



Parameters Describing Multimodal Interaction – Definitions and Three Usage Scenarios

Christine Kühnel and Benjamin Weiss and Sebastian Möller

Quality and Usability Lab, Technische Universität Berlin, Germany

[christine.kuehnel, bweiss, sebastian.moeller]@telekom.de

Abstract

While multimodal systems are an active research field, there is no agreed-upon set of multimodal interaction parameters, which would allow to quantify the performance of such systems and their underlying modules, and would therefore be necessary for a systematic evaluation. In this paper we propose an extension to established parameters describing the interaction with spoken dialog systems [1] in order to be used for multimodal systems. Focussing on the evaluation of a multimodal system, three usage scenarios for these parameters are given.

Index Terms: multimodal interaction, interaction parameters, evaluation

1. Introduction

One of the main approaches to the evaluation of human-computer interaction (HCI) is to parametrize individual interactions on the basis of manually annotated or automatically logged data. Such data can then be used for a predictive evaluation, as it is proposed by the PARADISE framework [2, 3, 4]. Apart from (1) modeling user satisfaction, such a description of individual interaction can serve to (2) find problems during the interaction in order to improve system modules to increase cooperativity, effectiveness, or efficiency, and (3) analyze the interaction to gain insight in human behavior and individual differences (e.g. to define user groups or task factors). For more than two decades of experience with spoken dialog systems, researcher and developer have defined, used, and evaluated so-called interaction parameters for the named purposes, summarized for example in [1]. Single metrics can be assigned to classes that relate to dialog- and communication (e.g. *dialog duration*), meta-communication (e.g. *number of help requests*), cooperativity (e.g. *percent of contextual appropriate utterances*), task (e.g. *task success*), or input (e.g. *word error rate*). With the emergence of multimodal systems, this approach of using interaction parameters has been stipulated for this new domain as well (cf. [5]). Several annotation schemes for multimodal interaction have been published, e. g. [6, 7, 8, 9, 10], but researcher still build ‘their own corpora, codification and annotation schemes’ mostly ‘ad hoc’ ([9], p. 121). None of the referred work defines multimodal interaction parameters or annotations for a systematic evaluation of dialog systems, as proposed with PROMISE [11], which can be seen as a multimodal extension of PARADISE. There exists no well-founded expansion to established sets of interaction parameters to assess multimodal HCI.

2. Multimodal Interaction Parameters

Most interaction parameters which have been proposed for spoken dialog systems [1] can be directly transferred to the context

of multimodal dialog systems, for instance, *system response delay* (e.g. the time until the response is displayed on a GUI). For some multimodal systems, *system feedback delay* (e.g. the time until a GUI’s progression bar appears) can be computed additionally. For other parameters, the definition has to be adapted, as exemplified by the parameter *words per turn*, which should be changed to *elements per turn* to accommodate multimodal input and output. In multimodal interaction an element could – for example – be a word, a gesture, a key pressed or an information carrying bit changed in a GUI. Some other parameters, such as speech input-related metrics, have to be mirrored for every input modality.

However, there are also new parameters inherent to multimodal interaction which should be considered. The concept of *modality appropriateness* is discussed in [12]. *Relative multimodal efficiency* and *multimodal synergy* have been defined in [13]. Other parameters have been derived from considerations given in [11]: We propose to parametrize “way of interaction” as *number of turns* for each modality and *lag of time* as a metric for “synchrony”. In addition, the parameters *number of modality changes*, *fusion accuracy*, and *multimodal accuracy* arise from the sheer concept of multimodality. Table 1 offers a short definition of the proposed set of new parameters, which directly refer to the multimodal aspect of the system. The complete list of revised and new parameters can be found in [14].

Not every proposed parameters will be relevant for all possible multimodal systems. Fusion accuracy, for example, will differ from multimodal accuracy only in the case of concurrent or synergistic multimodal input (cf. [15]). And we have considered only multimodal systems relying on directed input and output modalities – modalities, which are used intentionally by the user or perceived consciously.

3. Methodology

As proof of concept an interaction experiment with a multimodal smart-home system has been analyzed based on a subset of the proposed interaction parameters. The system and the experiment are briefly described in the following.

3.1. The multi-modal smart-home system

The smart-home system is set up inside a fully functional living room. Possible interactions include the control of lamps and blinds, the TV, an internet radio, an electronic program guide (EPG), video recorder, and hi-fi system. Furthermore, the system offers an archive for music and allows the generation of playlists. The system can be controlled via a lapel microphone for spoken interaction, and via a smartphone offering a combination of gestural interaction based on accelerometer data and touch interaction via a graphical user interface (GUI). For the

Table 1: Overview of multimodal interaction parameters.

Abbr.	Name	Definition
MA, MER	Multimodal Accuracy, Multimodal Error Rate	Percentage of user inputs (words, gestures, etc.), which have been correctly recognized, based on the hypothesized and the transcribed or coded reference input, averaged over all recognition moduls. $MER = 1 - MA$
SFD	System Feedback Delay	Average delay of system feedback, measured from the end of user input to the beginning of the system feedback in [ms].
# UT_{mod}	Number of User Turns per Modality	Average number of user turns per modality: number of voice inputs, number of gesture inputs, number of multimodal inputs, etc.
# MC	Number of Modality Changes	Overall number of modality changes by the user.
IMA	Input Modality Appropriateness	Overall number or percentage of input modalities chosen which are judged to be appropriate in their immediate dialog context. Determined by labeling user input according to whether they violate one or more of the modality properties defined in [12]: <ul style="list-style-type: none"> • IMA:AP: Appropriate. • IMA:PA: Partially appropriate. • IMA:IA: Inappropriate.
OMA	Output Modality Appropriateness	Overall number or percentage of output modalities chosen which are judged to be appropriate in their immediate dialog context. Determined by labeling system output according to whether they violate one or more of the modality properties defined in [12]: <ul style="list-style-type: none"> • OMA:AP: Appropriate. • OMA:PA: Partially appropriate. • OMA:IA: Inappropriate.
LT	Lag of Time	Overall lag of time between corresponding modalities, in [ms].
FA, FER	Fusion Accuracy, Fusion Error Rate	Percentage of fusion results that are correct. $FER = 1 - MA$. $FA \neq MA$ only if concurrent or synergistic input (cf. [15]).
RME	Relative Multimodal Efficiency	Number of information bits that are communicated correctly using each modality in time unit [13].
MS	Multimodal Synergy	Percent improvement in terms of time-to-task-completion achieved by the multimodal system compared to a system randomly combining modalities [13].

experiment the speech recognizer was replaced by a transcribing wizard, resulting in an additional average delay of 1.4 seconds for spoken input, but a nearly perfect recognition.

3.2. Test design

We asked 24 young adults ($M_{age}=26.1$, $SD_{age}=3.89$) – of which 12 were female – to participate in our experiment. On average, it took the participants 8.53 minutes to complete the task-guided interaction ($SD=2.45$). Some examples of the tasks to be performed with the system are given below, with the possible input modality given in brackets.

- Find out which movies are running tonight (gesture, voice) and record one of them. (touch, voice)
- Play the biathlon video. Mute the sound. (touch, voice)
- Delete two tracks from your ‘favorites’ playlist. Add two new titles. (touch, voice)
- Zap through the radio stations. Turn down the volume. Switch to the next station. (gesture, voice) Mute the sound. (touch, voice)

During the whole experimental session textual and video log data was recorded. For interaction parameters which could not be extracted from log-data, the videos were annotated using ELAN¹. To capture the participants perception of the systems quality, after each interaction they were asked to complete a questionnaire containing 10 semantic pairs rated on a 7-point scale, namely a short version of the AttrakDiff [16].

¹<http://www.lat-mpi.eu/tools/elan/>

4. Results

As stated above, not all multimodal interaction parameters are applicable for the system under consideration. We thus present in this section only results for selected parameters which are meaningful for our system.

4.1. Multimodal accuracy and number of user turns

While the system offers in principle three input modalities, namely spoken input, gestural input and touch input via a GUI presented on the smartphone, depending on the task the user could only select between two modalities: spoken input and gestural input, or spoken input and touch input. Spoken input could be used for every type of interaction. For simple and often repeated interactions, such as turning on the TV, gestural commands were possible. But more complex interaction tasks, for instance the generation of a playlist, could be solved by touch input. In terms of the CARE properties [17], gestural and touch input can therefor be described as assigned input while the combination of both smartphone-based input options are equivalent to spoken input. As the system does neither accept redundant nor complementary input, fusion accuracy is identical with multimodal accuracy.

The recognition accuracy of the three input modalities was 100% for spoken input (as the speech recognizer was replaced by a transcribing wizard) and touch input via the GUI, but only

Table 2: Minimum *min*, maximum *max*, mean *M*, and standard deviation *SD* for number of turns by modality, combined for the smartphone and over-all.

Modality	<i>min</i>	<i>max</i>	<i>M</i>	<i>SD</i>
voice	2	43	16.06	11.4
touch	1	75	30.92	14.47
gesture	0	40	12.58	8.41
smartphone	1	90	43.50	19.91
over-all	40	92	59.54	13.20

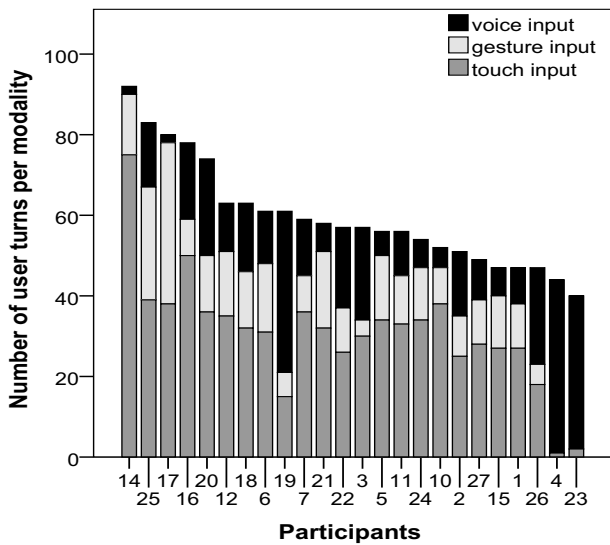


Figure 1: User input choice

42% on average for gesture recognition ($SD=19$). The combination of touch and gesture input (smartphone-based input) achieved an average recognition rate of 85.62% ($SD=6.89$). At the same time the number of turns participants completed with each input modality shows a high variability, as displayed in Table 2. This results in an acceptable average multimodal accuracy (MA) of 88% ($SD=6$).

4.2. System response delay

The system offers immediate feedback when interacting via the smartphone (the GUI reacts on user input with a graphical feedback and each recognized gesture triggers a vibratory feedback). No feedback, apart from the system response, is provided when using spoken input. Consequently, the system feedback delay is constant for each modality. We will analyze system response delay (SRD) instead. The average system response delay is 1.59 seconds with a standard deviation of .53. As stated above, the replacement of the speech recognizer with a wizard led to an additional system response delay for spoken input. This results in a positive correlation of the SRD with the number of user turns via spoken input (UT_{voice}) ($r = .72, p < .01$). Users who used mostly spoken input experienced a much higher SRD than users that preferred the smartphone-based input.

4.3. Number of modality changes

On the smartphone both, gestural and touch interaction both have to be used to solve the given tasks. Therefore, only switching from spoken input to smartphone-based input and vice versa is counted as modality change. Again, a high variability is found with a minimum of 1, a maximum of 35 and a mean of 11.87 ($SD=8.04$) modality changes.

5. Discussion

The parameters gained from the evaluation described above are used to exemplify the three above named applications, namely (1) modeling user ratings, (2) identifying problems appearing during the interaction, and (3) analyzing the interaction to gain insight in human behavior and individual differences.

5.1. Modeling user ratings

We will use the mean of our questionnaire items $\overline{AttDiff}$ as target for a linear regression, resulting in an $r = .72$; $RMSE = .573$. Using normalized parameter values and standardized coefficients, the model looks as follows:

$$\overline{AttDiff} = .314 \cdot UT_{voice} - .301 \cdot PA:IC - .270 \cdot DD + 1.347 \quad (1)$$

In Equation 1 a positive impact of the number of user turns via spoken input, and a negative impact of the number of incorrectly parsed user utterances ($PA:IC$) and the dialog duration (DD) is found. The interaction parameters explain about 50% of the variance in the data. From the list of multimodal parameters, only UT_{voice} shows a significant impact on the perceived quality of the system. The impact of other parameters – such as recognition rates for each modality – dissolves when using UT_{voice} . The relatively good fit of the model is promising. Albeit, it has to be kept in mind that this is the fit for the training data. In addition, the perception of a system’s quality is not only influenced by the course of the interaction between user and system, but also user characteristics and the users perception of hedonic aspects of the system have an influence on their ratings [18, 19].

5.2. Identification of problems

Obvious problems, easily identified by analyzing the interaction parameters, are the low recognition rate of the gesture recognition module and the high system response delay of the spoken input introduced by the wizard. In retrospect, it was found that a programming error in the recognition algorithm hindered an optimal alignment of gesture sequences. That is to say, the algorithm delivered suboptimal results in terms of false, false positive or not recognized gestures. After correction of the recognition algorithm, a short test with ten participants, every participant executing each gesture 10 times, delivered a recognition rate of 85.3%. Although this is a significant improvement on previous results, there is still room for further enhancement. This may be achieved by investigating the use of different algorithms, e.g. Hidden Markov Models (HMMs) as described by Turunen et al. [20] or Neural Networks that possibly provide a better fit in this field of application.

5.3. Analysis of the interaction

The data shows a clear preference of the participants for touch input, as measured by number of user turns (see Table 2). But not every task could be solved via touch. For simple tasks, such

as turning on or off the radio, the TV or the lights, participants had to choose between voice and gestural input. For these tasks gestural input was used 272 times in total and thus clearly more often than voice, which was used 189 times. Only in 27% of the time did participants switch to voice after a gesture recognition error occurred. Therefore, every recognition error led to a second and even third or fourth user turn. Obviously is the high error rate one reason for the high number of gestural input. In the limits of what we could observe in the experiment, the preference for one input modality is not affected by the recognition rates. Based on these parameters, the modality-affected experience with the system apparently had no impact on the modality preferences or usage of the participants.

Only two participants chose to solve nearly all tasks via spoken input, two participants did hardly use spoken input at all, and 15 participants decided to use spoken input for less than half of the tasks (see Figure 1), mainly those, that could not be solved by touch input via the GUI.

Spoken input is still unfamiliar to most users, therefore it is not surprising that they choose the already widely spread touch interaction. There are two possible reasons for the avoidance of spoken input – even for tasks that had to be solved with gestural interaction instead. Firstly, users are known to select the modality, that is more efficient [21, 22]. Due to the additional delay introduced by the typing wizard, system response is much faster when the smartphone was used for interaction. The higher overall efficiency achieved with spoken input due to the higher recognition rate, resulting e.g. in an overall shorter dialog duration and smaller number of user turns, does not counterbalance this effect. Secondly, it has been found before that users avoid spoken input, if they can use a different modality [22].

6. Conclusions

Based on interaction parameters from the field of spoken dialog systems, a set of multimodal interaction parameters is presented in this paper. By exemplifying its usefulness in the multimodal domain, we aim for establishing this extension of common parameters for multiple purposes. One major issue is the transfer of approaches like PARADISE to model user satisfaction for multimodal human-computer interaction. Although multimodal systems are an established topic in the scientific community, a usable approach as proposed with PROMISE has not been presented yet. Other issues addressed include parametrization of interaction for system evaluation and user analysis.

As a next step, correlations between multimodal interaction parameters on the one hand, and ratings and user strategies or characteristics on the other hand have to be investigated and validated. But also a suitable annotation scheme has to be adopted or developed for extracting the multimodal interaction parameters efficiently and comparably. Here, information about the tasks (start, end, success) have to be taken into account to fully use the interaction parameters.

7. Acknowledgments

The project was financially supported by the Deutsche Forschungsgemeinschaft DFG (German Research Community), grant MO 1038/6-1.

8. References

[1] S. Möller, *Quality of telephone-based spoken dialogue systems*. New York, NY, USA: Springer, 2005.

[2] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella, “PARADISE: a framework for evaluating spoken dialogue agents,” in *Proc. of the ACL/EACL 35th Meeting of the Assoc. for Computational Linguistics, Madrid*, 1997, pp. 271–280.

[3] ———, “Evaluating spoken dialogue agents with PARADISE: two case studies,” *Computer Speech and Language*, vol. 12, pp. 317–347, 1998.

[4] M. A. Walker, C. Kamm, and D. J. Litman, “Towards developing general models of usability with PARADISE,” *Natural Language Engineering*, vol. 6, pp. 363–377, 2000.

[5] 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.

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

[7] J.-P. Thiran, F. Marqués, and H. Bourlard, *Multimodal Signal Processing. Theory and applications for human-computer interaction*. Oxford: Academic Press, 2010.

[8] W. Wahlster, *SmartKom: Foundations of Multimodal Dialogue Systems*. Berlin: Springer, 2006.

[9] R. López-Cózar Delgado and M. Araki, *Spoken, Multilingual and Multimodal Dialogue Systems: Development and Assessment*. Chichester: John Wiley & Sons, 2005.

[10] D. Gibbon, I. Mertins, and R. Moore, Eds., *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation*. Norwell, MA, USA: Kluwer, 2000.

[11] N. Beringer, U. Kartal, K. Louka, F. Schiel, and U. Türk., “PROMISE: A procedure for multimodal interactive system evaluation,” in *Proc. of LREC Workshop on Multimodal Resources and Multimodal Systems Evaluation*, 2002, pp. 77–80.

[12] N. O. Bernsen, “From theory to design support tool,” in *Multimodality in Language and Speech Systems*. Dordrecht: Kluwer, 2002, pp. 93–148.

[13] M. Perakakis and A. Potamianos, “Multimodal system evaluation using modality efficiency and synergy metrics,” in *Proceedings of IMCI*, 2008, pp. 9–16.

[14] S. Möller, C. Kühnel, and B. Weiss, “Extending Suppl. 24 to P-Series towards multimodal systems and services,” 2010, source: Deutsche Telekom Laboratories, ITU-T SIG12 Meeting 18–27 May.

[15] L. Nigay and J. Coutaz, “A design space for multimodal systems: concurrent processing and data fusion,” in *Proc. of the INTERACT and CHI*, 1993, pp. 172–178.

[16] M. Hassenzahl and A. Monk, “The inference of perceived usability from beauty,” *Human-Computer Interaction (submitted)*, 2010.

[17] J. Coutaz, L. Nigay, D. Salber, A. Blandford, J. May, and R. Young, “Four easy pieces for assessing the usability of multimodal interaction: The CARE properties,” in *Human-Computer Interaction, Interact ’95*, K. Nordby, P. Helmersen, D. Gilmore, and S. Arnesen, Eds. London: Chapman & Hall, 1995, pp. 115–120.

[18] K. Jokinen and T. Hurtig, “User expectations and real experience on a multimodal interactive system,” in *Proc. of INTERSPEECH*, 2006, pp. 1049–10523.

[19] N. Tractinsky, “Aesthetics and apparent usability: empirically assessing cultural and methodological issues,” in *Proc. of CHI*, 1997, pp. 115–122.

[20] M. Turunen, A. Melto, J. Hella, T. Heimonen, J. Hakulinen, E. Mäkinen, T. Laivo, and H. Soronen, “User expectations and user experience with different modalities in a mobile phone controlled home entertainment system,” in *Proc. of MobileHCI*, 2009, pp. 1–4.

[21] F. Metzke, I. Wechsung, S. Schaffer, J. Seebode, and S. Möller, “Reliable evaluation of multimodal dialogue systems,” in *Proc. of HCII. Part II*, 2009, pp. 75–83.

[22] A. B. Naumann, I. Wechsung, and S. Möller, “Factors influencing modality choice in multimodal applications,” in *Proc. of PIT*, 2008, pp. 37–43.