1: Gesture-command mapping.

Gesture	Command	Device
Swing up	Volume up Brighter Open	TV, Radio Lamps Blinds
Swing down	Volume down Dim Close	TV, Radio Lamps Blinds
Point forward	Turn on/off	TV, Radio
Swing to the right	Next channel Stop	TV, Radio Blinds
Swing to the left	Previous channel	TV, Radio

A camera-based gesture recognition for simple and often used gestures (TV, radio, lamps, blinds) was simulated by placing two cameras in front of the participants at a distance of approximately two meters below the TV screen. The actual recognition was done by the wizard who could monitor the participant via the cameras and enter the recognition result as attribute-value pairs (e. g. [device:blinds; action:down]) into the system. A set of five three-dimensional gestures was used in this experiment (see Table 1). By pointing towards a device with the hand this device is selected. The same gesture could thus be used to initiate the same effect for different devices. Reusing the same concept for different system controls reduces the gesture set considerably. For more detailed information please refer to [12].

3.1. Test design

3.1.1. Participants

We asked 17 young and 16 older adults to participate in the study. The younger group of participants (20–29 years, M=26, SD=2.74, 9 female) was recruited at the university campus. The older group of participants (51-67 years, M=59, SD=4.60, 9 female) was recruited via notices placed in supermarkets and near employment offices. All participants were paid for their time. None of the participants was familiar with the system used in the study.

3.1.2. Procedure

The main experiment was split into three parts: judgment of A) system output (passive), B) unimodal input (interactive) and C) multimodal input (interactive).

In the first part (Part A) participants were asked to rate each of the three different output channels (TTS, touch screen and TV screen) after the presentation of three to seven examples of one output channel in a row. According to [18] it is sufficient to show a web page for less than one second to judge its aesthetics. Thus, each interface was presented only very shortly to the participants.

In the second part (Part B) the participants were guided through three identical task-based interactions, each time using a different input (touch, voice and gesture). The tasks were short, simple, and closely defined, such as ”Lower the blinds and stop them midway.” or ”Turn on the radio and switch to the next station.”. This part was used to collect judgments for each input modality and to train the participants in the use of the modalities and the system. The sequence of output and input in Part A and B followed a full Latin square design to counterbalance order effects.

In the last part (Part C) the user was guided by four tasks displayed one at a time on the screen in front of them. This time participants could choose freely which modality they wanted to use and change the modality whenever they felt like it. The first task consisted of all the interactions that had been conducted in Part B, but in this part the subtasks were less precisely defined (e.g. ”Find a radio station you like”). The second and third task asked the participants to do something they had not done before, such as recording a movie or adding songs to their playlist. These tasks could not be solved via gestural interaction. As participants were not explicit informed about this, some tried nonetheless. The fourth task was open; users were asked to ’play’ with the system, try something they had not done yet or use a modality they had not used often.

3.1.3. Assessments

All participants were asked for their judgments of the three output channels (Part A), the three unimodal input channels (Part B) and the multimodal interface (Part C) via a short version of the AttrakDiff questionnaire [10], resulting in seven questionnaires filled in per participant (3,3,1). The AttrakDiff questionnaire contains antonym pairs rated on a 7-point scale ([-3,+3]), yielding the subscales ’Attractiveness’ (ATT), ’Pragmatic Qualities’ (PQ), ’Hedonic Quality – Stimulation’ (HQS) and ’Hedonic Quality – Identity’ (HQI).

According to [19] overall ’attractiveness’ (i. e., valence, beauty) is the result of a simple linear combination of ’pragmatic qualities’ (i. e., simple and functional), ’hedonic quality-stimulation’ and ’hedonic quality-identity’. Of the hedonic qualities, ‘identity’ describes how well a user identifies with the product. ‘Stimulation’ indicates the extent to which a product supports the needs to develop and move forward by offering novel, interesting and stimulating functions, contents, interactions and styles of presentation.

4. Results

Based on [14] we assume a simple model for the usability measure ‘attractiveness’ (ATT), described as a linear combination of ‘pragmatic quality’ (PQ) and the hedonic qualities ‘identity’ (HQI) and ‘stimulation’ (HQS):

$$\text{ATT} = \beta_{PQ} \cdot \text{PQ} + \beta_{HQI} \cdot \text{HQI} + \beta_{HQS} \cdot \text{HQS} \qquad (1)$$

The weights (β_{PQ}, β_{HQI}, β_{HQS}) indicate the respective importance of the attributes and the goodness of the model can be described using Pearson’s R measure. For the multimodal system (mm, part C) a linear regression with $\beta_{PQ} = .33$, $\beta_{HQI} = .40$, and $\beta_{HQS} = .30$ correlates strongly with ATT$_{mm}$ ($R = .92$).

Now, output quality is analyzed by examining the questionnaire results of Part A concerning the weights of PQ, HQI and HQS. We assume that the ‘attractiveness’ (ATT$_o$) of the three output channels would be highly dependent on the ‘hedonic quality-identity’ (HQI$_o$) as it is the ‘skin’ of the interfaces which is presented for evaluation and which is affecting a possible identification with the system. The data shown in Table 2 confirms this assumption. ‘Stimulation’ (HQS$_o$) has no influence at all and the impact of ‘pragmatic quality’ (PQ$_o$) is clearly weaker than the impact of HQI$_o$.

Table 2: Influence of PQ$_o$, HQI$_o$ and HQS$_o$ on ATT$_o$

output	β_{PQ}	β_{HQI}	β_{HQS}	R
TTS	.30	.68	**.00**	.84
touch screen	.38	.64	**.00**	.85
TV screen	.36	.78	**.00**	.92

In Part B the input modalities were rated but here an interaction took place. As the interaction was task-guided, a strong impact of ‘pragmatic quality’ on ‘attractiveness’ (ATT$_i$) should be expected [20]. Again, the model introduced above (Eq. 1) is used to analyze the ratings of the input modalities. The results are displayed in Table 3. When interacting with gestures hardly any output was given, apart from device response, such as lamps turning on. This explains why ‘identity’ has no impact on the rating of the gesture interface. But gestures are relatively new as an interaction mode, therefore, the interface is rated as highly stimulating. Interestingly, for spoken input the pragmatic term disappears and ‘identity’ has a stronger influence than stimulation. For touch interaction ‘identity’ is again the strongest influence, accompanied by ‘pragmatic quality’. This can probably be explained by the overall impression of the smartphone interface. It is already an established interface (not stimulating) but is seen as stylish and worthwhile.

Table 3: Influence of PQ$_i$, HQI$_i$ and HQS$_i$ on ATT$_i$

input	β_{PQ}	β_{HQI}	β_{HQS}	R
gesture	.52	**.00**	.50	.86
voice	**.00**	.60	.35	.91
touch	.35	.63	**.00**	.86

Finally, the influence of the output and input modalities on the overall impression of the multimodal system is examined. We expected that a linear combination of the hedonic quality aspect ‘identity’ of the output channels (TTS, touch screen and TV screen) would be a good predictor for ‘hedonic quality-identity’ of the multimodal system. It turned out that the rating of the touch screen (HQI$_{o-t}$) is sufficient.

‘Hedonic quality-stimulation’ should relate to the interfaces offered for user input, namely voice, touch and gesture. Here, the rating of the spoken input (HQS$_{i-v}$) is the only metric related to multimodal stimulation.

In [21] parameters describing multimodal interaction have been defined. For ‘pragmatic quality’ we assumed that efficiency (e. g., dialog duration) and effectiveness (e. g., task success) measures would be appropriate metrics. In fact, task success ts, measured by the number of unsuccessful tasks, shows the strongest correlation.

In order to integrate ts, the data was transformed to achieve a standardized normal distribution (Z-transformation, $M = 0$ and $SD = 1$):

$$\text{ATT}_{mm} = -.31 \cdot ts + .51 \cdot \text{HQI}_{o-t} + .27 \cdot \text{HQS}_{i-v} \qquad (2)$$

The final model correlates not as strong with ATT$_{mm}$ as the model with factors from the same questionnaire, but is still highly significant ($R = .76$).

5. Discussion

Our findings provide evidence for the claim that the usability of a multimodal dialog system as measured by the AttrakDiff subscale 'attractiveness' is dependent on pragmatic and hedonic aspects. Furthermore, a clear relationship between the hedonic qualities of input and output channels with the overall 'attractiveness' of the multimodal system could be shown.

The 'attractiveness' of the output modalities is mostly defined by their contribution on 'hedonic quality-identity' followed by 'pragmatic quality'. It is the 'look and feel' that determines how much the user identifies with the system. Furthermore, the first impression of an interface is already a good indicator of its' usability, as has been claimed by [13]. And, it appears that the approach described in [18] of presenting an interface for a short time is sufficient for a good prediction of both the final hedonic quality and the usability for multiple interfaces as well.

The characteristics of input quality depend very much on the interface in question. As input quality can only be judged after an interaction, we expected 'pragmatic quality' to be of high importance. But this is only the case for gestural interaction. And here it is nearly of the same importance as 'hedonic quality-stimulation'. This might be due to the comparatively high novelty of this interface. For spoken input, 'pragmatic quality' has no impact at all. This interface, although not new, is quite unfamiliar to most users which might explain the high impact of HQS. And participants showed either a strong liking or disliking of this interface. While some experienced spoken input to be comfortable, others said that they 'felt funny speaking to the air'. Thus, some participants might show a strong identification with this interface while others react negatively. The evaluation of the touch-based input seems to be strongly influenced by the device offered for input, namely an iPhone. This is supported by a moderate correlation of a Pearson's r=.40 (p=.019) between HQI$_{o-t}$ and HQI$_{i-t}$. Finally, we found that the pragmatic aspect of 'attractiveness' can be measured to a good extend by a metric for task success.

6. Conclusions

Analyzing a study with a multimodal system – offering sequential spoken, gestural and smartphone-based touch input and spoken as well as two kinds of graphical output – a simple model for the systems' usability was found. It is possible to predict multimodal usability – as measured, for example, by 'attractiveness' – based on the 'identity' transported by its' output channels, the 'stimulation' offered by the input modalities and interaction parameters such as task success (a measure for 'pragmatic quality'). The parametrization of 'pragmatic quality' by the number of task failures, 'hedonic quality-identity' (HQI) of the output by the HQI of the touch screen, and 'hedonic quality-stimulation' (HQS) of the input by HQS of speech is certainly specific to our system. Nonetheless, we argue that the general model is theoretically well-founded and we are eager to test for generalizability with other multimodal systems.

7. References

[1] S. Oviatt, "Multimodal interfaces," in *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications*. Hillsdale, USA: L. Erlbaum Associates Inc., 2003, pp. 286–304.

[2] R. López-Cózar Delgado and M. Araki, *Spoken, multilingual and multimodal dialogue systems: Development and assessment*. Chichester: John Wiley & Sons, 2005.

[3] L. Dybkjær, N. O. Bernsen, and W. Minker, "Evaluation and usability of multimodal spoken language dialogue systems," *Speech Communication*, vol. 43, pp. 33–54, 2004.

[4] S. Möller, K.-P. Engelbrecht, C. Kühnel, I. Wechsung, and B. Weiss, "Evaluation of multimodal interfaces for ambient intelligence," in *Human-Centric Interfaces for Ambient Intelligence*, H. Aghajan, R. López-Cózar Delgado, and J. C. Augusto, Eds. Amsterdam: Elsevier, 2010, pp. 347–370.

[5] N. O. Bernsen and L. Dybkjær, *Multimodal Usability*. London, UK: Springer-Verlag, 2009.

[6] ISO 9241-11, "Ergonomic requirements for office work with visual display terminals. Part 11: Guidance on usability," International Organization for Standardization (ISO), Switzerland, 1999.

[7] ISO 9241-210, "Ergonomics of human system interaction – Part 210: Human-centred design for interactive systems (formerly known as 13407)," International Organization for Standardization (ISO), Switzerland, 2010.

[8] N. Bevan, "What is the difference between the purpose of usability and user experience evaluation methods?" in *Proc. UXEM'09 Workshop, INTERACT*, 2009.

[9] M. Hassenzahl, "User experience (UX): Towards an experiential perspective on product quality," in *Proc. International Conference of the Association Francophone d'Interaction Homme-Machine*. New York, USA: ACM, 2008, pp. 11–15.

[10] M. Hassenzahl and A. Monk, "The inference of perceived usability from beauty," *Human-Computer Interaction*, vol. 25, no. 3, pp. 235–260, 2010.

[11] A. Naumann and I. Wechsung, "Developing usability methods for multimodal systems: The use of subjective and objective measures," in *Proc. VUUM*, 2008, pp. 8–12.

[12] C. Kühnel, B. Weiss, and S. Möller, "Evaluating multimodal systems – A comparison of established questionnaires and interaction parameters," in *Proc. NordiCHI*, 2010, pp. 286–293.

[13] N. Tractinsky, A. S. Katz, and D. Ikar, "What is beautiful is usable," *Interacting with Computers*, vol. 13, pp. 127–145, 2000.

[14] M. Hassenzahl, A. Platz, M. Burmester, and K. Lehner, "Hedonic and ergonomic quality aspects determine a software's appeal," in *Proc. SIGCHI*, 2000, pp. 201–208.

[15] I. Wechsung, K.-P. Engelbrecht, A. Nauman, S. Schaffer, J. Seebode, F. Metze, and S. Möller, "Predicting the quality of multimodal systems based on judgements of single modalities," in *Proc. Interspeech*, 2009, pp. 1827–1830.

[16] S. Möller and J. Skowronek, "Quantifying the impact of system characteristics on perceived quality dimensions of a spoken dialogue service," in *Proc. European Conference on Speech Communication and Technology*, 2003, pp. 1953–1956.

[17] C. Kühnel, B. Weiss, and S. Möller, "Talking heads for interacting with spoken dialog smart-home systems," in *Proc. Interspeech*, 2009, pp. 304–307.

[18] N. Tractinsky, A. Cokhavi, M. Kirschenbaum, and T. Sharfi, "Evaluating the consistency of immediate aesthetic perceptions of web pages," *International Journal on Human-Computer Studies*, vol. 64, pp. 1071–1083, November 2006.

[19] M. Hassenzahl, "The interplay of beauty, goodness, and usability in interactive products," *Human-Computer Interaction*, vol. 19, pp. 319–349, December 2008.

[20] M. Hassenzahl, R. Kekez, and M. Burmester, "The importance of a software's pragmatic quality depends on usage modes," in *Proc. WWDU*, 2002, pp. 275–276.

[21] C. Kühnel, B. Weiss, and S. Möller, "Parameters describing multimodal interaction – Definitions and three usage scenarios," in *Proc. Interspeech*, 2010, pp. 2014–2017.