New Challenges in Usability Evaluation - Beyond Task-Oriented Spoken Dialogue Systems

Laila Dybkjær, Niels Ole Bernsen* and Wolfgang Minker***

* Natural Interactive Systems Laboratory
University of Southern Denmark
Campusvej 55, 5230 Odense M, Denmark
laila@nis.sdu.dk, nob@nis.sdu.dk

** Department of Information Technology
University of Ulm
Albert-Einstein-Allee 43, 89081 Ulm, Germany
wolfgang.minker@e-technik.uni-ulm.de

Abstract

There is a fairly good baseline for usability evaluation of task-oriented unimodal spoken dialogue systems (SDSs) but much is still unknown regarding the usability of multimodal and non-task-oriented SDSs. This paper reviews and discusses approaches to usability evaluation of these kinds of SDSs.

1. Introduction

We have eventually achieved a rather strong baseline for evaluating the usability of task-oriented unimodal spoken dialogue systems (SDSs) although some important gaps in our knowledge remain. The knowledge we have comes from national and international projects which have contributed in-the-small via usability evaluation of the systems built in these projects, and, not least, from projects which - based on such individual projects and evaluation contributions – have tried to generalise and propose usability evaluation recommendations. EAGLES [19] and DISC [16] are well-known examples of projects that have collected and built on experience and results from many other projects and proposed guidelines for usability evaluation as well as for technical evaluation of SDSs and their components. The PARADISE framework [33] is also well-known, focusing on a particular metrics for usability evaluation. The framework views user satisfaction as a measure of system usability and seeks to predict user satisfaction by quantitative metrics. See [17] for a review of EAGLES, DISC, PARADISE and several other projects.

However, since research systems for several years have been moving beyond task-oriented unimodal SDSs towards multimodal task-oriented SDSs and towards non-task-oriented conversational SDSs, there is an increasing need for knowledge of how to evaluate the usability of these systems. In many respects this remains an open research issue. We are not starting from scratch, however, since it would seem obvious to draw on methods and criteria from task-oriented unimodal SDS usability evaluation. But we still need to decide – not least for non-task-oriented SDSs – what exactly is transferable and which new evaluation criteria and metrics are required.

This paper discusses current trends and reviews some existing experiences and results in usability evaluation of multimodal as well as non-task-oriented SDSs.

2. Challenges in usability evaluation

Usability evaluation of SDSs is to a large extent based on qualitative and subjective methods and criteria and (mostly) concerns the system as a whole, such as the adequacy of its error handling or the spoken interaction naturalness. As mentioned, gaps remain in our knowledge of usability evaluation of unimodal task-oriented systems. A major gap concerns what usability actually is and what exactly makes a user like a system. We know that there are several contributors to user satisfaction but we hardly know them all nor the extent to which each of them contributes. Moreover, the importance of each criterion may differ across users and user groups.

In addition, we are faced with a number of new usability evaluation issues depending on the type of system we are dealing with. For task-oriented *multimodal* SDSs, a main challenge is to find criteria for evaluating the combinatorial contribution to usability and user satisfaction of the non-speech input and/or output modalities. For *non-task-oriented* unimodal or multimodal SDSs, usability evaluation must be based on the nature of conversation rather than that of shared-goal information-exchange dialogue, which poses new questions as to which of the criteria typically used in evaluating task-oriented SDSs are relevant at all.

Furthermore, the increasing sophistication of SDSs, whichever their modalities and whether task-oriented or not, continues to demand new evaluation metrics. For example, SDSs may be operated in mobile environments and not only in a static environment. There are now research systems which include on-line user modelling to provide more flexible and adaptive dialogue behaviour. Some systems aim to recognise the user's emotional state to provide more appropriate and natural system reactions. User preferences and priorities raise new issues in such systems. Some implications for usability evaluation are outlined in the following.

Speech may be a good choice in mobile environments due to its modality properties of being hands-free and eyes-free, but speech is not very private in public spaces and speech recognisers are sensitive to noise. Thus, the consideration of complementary modalities becomes highly relevant. Mobile SDSs raise several evaluation issues which have not been fully solved, including how (not) to use, and when (not) to use, (very) small screens in combination with speech, see e.g. [31]; for which purposes (not) to use location awareness and situation awareness; and when and for which purposes it is (not) safe to use displays in, e.g., cars [10][28].

On-line user modelling for SDSs is receiving increasing attention for several reasons [5]. Users of mobile devices, which are usually personal belongings, may benefit from functionality which builds knowledge of the individual user. Generic user modelling may also be useful. For instance, novice users could receive more extensive interaction guidance and users who repeatedly make particular types of error could be helped by explicit advice or by adaptation of dialogue structure or initiative distribution. General on-line user modelling is an active research area, see, e.g., [9]. Some key evaluation questions regarding on-line user modelling concern: (i) if the user modelling functionality is feasible and (ii) if it will be of benefit rather than a nuisance to the majority of users of the application. For instance, even if the system has enough information on an individual user, adaptation may fail because of too primitive update algorithms or insufficient information about when the user model has been used.

Not only recognition of users' emotional states but also system expression of emotion is an active research area [1]. For spoken input, the main focus is on prosody [2][23]. Regarding multimodal interaction, research addresses areas, such as the recognition of facial expressions of emotion [11], or speech-cum-facial emotion, as in the ERMIS project (www.image.ntua.gr/ermis/) on emotionally rich interaction systems. Usability evaluation must consider which impacts (positive and negative) emotion modelling has on users.

User preferences can make life hard for the developer as they may contradict what is empirically the most efficient solution. Some users may, e.g., prefer pen-based input to spoken input or keypad-based input to spoken input, simply because they feel more familiar with GUI-style interfaces [25][31]. Depending on the target user group(s), alternative modalities may be needed because it is likely that each of them will be preferred by some users. This is just one reason why user involvement from early on is recommended and why on-line user modelling appears attractive. Some preferences we can design for, such as modality preferences. Others, however, are hard to cope with. Thus, some users may prioritise speed (no queues on the line) or economical benefit (queues but cheap or free calls), while others prioritise human contact. The question is whether we can build systems with a usability profile that will make these users change their priorities, and exactly which usability issues must be resolved to do so.

There is a growing body of results from very different projects which have built and evaluated various aspects of task-oriented multimodal SDSs. Often, the evaluation is done in much the same way as for unimodal SDSs but with additional focus on what the novel modalities might contribute. For non-task-oriented SDSs, there are still few results. Below, we review some approaches to the usability evaluation of such next-generation systems, being aware that this overview is far from complete due to space limitations. Rather, we try to exemplify different trends today.

3. Evaluation of multimodal SDSs

Broadly speaking, we may distinguish between at least the following approaches to usability evaluation of task-oriented multimodal SDSs: (i) Empirical investigations of modality appropriateness, including comparison of SDSs with different modality combinations, and evaluation of user preferences. Focus is on deciding which combination is best suited for a concrete application, user group, environment, etc. (ii) Empirical evaluation of the effects on interaction of animated talking agents. (iii) Theory-based evaluation of SDSs. This is typically done early in the development process and is a relatively cheap method, but it does require an appropriate theory.

3.1. Empirical approaches to modality appropriateness

3.1.1. System comparisons and frameworks

To get an idea of how well different modalities work in combination and of their effect on users, several comparative studies have been made of users interacting with different systems. Often, the three ISO-recommended usability parameters are used in the evaluation, i.e. effectiveness (measured as dialogue success rate), efficiency (measured as time to task completion), and user satisfaction (measured by a questionnaire) [24]. For example, Sturm et al. [30] compared a user-driven and a mixed initiative multimodal SDS on a train timetable information task. Both interfaces offered spoken and pen-based input and display output. The mixed initiative version used speech to guide the dialogue whereas, in the user-driven version – mainly for expert users - the user communicated via tap-and-talk, i.e., the user indicated on the screen which field to fill in next. The effectiveness was found to be approximately the same for the two interfaces whereas the efficiency was higher for the user-driven interface which was also the interface preferred by most users.

Cohen et al. [12] compared the use of a standard GUI interface and an interface with pen and voice input and graphics and voice output. The application was a military task in which units and control measures had to be placed on a map. They showed that the pen/voice SDS interface was faster, also regarding error correction, and strongly preferred by users.

The parameters of efficiency, effectiveness and user satisfaction are basically also those we find at the bottom of the PARADISE framework [33]. In the German SmartKom project, PARADISE has been extended for the purpose of usability evaluation of task-oriented multimodal SDSs. SmartKom allows input speech and gesture and output via speech and screen graphics. SmartKom operates in three environments, i.e. home, mobile, and public. The questionnaire used was adapted to collect information on the different SmartKom scenarios. It includes and extends the usability survey developed in PARADISE. Also, the measurement of dialogue costs, such as dialogue quality, is modified to take into account that the system includes several modalities which may be used in different combinations [3].

3.1.2. User preferences

When speech is the only input/output option, the user is in no doubt about which modality to use, no modality is ignored, and no modality preferences are catered for. The addition of modalities creates the need for usability evaluation of the appropriateness of the offered modalities in relation to application and user group, and of the clarity in presentation to the user of what they can be used for.

den Os et al. [15] conducted an expert evaluation of a speech and pen input, text and speech output directory assistance service running on an iPAQ. The evaluation showed that it must be unambiguous which modalities are available when during interaction, if this may vary. If, e.g., speech has been available at some point, users will expect speech to remain available unless explicitly told that this is no longer the case. It is a design challenge to clearly convey which modalities are available, and when. The authors subsequently made a user test of the same system. The test showed that users have different modality preferences, which affect the way they interact with an application. Several other studies confirm that users have different modality preferences. Sturm et al. [31] analysed the behaviour, preferences, and satisfaction of subjects interacting with an SDS using speech input/output, pointing input and graphics output. Jameson and Klöckner [25] made an experiment showing different modality preferences in a mobile phone task. The task of calling someone while walking around could be carried out using speech and/or keypad input and acoustic and spoken output and/or display.

3.2. Animated talking agents

Animated talking agents (face-only or embodied) have become a popular research area. Usability evaluation of these systems often concern issues such as life-likeness, perceived intelligence, credibility, reliability, efficiency, personality, ease of use, and understanding quality [8][14][22]. The effect of this kind of systems is typically measured either in terms of the user's preferences or via the user's performance. Dehn and van Mulken [14] conclude that, so far, there is no evidence of any general advantage of having an interface with an animated agent over one without. This is supported by [13]. It is also in line with the findings in [8] who evaluated the effects on communication of a real-estate talking agent vs. an over-the-phone version of the same system, in which only the apartments and not the agent could be seen on a screen next to the phone. The perception of efficiency seemed to be gender-dependent, but users generally liked the system better in the speech-only condition. Probably, the lack of natural human behaviour of the agent had a negative effect on users. That this may have an effect is to some degree confirmed by the findings in [22] where controlled experiments were made on the effects of different eye gaze behaviours of a cartoon-like talking face on the quality of human-agent dialogues. The most human-like behaviour led to higher appreciation of the agent and more efficient task performance.

Despite the general conclusion in [14] mentioned above, agents do exist which appear to improve, e.g., intelligibility for users with special problems. Granström and House [20] have used a talking head in several applications, including tourist information, real estate (apartment) search, aid for the hearing impaired, education, and infotainment. Evaluation has shown a significant gain in intelligibility for the hearing impaired when a talking face is added. Eyebrow and head movements enhance perception of emphasis and syllable prominence. Over-articulation may be useful as well when there are special needs for intelligibility. The findings in [26] support these promising conclusions, focusing on applications for the hard-of-hearing, children with autism, and child language learning more generally.

3.3. Theory-based approaches

Usability evaluation is often done by some kind of user testing, cf. the descriptions above. However, the approach of [18] in the Embassi project is a heuristic one. The Embassi system is meant for interaction with home entertainment systems and allows for speech and gesture input and acoustic and graphical output. Heuristic evaluation is motivated as being less time-consuming and expensive than user testing. Based on the modality properties in [4], a set of guidelines is derived and used together with GUI design guidelines [29] to evaluate modality appropriateness.

Given the overwhelming number of modality combinations which could be compared in principle, it may be worthwhile to further explore theory-based approaches. Potentially, much effort could be saved on comparative studies if we can establish a solid set of guidelines based on, e.g., modality theory as suggested by [18].

4. Evaluation of non-task-oriented SDSs

Despite the frequent use of the term 'conversational' by researchers today [32] only few non-task-oriented SDSs have been developed so far and little has been done regarding their usability evaluation. Some of the usability criteria typically used for task-oriented systems become irrelevant, such as sufficiency of task coverage, and probably also efficiency and informativeness. Instead, other issues arise, such as conversation success and naturalness.

The August system [21] allowed users to interact with the Swedish author August Strindberg about various topics via spoken input, and speech and facial output. It was developed in the late 1990s but did not lead to novel usability evaluation metrics. The NICE project [6] develops a non-task-oriented multimodal SDS enabling "real" conversation with life-like fairytale author Hans Christian Andersen via speech and pointing gesture input and speech and graphics output. The usability evaluation criteria proposed in the project include several known from unimodal task-oriented SDSs, but also include criteria for evaluating modalities other than speech, e.g., quality and adequacy of all input and output modalities. However, new challenges are being considered, including metrics for conversation naturalness, such as conversation success, common ground, interlocutor contribution symmetry, topic shift adequacy, educational value, and entertainment value [7].

It is clearly too early to make any firm conclusions regarding usability evaluation of non-task-oriented SDSs but, surely, novel and, in some cases re-defined, metrics will be needed as suggested by NICE.

Cole et al. [13] present ongoing work on tutoring systems and envision that it will be possible in the near future to build life-like characters that interact with people much like people interact with each other. Spoken dialogue technology must be combined with computer vision and animated agent technology to achieve this goal. An important evaluation criterion of such tutoring systems will be if there is any learning benefit.

5. Conclusion

We have briefly addressed the current usability evaluation baseline for unimodal task-oriented SDSs. We have discussed some remaining gaps in our knowledge of usability evaluation and some new challenges ahead caused by the increasing sophistication of SDSs as well as by research moving into multimodal and to non-task-oriented SDSs. We have reviewed approaches to usability evaluation in several finished and ongoing projects on multimodal task-oriented and non-task-oriented SDSs.

There seems to be a broad need for usability evaluation that can help us find out how users perceive these new kinds of SDSs and how well users perform with them, possibly compared to other types of system. There is a strong wish to find ways in which usability and user satisfaction might be correlated with technical aspects in order for the former to be derived from the latter. We don't have methods today that enable prediction of how well users will receive a system. We just know that a technically optimal system is not enough to produce user satisfaction. Regarding modality appropriateness which is a central issue in multimodal SDSs, modality theory may be a promising and powerful approach to usability evaluation of modalities at an early stage. However, user tests of the actual design will still be needed, as for unimodal SDSs.

6. References

[1] André, E., Dybkjær, L., Minker, W. and Heisterkamp, P. (Eds.): Affective Dialogue Systems. LNAI 3068, Springer 2004.

[2] Batliner, A., Fischer, K., Huber, R., Spilker, J. and Nöth, E.: Desperately Seeking Emotions: Actors, Wizards, and Human Beings. Proc. of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research, Belfast, 2000, 195-200.

[3] Beringer, N., Kartal, U., Louka, K., Schiel, F. and Türk, U.: PROMISE - a procedure for multimodal interactive system evaluation. Proc. of the LREC Workshop on Multimodal Resources and Multimodal Systems Evaluation, Las Palmas, 2002, 77-80.

[4] Bernsen, N.O.: Multimodality in Language and Speech Systems - from Theory to Design Support Tool. In Granström, B., House, D., and Karlsson, I. (Eds.): Multimodality in Language and Speech Systems, Kluwer Academic Publishers, Dordrecht, 2002, 93-148.

[5] Bernsen, N.O.: User Modelling in the Car. In [9], 2003, 378-382.

[6] Bernsen, N.O., Charfuelàn, M., Corradini, A., Dybkjær, L., Hansen, T., Kiilerich, S., Kolodnytsky, M., Kupkin, D. and Mehta, M.: First Prototype of Conversational H. C. Andersen. Proc. of the International Working Conference on Advanced Visual Interfaces (AVI 2004), Gallipoli, Italy, 2004, 458-461.

[7] Bernsen, N.O., Dybkjær, L. and Kiilerich, S.: Evaluating Conversation with Hans Christian Andersen. Proc. of LREC, Lisbon, Portugal, 2004, 1011-1014.

[8] Bickmore, T. and Cassell, J.: Social Dialogue with Embodied Conversational Agents. In [32], 2004.

[9] Brusilovsky, P., Corbett, A. and de Rosis, F. (Eds.): User Modeling 2003. Proc. of the 9th International Conference, UM 2003, Johnstown, PA, USA, Springer Lecture Notes in Artificial Intelligence, Vol. 2702, 2003.

[10] Bühler, D., Minker, W., Häussler, J., and Krüger, S.: Flexible Multimodal Human-Machine Interaction in Mobile Environments. Proc. of the ECAI Workshop on Artificial Intelligence in Mobile System (AIMS), Lyon, 2002, 66-70.

[11] Cohen, I., Sebe, N., Chen, L., Garg, A. and Huang, T.: Facial Expression Recognition from Video Sequences: Temporal and Static Modeling. In Computer Vision and Image Understanding, Special Issue on Face Recognition, Vol. 91, Issues 1-2, 2003, 160-187.

[12] Cohen, P., McGee, D., and Clow, J.: The Efficiency of Multimodal Interaction for a Map-based Task. Proc. of the Applied Natural Language Processing Conference, Morgan Kaufmann, 2000, 331-338.

[13] Cole, R., van Vuuren, S., Pellom, B., Hacioglu, K., Ma, J., Movellan, J., Schwartz, S., Wade-Stein, D., Ward, W. and Yan, J.: Perceptive Animated Interfaces: First Steps towards a New Paradigm for Human-Computer Interaction. In [32], 2004.

[14] Dehn, D. and van Mulken, S.: The Impact of Animated Interface Agents: A Review of Empirical Research. Int. Journal of Human-Computer Studies 52, 2000, 1-22.

[15] den Os, E., de Koning, N., Jongebloed, H. and Boves. L.: Usability of a Speech-Centric Multimodal Directory Assistance Service. Proc. of the CLASS Workshop on Information Presentation and Natural Multimodal Dialogs, Verona, Italy, 2001, 65-69.

[16] Dybkjær, L. and Bernsen, N.O.: Usability Issues in Spoken Language Dialogue Systems. Natural Language Engineering, Special Issue on Best Practice in Spoken Language Dialogue System Engineering, Vol. 6 Parts 3 & 4, 2000, 243-272.

[17] Dybkjær, L., Bernsen, N.O. and Minker, W.: Evaluation and Usability of Multimodal Spoken Language Dialogue. Speech Communication, Elsevier, Amsterdam, 2004.

[18] Elting, C, Strube, S, Möhler, G, Rapp, S. and Williams, J: The Use of Multimodality within the EMBASSI system. Proc. of M&C2002, Usability Engineering Multimodaler Interaktionsformen, Hamburg 2002.

[19] Gibbon, D., Moore, R. and Winski, R. (Eds.): Handbook of Standards and Resources for Spoken Language Systems. Mouton de Gruyter, Berlin, New York, 1997.

[20] Granström, B. and House, D.: Effective Interaction with Talking Animated Agents in Dialogue Systems. In [32], 2004.

[21] Gustafson, J., Lindberg, N. and Lundeberg, M.: The August Spoken Dialogue System. Proc. of Eurospeech, 1999, 1151-1154.

[22] Heylen, D., van Es, I., Nijholt, A. and van Dijk, B.: Controlling the Gaze of Conversational Agents. In [32], 2004.

[23] Hirschberg, J., Swerts, M. and Litman, D.: Labeling Corrections and Aware Sites in Spoken Dialogue Systems. Proc. of the 2nd SIGdial Workshop on Discourse and Dialogue, Aalborg, Denmark, 2001, 72-79.

[24] ISO (International Standardisation Organisation): ISO 9241: Ergonomic requirements for office work with visual display terminals (VDTs), Part 11: Guidance on usability. http://www.iso.org

[25] Jameson, A. and Klöckner, K.: User Multitasking with Mobile Multimodal Systems. In [27], 2004.

[26] Massaro, D. W.: The Psychology and Technology of Talking Heads: Applications in Language Learning. In [32], 2004.

[27] Minker, W., Bühler, D. and Dybkjær, L. (Eds.): Spoken Multimodal Human-Computer Dialogue in Mobile Environments. Kluwer Academic Publishers, to appear, 2004.

[28] Minker, W., Haiber, U., Heisterkamp P. and Scheible, S.: The Seneca Spoken Language Dialogue System. Speech Communication, Elsevier, Amsterdam, 2004.

[29] Nielsen, J.: Heuristic Evaluation. In Nielsen, J. and Mack, R.L. (Eds.): Usability Inspection Methods. John Wiley & Sons, New York, 1994.

[30] Sturm, J., Bakx, I., Cranen, B. and Terken, J.; Comparing the Usability of a User Driven and a Mixed Initiative Multimodal Dialogue System for Train Timetable Information. Proc. of Eurospeech, 2003, 2245-2248.

[31] Sturm, J., Cranen, B., Wang, F., Terken, J. and Bakx, I: Effects of Prolonged Use on the Usability of a Multimodal Form-filling Interface. In [27], 2004.

[32] van Kuppevelt, J., Dybkjær, L. and Bernsen, N.O. (Eds.): Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems. Kluwer Academic Publishers, to appear, 2004.

[33] Walker, M., Litman, D., Kamm, C. and Abella, A.: PARADISE: A Framework for Evaluating Spoken Dialogue Agents. Proc. of the Association of Computational Linguistics (ACL), 1997, 271-280.