Describing Multimodal Human-Computer Interaction

Benjamin Weiss
Technische Universität Berlin
Deutsche Telekom Innovation Laboratories
benjamin.weiss@tu-berlin.de

Tatjana Scheffler
Deutsches Forschungszentrum für Künstliche Intelligenz
Projektbüro Berlin
tatjana.scheffler@dfki.de

Sebastian Möller
Technische Universität Berlin
Deutsche Telekom Innovation Laboratories
sebastian.moeller@telekom.de

Norbert Reithinger
Deutsches Forschungszentrum für Künstliche Intelligenz
Projektbüro Berlin
norbert.reithinger@dfki.de

ABSTRACT

In this paper we argue that approaches to annotate multimodal human face-to-face interaction are not suitable for current device-based human-computer interaction. Instead, existing extensions proposed to established parameters describing the interaction with spoken dialog systems are presented and discussed. A standardization activity for annotating user interactions with multimodal systems is needed, e.g. to efficiently extract multimodal interaction parameters useful for evaluation.

Author Keywords

Annotation, Interaction Parameters, Logging.

INTRODUCTION

One of the main approaches to study and evaluate human-computer interaction (HCI) is to record data from individual interactions between user and machine by either using automatic logging functions or by recording audio and video of the interaction as a basis for manual annotation. The data can then be analyzed qualitatively or quantitatively for several purposes.

As an additional step, such data of individual interactions can be aggregated to obtain mainly time-independent parameters, e.g. overall duration, average duration of one interactive action, number of misunderstandings etc. With such a description of individual interactions several purposes are served, for example: (1) finding interaction problems in order to improve system modules to increase cooperativity, effectiveness, or efficiency; (2) analyzing the interaction to gain insight in human behavior and individual differences (e.g. to define user groups or task factors); (3) building models to automatically evaluate interactive systems.

ANNOTATING MULTIMODAL INTERACTION

One major difference between unimodal and multimodal annotation is the need to separate modality input on different layers for each modality [2, 21]. With the NITE project[1], for example, a tool is provided to support annotating audio-video corpora on multiple layers without providing a unified annotation scheme. EMMA[2] offers a coding scheme to be used with automatic recognizer and fusion modules only. Mostly, annotation of multimodal interaction is presented in the domain of human-human interaction (e.g. AMicorpus[3] or data presented at the LREC conferences). For HCI, such corpora are used for example to train recognition modules or build realistic embodied conversational agents.

Annotation schemes for multimodal interaction with computers have been proposed in e.g. [2, 22, 21, 14, 4]. However, researchers build 'their own corpora, codification and annotation schemes' mostly 'ad hoc' [14, p. 121]. This statement seems to be still valid even today.

Another issue is the lack of unified ways to automatically log data. With the 'turn to the wild' [19] researchers have to deal with data from the field, especially from mobile applications, which often provide multimodal interfaces. This increase in mobility results in more and more incomplete data as well as a higher dependency on automatic logging, as laboratory-like audio or video recordings of the users for manual annotation do not exist. We identify three problematic areas when dealing with multimodal HCI:

1. Direct measures of User Experience. With multimodal interfaces, a narrow definition of usability (efficiency, effectiveness, user satisfaction) [7] seems to be inadequate. Current views broaden the focus to user experience [8] which allows to take into account affective aspects of the user, most relevant to distinguish e.g. aesthetics of multimedial system output, stimulation of multimodal user input, or degree of complementation of modalities. As *there exists no comprehensive set of quality aspects for multimodal HCI* [20],

[1] http://www.dfki.de/nite
[2] http://www.w3.org/TR/emma
[3] http://corpus.amiproject.org

systematic research on the relation between such User Experience aspects and descriptions of HCI is currently restricted to those aspects already established like e.g. stimulation, identity, and pragmatic quality [5].

Also, with established subjective methods, the actual benefit of multimodality is not explicitly assessed. But experiencing fusion, fission or sensor-based adaption of a system and interface is one of the main questions in development and research, so explicit user feedback is often desired.

2. Annotation of interaction. Current research directed towards standards for annotating multimodal behavior have been quite successful for human-human communication. For example, standards exist for dialog acts [6], which are multimodal by nature, and a lot of shared experience is available for annotating facial expressions, hand gestures etc. [12]. For this domain of multimodal human-human communication, as well as related HCI (interaction with robots or embodied virtual persons), there seem to be no bigger obstacles for applying existing approaches of aggregating data (e.g. PARADISE, [23]), albeit extended by nonverbal signals [3].

However, such annotation schemes cannot be directly applied to state-of-the-art multimodal interfaces with no embodied system. In human-human communication multimodal (i.e., non-speech) input can often be ignored as redundant or supplementary, and less important, as long it is natural and not the topic of research. In contrast, multimodal interaction is central to many state-of-the-art HCI systems. In HCI, gesture and speech recognition, as well as touch interaction typically provide optimized sets of commands, which are usually system dependent. In order to derive meaningful interaction parameters like modality changes, *single user actions have to be annotated for each modality* providing each piece of information conveyed. Also, user strategies can only be assessed by knowing *when a user switches modalities* to improve effectiveness (i.e. after an error or to cope with environmental changes) or efficiency (e.g. dependent on the amount of information provided).

There is an important separation between non-linguistic and linguistic user information in HCI: On the one hand, systems usually do not understand nonverbal information, but will react with an error instead. On the other hand, hand gestures or voice commands should represent a command known to a system. *How should nonverbal (i.e. sometimes not intentional) user action be annotated? And is it important to also annotate user commands which the system (and even a human annotator) cannot interpret?* Interaction happens in an environment, and multimodal input (e.g. 3D gestures) can only be interpreted considering also this environment. This aspect is still missing.

3. Asymmetry between Human and Computer. For multimodal input, current systems increasingly use sensor data to adapt the interface to the situation (context and environment). In many cases this input data can replace or supplement (spoken) user turns, e.g. as a GPS location can provide the starting point of a bus information query, which otherwise would have to be provided explicitly. But how should such data be handled?

On the output side, *can automatic display adaption be treated as human recipient signals (back-channeling)* or should such system changes be neglected in the first place? On the other hand, *there is an inherent asymmetry between system* and user not evident in human-human communication, as the abilities of both interaction partners are not comparable. Does this difference have to result in separate definitions of annotation concepts for system and user? *Is it, for example, accurate to talk about a system turn* or do we need a different definition of annotating intentional interactive behavior for annotation?

MULTIMODAL INTERACTION PARAMETERS

As starting point for further work, definitions proposed by the ITU [11] are briefly presented here. As a first step, this list of rather abstract information has to be consolidated to meet the demands of many. In a top-down approach, requirements and finally standards how to annotate multimodal HCI could be derived from such a list. Confer also [13] for its application. There is also a related action started at ITU for assessing subjective quality of multimodal services to update the current recommendation for spoken dialog systems [9].

Speech, handwriting and keyboard input employ roughly the same code, namely natural language and can therefore be annotated using the same parameters. At the same time a gesture is defined as every kind of user input executed by parts of the body that is neither handwriting nor keyboard input. This also includes GUI input such as a button press. Modalities are considered as 'directed', if they are used intentionally by the user or perceived consciously. Furthermore, 'undirected' (unconscious or subconscious) information can be obtained from the user: Facial expressions and prosody can be observed to infer the user's mental state and different kinds of sensors can be used to track the user's position. So far, only parameters for directed input and output modalities are specified. Some of the established parameters defined in [10] have been adapted and the existing list was extended.

Adapted Parameters

Among the established parameters for the assessment of spoken dialog systems are time-related parameters such as system and user turn duration [4] and system and user response delay [18]. While the concepts can be transfered to multimodal interaction without modifications, their definition is based on user and system turns, which – in the case of spoken dialog – are equivalent with utterances. It is therefore crucial to define these terms for multimodal interaction, where a turn may take different forms.

It is suggested to measure the duration of the user turn from the beginning of observable user input (in the case of a gesture: the beginning of the observable preparation phase) to the end of a click or the end of the retraction phase of a gesture. Concerning system output a system turn (e. g. the display of an updated GUI) is distinguished from system feedback, such as the display of the loading status of a GUI or vibration feedback indicating the successful receiving of user input (see Figure 1). System feedback is not proposed to be counted as

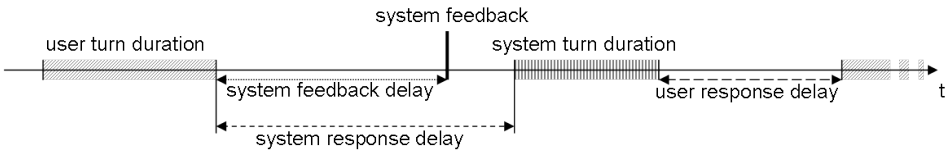

Figure 1. Time-related parameters.

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 modules. $MER = 1 - MA$
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**	Total 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 it violates one or more of the modality properties presented in [2]: • 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 it violates one or more of the modality properties defined in [1]: • 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. [16].
RME	**Relative Multimodal Efficiency**	Number of information bits that are communicated correctly using each modality in time unit [17].
MS	**Multimodal Synergy**	Percent improvement in terms of time-to-task-completion achieved by the multimodal system compared to a system randomly combining modalities [17].

Table 1. Overview of multimodal interaction parameters.

a turn, unless the absence of expected feedback or the occurrence of negative feedback interrupts the user input and / or leads to a repetition. The parameter 'words per turn' is changed to '*elements per turn*' to accommodate multimodal input and output. An element could be a word, a gesture, a key pressed or a piece of information changed in a GUI. Unfortunately, the renaming of words into elements does not mean such values will be comparable between modalities without defining a universal semantic concept to compare the amount of information conveyed.

New Parameters

In parallel to the performance of the speech recognition engine the performance of each recognizer can be quantified. It is also possible to define an average recognition accuracy (*multimodal accuracy*, MA) per user turn by computing the mean of accuracy or error rates of all user inputs.

Multimodal systems offer the possibility to indicate to the user the successful reception of user input or the time needed for the system to process this input. This is considered as positive feedback, which is not counted as a system turn (see above) but which may have a certain delay. This *system feedback delay* should be measured from the end of user input to the start of the system feedback, e.g. from a button click to the display of the loading status of a GUI.

When analyzing multimodal interaction the users' choice of input modality (annotation of *number of user turns for each modality*) and the change of modalities is of interest [15]. In the case of user turns that are truly multimodal (composite or redundant usage of more than one modality) these turns can be annotated as 'multimodal'. For the second point the overall *number of modality changes* by the user and the system are counted.

Depending on the content, the environment, and the user, it can be determined if the offered input and output modalities are appropriate for every given turn (for example guided by modality properties as described by Bernsen [1]). This can be annotated per modality or – in the case of synergistic input and output – for the multimodal input and output as a whole, resulting in *modality appropriateness*. In the first case, each modality can be appropriate or inappropriate. In the second case, the multimodal input and output can be appropriate, partially appropriate or inappropriate.

For multimodal systems the synchrony of related output can be measured by the *lag of time* between corresponding modal-

ities or by the overall number of times different output modalities are asynchronous. Furthermore, the *relative multimodal efficiency* of correctly communicated information of a system and the improvement due to reasonably combining modalities (*multimodal synergy*) can be assessed [17].

Multimodal fusion can be evaluated by comparing the fused result with the result of the individual recognition modules. The total number of fusions computed can be compared to the number of correctly fused concepts (*fusion accuracy*) or to the number of incorrectly fused concepts as a result of ignored and wrongly included recognition results (*fusion error rate*). The performance of the signal level fusion can be measured by comparing the fusion results with the recognition results obtained with each modality separately, see [14] for examples. The parameters are summarized in Table 1. A full list of all parameters is available as ITU-T supplement [11].

SUMMARY AND CONCLUSION

Several issues exist which complicate intuitive annotation and logging of multimodal HCI. We argue for finding an agreement on annotating and logging interactive sessions considering these issues. A basis could be the recommendation of parameters describing multimodal interaction proposed by the ITU following a top-down approach. Separating number and length of turns for each modality to estimate modality changes and proportions of usage and duration for each modality is just one example of the necessity of this work.

REFERENCES

1. Bernsen, N. From theory to design support tool. In *Multimodality in Language and Speech Systems*. Kluwer, Dordrecht, 2002, 93–148.
2. Bernsen, N., and Dybkjær, L. *Multimodal Usability*. Springer, London, 2009.
3. Foster, M., Giuliani, M., and Knoll, A. Comparing objective and subjective measures of usability in a human-robot dialogue system. In *Annual Meeting of the ACL* (2009), 879–887.
4. Gibbon, D., Mertins, I., and Moore, R., Eds. *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation*. Kluwer, Norwell, 2000.
5. Hassenzahl, M. The interplay of beauty, goodness, and usability in interactive products. *Human-Computer Interaction 19* (2008), 319–349.
6. ISO 24617-2. Language resources management —Semantic annotation framework (SemAF) – Part 2: Dialogue acts, 2012. International Organization for Standardization, Geneva.
7. ISO 9241-11. Ergonomic requirements for office work with visual display terminals (VDTs). Part 11: guidance on usability, 1999. International Organization for Standardization, Geneva.
8. ISO DIS 9241-210. Ergonomics of human system interaction—part 210: human-centred design for interactive systems (formerly known as 13407), 2010. International Organization for Standardization, Geneva.
9. ITU Rec. P.851. Subjective quality evaluation of telephone services based on spoken dialogue systems, 2003. International Telecommunication Union, Geneva.
10. ITU Supplement 24 to P-Series Rec. Parameters describing the interaction with spoken dialogue systems, 2005. International Telecommunication Union, Geneva.
11. ITU Supplement 25 to P-Series Rec. Parameters describing the interaction with multimodal dialogue systems, 2011. International Telecommunication Union, Geneva.
12. Kipp, M., Martin, J.-C., Paggio, P., and Heylen, D. *Multimodal Corpora: From Models of Natural Interaction to Systems and Applications*. LNAI 5509. Springer, Berlin, 2009.
13. Kühnel, C. *Quantifying Quality Aspects of Multimodal Interactive Systems*. T-Labs Series in Telecommunication Services. Springer, Berlin, 2012.
14. López-Cózar Delgado, R., and Araki, M. *Spoken, multilingual and multimodal dialogue systems: development and assessment*. John Wiley & Sons, Chichester, 2005.
15. Naumann, A., Wechsung, I., and Hurtienne, J. Multimodal interaction: Intuitive, robust, and preferred? In *Proc. INTERACT* (2009), 93–96.
16. Nigay, L., and Coutaz, J. A design space for multimodal systems: concurrent processing and data fusion. In *Proc. INTERACT & CHI* (1993), 172–178.
17. Perakakis, M., and Potamianos, A. Multimodal system evaluation using modality efficiency and synergy metrics. In *Proc. IMCI* (2008), 9–16.
18. Price, P., Hirschman, L., Shriberg, E., and Wade, E. Subject-based evaluation measures for interactive spoken language systems. In *DARPA Workshop* (1992), 34–39.
19. Rogers, Y. *Quantifying Quality Aspects of Multimodal Interactive Systems*. Synthesis Lectures on Human-Centered Informatics. Morgan & Claypool, 2012.
20. Scapin, D., Senach, B., Trousse, B., and Pallot, M. User experience: Buzzword or new paradigm? In *5th ACHI, Valencia* (2012), 336–341.
21. Steininger, S., Schiel, F., and Rabold, S. Annotation of multimodal data. In *SmartKom: Foundations of Multimodal Dialogue Systems*, W. Wahlster, Ed., Cognitive Technologies. Springer Berlin Heidelberg, 2006, 571–596.
22. Thiran, J.-P., Marqués, F., and Bourlard, H. *Multimodal Signal Processing. Theory and applications for human-computer interaction*. Academic Press, Oxford, 2010.
23. Walker, M., Litman, D., Kamm, C., and Abella, A. Evaluating spoken dialogue agents with PARADISE: two case studies. *Computer Speech and Language 12* (1998), 317–347.