Usability Evaluation of Multimodal and Domain-Oriented Spoken Language 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

Considerable work has been done regarding usability evaluation of task-oriented unimodal spoken language dialogue systems (SLDSs). However, there are still important gaps in our knowledge even in this area. If we move to multimodal task-oriented SLDSs, there are more challenges ahead primarily due to the combination of different modalities. For non-task-oriented conversational SLDSs, a major challenge is the new evaluation issues brought up by the nature of conversation. This paper presents a state-of-the-art in usability evaluation of these new types of SLDSs and conclusions based on the experience we have today.

1. Introduction

Various metrics and strategies have been proposed for evaluating the usability of spoken language dialogue systems (SLDSs). By contrast with technical evaluation, usability evaluation of SLDSs is to a large extent based on qualitative and subjective methods and (mostly) concerns the system as a whole. Ideally, we would prefer quantitative and objective usability evaluation scores which, furthermore, can be compared to scores obtained from evaluation of other SLDSs. However, many important usability issues seem unlikely to be subjected to objective quantification in the foreseeable future, and expert evaluation is sometimes highly uncertain or unavailable. Nevertheless, although important problems remain, there actually exists a rather strong baseline for evaluating the usability of task-oriented unimodal SLDSs, cf. Section 2.

The picture is quite different when we consider multimodal task-oriented SLDSs and non-task-oriented conversational SLDSs, whether unimodal or multimodal. We call the latter domain-oriented systems since these systems have to work without the powerful constraints provided by the task. Multimodal SLDSs which allow the use of other modalities in addition to speech are proliferating in current research but not yet in commercial applications. How to evaluate their usability remains an open research issue in many respects. We are not clueless in addressing this issue, however, since it would seem obvious, for a start, to draw on methods and criteria from SLDS usability evaluation. The issue then becomes one of deciding what is (not) transferable and which new evaluation criteria and metrics are required. An analogous situation may obtain as regards usability evaluation of (unimodal and multimodal) domain-oriented SLDSs, except for the intriguing fact that conversation is very different from task-oriented dialogue.

This paper presents what we view as the current baseline in SLDS usability evaluation, including some unsolved problems (Section 2). Section 3 reviews some existing experiences and results in multimodal SLDSs evaluation. Section 4 concludes on today's challenges in developing usable multimodal and domain-oriented SLDSs.

2. Evaluation baseline

A usable SLDS must not only offer appropriate functionality but must also satisfy user needs which go beyond functionality needs, and it must be easy to understand and interact with, not least when it comes to walk-up-and-use systems. As a rule, usability should be factored in from the very beginning of the SLDS development process. It is therefore recommended to have close interaction with representative users throughout the development process although this does not in itself guarantee good usability. To build usable SLDSs we need knowledge about issues, such as users' linguistic, para-linguistic and non-linguistic behaviour, their comprehension of a given system's behaviour, user reactions to SLDSs in the field, and the main factors which determine overall user satisfaction.

Today's baseline for SLDSs usability evaluation may be viewed as a range of evaluation methods and approaches and a series of more specific evaluation criteria. Methods and approaches support usability evaluation from early on and throughout the development process. Design analysis using, e.g., mock-ups and walkthroughs, can be done early in the process to evaluate dialogue model sketches. Wizard of Oz data analysis may also help evaluate the dialogue model prior to implementation, while controlled user tests and field studies require a largely implemented SLDS. Following methods like these, usability information is collected by means of, e.g. logfiles, transcriptions, observations, notes, interviews and questionnaires.

Many projects have carried out usability evaluations of task-oriented SLDSs on a small scale focusing on different evaluation criteria, see [Dybkjær et al. 2004] for an overview. A few projects have gone a step further and proposed recommendations and guidelines for usability evaluation by collecting and building on experience and results from many other projects. These include the EAGLES and DISC projects and the PARADISE framework.

2.1. The EAGLES project

The EAGLES (Expert Advisory Group on Language Engineering Standards) project (1993-1998, http://lingue.-ilc.pi.cnr.it/EAGLES96/home.html) included several working groups one of which addressing evaluation recommendations, cf. [Gibbon et al. 1997]. A drawback for this group was that only a few SLDSs had been evaluated by the mid-1990s when EAGLES collected their information. Focus was on both technical and usability evaluation in terms of glass-box and black-box evaluation. The proposed black-box evaluation includes the following quantitative and qualitative measures, many of which are of clear relevance to usability although some of them may be (too) cumbersome to follow: average number of exchanges to obtain relevant responses, task completion rate, transaction success rate, system response time, terseness of the system's answers, user satisfaction, ability to adapt to new users, ability to adapt to the same user, and ability to handle multimodality. Black-box evaluation is also recommended for comparative system evaluation. The proposed key comparative evaluation measures include dialogue duration, turn duration, contextual appropriateness, correction rate, and transaction success rate.

2.2. The DISC project

The EU DISC project (1997-1999, www.disc2.dk) analysed the effort made in a wide range of unimodal and multimodal projects, producing an overview of current development and evaluation practice and proposing a best practice guide for the development and evaluation of SLDSs. DISC had the advantage of being able to draw on more finished SLDS projects than EAGLES. Moreover, DISC went a step further and took a comprehensive look at development and evaluation best practice for all major SLDSs components as well. SLDSs usability evaluation was among the issues investigated by DISC. The results are reported in [Dybkjær and Bernsen 2000]. They propose a set of 15 quantitative, qualitative, and subjective usability evaluation criteria, i.e.: modality appropriateness; input recognition adequacy; coverage of user vocabulary and grammar; output voice quality; output phrasing adequacy; feedback adequacy; adequacy of dialogue initiative relative to the task(s); naturalness of the dialogue structure relative to the task(s); sufficiency of task and domain coverage; sufficiency of the system's reasoning capabilities; sufficiency of interaction guidance (information about system capabilities, limitations and operations); error handling adequacy; sufficiency of adaptation to user differences; number of interaction problems; and user satisfaction. Some of these criteria are based on existing theory, such as modality theory for modality appropriateness [Bernsen 2002] and co-operativity principles for output phrasing and interaction problems [Bernsen et al. 1998, Grice 1975], while most of them are empirical and judgmental.

The list is not complete – e.g., politeness and cultural differences are not included – but it probably covers a good deal of usability basics for task-oriented SLDSs. An important problem is that we still know too little about how each of the listed system properties affect general system usability.

2.3. The PARADISE framework

Most of the criteria listed above require qualitative or even subjective evaluation. Subjective evaluation is less reliable than objective evaluation. Thus, it is not surprising that several attempts have been made to avoid subjectivity in the evaluation of user satisfaction and use quantitative metrics instead, such as elapsed time, number of turns, and number of repairs, each of which in some way may contribute to user satisfaction.

The most well-known attempt to measure user satisfaction by quantitative metrics is probably the PARADISE framework [Walker et al. 1997] which has been used for the evaluation of several SLDSs, e.g., the DARPA Communicator systems [Walker et al. 2002]. The PARADISE framework views user satisfaction as a measure of system usability and assumes that the primary objective of an SLDS is to maximise user satisfaction. Task success and various dialogue costs relating to efficiency and quality contribute to user satisfaction. To maximize user satisfaction, one must maximise task success and minimise dialogue costs. Modelling user satisfaction as a function of task success and dialogue cost is intended to lead to a predictive performance model of SLDSs, enabling prediction of user satisfaction based on measurable parameters which can be found in log-files. Thus, the idea is that, eventually, costly and hard-to-interpret subjective user evaluation can be avoided. For the moment – since the predictive performance model does not yet exist - users are asked questions on various aspects of their interaction with the system and have to rate the aspects on a five-point multiple choice scale. The response values are summed, resulting in a user satisfaction measure for each dialogue.

Currently, there is no better proposal of its kind. However, the framework has several weaknesses: (i) In real life, user satisfaction is not identical to usability. (ii) It is not known exactly which are the criteria that contribute to user satisfaction and by which weight they do so. (iii) As long as there is no predictive model, the framework can only be used in controlled user tests since users must fill in a questionnaire. Moreover, we still know too little about user uptake of commercial applications. Compared with test users, real users may react quite differently to an application. If this is the case, the predictive model to be generated may be wrong. (iv) A fourth unsolved problem relates to the questionnaire which is probably the most commonly used tool for gathering users' opinions. Questionnaire results are interpreted as expressions of how satisfied users are with the system. In common practice, however, questionnaires are not validated before being used, to verify that the "right questions" are being asked nor is it verified that the results obtained are representative [Larsen 2003]. (v) Questionnaire interpretation is subjective, adding an additional layer of uncertainty.

EAGLES, DISC and PARADISE illustrate that substantial work has already been done regarding usability evaluation of task-oriented unimodal SLDSs. Important gaps in our knowledge remain to be filled, however, such as which parameters actually contribute to usability and to which extent.

3. Towards usability evaluation of multimodal and domain-oriented systems

When we turn to domain-oriented conversational SLDSs or to multimodal SLDSs we can re-use (part of) what is known about usability evaluation of task-oriented unimodal SLDSs. We are faced, however, with a number of new issues depending on the type of system we are dealing with. For task-oriented multimodal SLDSs, a main challenge is criteria for evaluating the combinatorial contribution to usability and user satisfaction of the non-speech input and/or output modalities. For domain-oriented unimodal or multimodal SLDSs, usability evaluation must be based on the nature of conversation rather than that of information-seeking dialogue, which poses new requirements as to which familiar criteria are relevant at all. In the following, we review experience and results from usability evaluation of a number of multimodal and domain-oriented SLDSs. Due to space the overview of multimodal SLDSs is quite limited and only gives examples. But we hope it gives an idea of where we are today.

3.1. Evaluation of multimodal SLDSs

When speech is the only input/output option, there is no doubt for the user about which modality to use, no modality is ignored, and no modality preferences are catered for. With the addition of modalities, this situation changes, raising 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. In multimodal SLDS projects, modality evaluation is often a focal point and has been done from different perspectives.

[den Os et al. 2001] 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. Some users prefer pen-based input to spoken input simply because they feel more familiar with GUI-style interfaces as confirmed by [Sturm et al. 2002] who analysed the behaviour and satisfaction of subjects interacting with an SLDS offering speech input/output, pointing input and graphics output. Depending on the target user group(s), alternative modalities may have to be enabled due to different user preferences. This is just one reason why user involvement from early on is recommendable and why on-line user modelling may be attractive.

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 two 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) [ISO]. Thus, [Sturm et al. 2003] compared a user-driven and a mixed initiative multimodal SLDS 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 – which is 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. 2000] 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 SLDS interface was faster – even regarding error correction - and strongly preferred by users.

[Heylen et al. 2002] made controlled experiments 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.

[Bickmore and Cassell 2004] evaluated the effects on communication of an embodied conversational real-estate agent vs. an over-the-phone version of the same system where 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.

The parameters of efficiency, effectiveness and user satisfaction are basically also those we find at the bottom of the PARADISE framework. In the German SmartKom project, PARADISE has been extended for the purpose of usability evaluation of task-oriented multimodal SLDSs. 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 [Beringer et al. 2002]. However, apart from catering for more modalities, the adapted framework would seem to suffer from the same weaknesses as PARADISE.

Usability evaluation is often done by some kind of user testing, cf. the descriptions above. However, the approach of [Elting et al. 2002] 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 [Bernsen 2002], they derive a set of guidelines which are used together with GUI design guidelines [Nielsen 1994] to evaluate modality appropriateness. There is an overwhelming number of modality combinations which could be compared. Maybe, 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 [Elting et al. 2002]. This would seem to be a powerful approach to usability evaluation of modalities at an early stage. User tests of the actual design will still be needed, just as for unimodal SLDSs.

3.2. Evaluation of domain-oriented SLDSs

Few domain-oriented SLDSs have been developed and little has been done so far regarding their usability evaluation. Some of the usability criteria mentioned in Section 2 are clearly irrelevant, such as sufficiency of task coverage, and probably also efficiency and informativeness. Instead, other issues arise, such as conversational naturalness.

The August system [Gustafson et al. 1999] which allowed users to interact via speech input, and speech and graphics output with the Swedish author August Strindberg about various topics, was developed in the late 1990s but did not lead to novel usability evaluation metrics.

The NICE project [Bernsen et al. 2004a] develops a domain-oriented multimodal SLDS enabling interaction 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 SLDSs (Section 2.2), but extended to include modalities other than speech, e.g., quality and adequacy of all input and output modalities. However, new challenges are being considered, including metrics for conversational adequacy and naturalness, such as common ground, interlocutor contribution symmetry, and topic shift adequacy; educational value; entertainment value; and a novel notion of transaction (or concept) success [Bernsen et al. 2004b].

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

4. Conclusion

We have presented a baseline for usability evaluation of task-oriented unimodal SLDS and reviewed the approach to usability evaluation in several finished and ongoing projects on multimodal task-oriented and domain-oriented SLDSs. There is a growing body of results from very different projects which have built and evaluated various aspects of next-generation task-oriented multimodal SLDSs. Often, the evaluation is done in much the same way as for unimodal SLDSs but with additional focus on modalities. It may be worthwhile to try to establish a set of theory-based guidelines which can save the effort of comparative studies of modality combinations. Concerning domain-oriented SLDSs, usability evaluation is in its infancy. There are proposals for usability metrics but no clear trend yet.

References

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

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.

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. *Proceedings of the International Working Conference on Advanced Visual Interfaces* (AVI 2004), Gallipoli, Italy, 2004a (to appear).

Bernsen, N.O., Dybkjær, H. and Dybkjær, L.: *Designing Interactive Speech Systems. From First Ideas to User Testing.* Springer Verlag, London, 1998.

Bernsen, N.O., Dybkjær, L. and Kiilerich, S.: Evaluating Conversation with Hans Christian Andersen. Proceedings of LREC, Lisbon, Portugal, 2004b (to appear).

Bickmore, T. and Cassell, J.: Social Dialogue with Embodied Conversational Agents. van Kuppevelt, J., Dybkjær, L. and Bernsen, N.O. (Eds.): *Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems.* Kluwer Academic Publishers, 2004 (to appear).

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

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

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

Dybkjær, L., Bernsen, N.O. and Minker, W.: Evaluation and Usability of Multimodal Spoken Language Dialogue. Accepted for publication by *Speech Communication*, Elsevier, Amsterdam, 2004.

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

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

Grice, P.: Logic and conversation. In P. Cole and J. L. Morgan (Eds.), *Syntax and Semantics* Vol. 3: *Speech Acts*. New York: Academic Press 1975, 41-58.

Gustafson, J., Lindberg, N., and Lundeberg, M.: The August Spoken Dialogue System. *Proceedings of Eurospeech*, 1999, 1151-1154.

Heylen, D., van Es, I., Nijholt, A. and van Dijk, B.: Experimenting with the Gaze of a Conversational Agent. *Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems.* Copenhagen, 2002, 93-100.

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

Larsen, L. B.: Assessment of Spoken Dialogue System Usability – What are We really Measuring? *Proceedings of Eurospeech* 2003, 1945-1948.

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

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. *Proceedings of Eurospeech,* 2003, 2245-2248.

Sturm, J., Cranen, B., Wang, F., Terken, J., and Bakx, I: The Effect of Prolonged Use on Multimodal Interaction. *Proceedings of the ISCA Workshop on Multi-Modal Spoken Dialogue in Mobile Environments,* Bonn: ESCA, 2002.

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

Walker, M., Rudnicky, A., Prasad, R., Aberdeen, J., Bratt, E., Garofolo, J., Hastie, H., Le, A., Pellom, B., Potamianos, A., Passonneau, R., Roukos, S., Sanders, G., Seneff, S., and Stallard, D.: DARPA Communicator: Cross-system results for the 2001 evaluation. *Proceedings of the International Conference of Spoken Language Processing* (ICSLP 2002), 2002, 269-272.