Chapter 13

OVERVIEW OF EVALUATION AND USABILITY*

Laila Dybkjær, Niels Ole Bernsen
Natural Interactive Systems Laboratory, Odense, Denmark
{laila,nob}@nis.sdu.dk

Wolfgang Minker
University of Ulm, Department of Information Technology, Ulm/Donau, Germany
wolfgang.minker@e-technik.uni-ulm.de

Abstract With the technical advances and market growth in the field, the issues of evaluation and usability of spoken dialogue systems, unimodal as well as multimodal, are as crucial as ever. This chapter discusses those issues by reviewing a series of European and US initiatives which have produced major results on evaluation and usability. Whereas significant progress has been made on unimodal spoken dialogue systems evaluation and usability, the emergence of, among others, multimodal, mobile, and non-task-oriented systems continues to pose entirely new challenges to research in evaluation and usability.

Keywords: Multimodal systems; Spoken dialogue systems; Evaluation; Usability.

1. Introduction

Spoken dialogue systems (SDSs) are proliferating in the market for a large variety of applications and in an increasing number of languages. As a major step forward, commercial SDSs have matured from technology-driven prototypes to business solutions. This means that systems can be copied, ported, localised, maintained, and modified to fit a range of customer and end-user needs

*This chapter is a modified version of the article entitled "Evaluation and Usability of Multimodal Spoken Language Dialogue Systems" published in Speech Communication, Vol. 43/1-2, pp. 33–54, Copyright (2004), reprinted with permission from Elsevier.

without fundamental innovation. This is what contributes to creating an emerging industry. At the same time, increasingly advanced SDSs are entering the market, drawing on experience from even more sophisticated research systems and continuous improvements in SDS technologies. Furthermore, in many research laboratories, focus is now on combining speech with other modalities, such as pen-based hand-writing and 2D gesture input, and graphics output, such as images, maps, lip movements, animated agents, or text (Wahlster et al., 2001; Bickmore and Cassell, 2002; Oviatt, 1997; Gustafson et al., 2000; Sturm et al., 2004; Oviatt et al., 2004; Whittaker and Walker, 2004). An additional dimension which influences development is the widening context of use. Mobile devices, in particular, such as mobile phones, in-car devices, PDAs and other small handheld computers open up a range of new application opportunities for unimodal as well as multimodal SDSs as witnessed by several chapters in this book. In this continually expanding field of unimodal and multimodal, mobile and non-mobile SDSs, many research issues still remain to be solved. Two issues of critical importance are evaluation and usability. Systems *evaluation* is crucial to ensure, e.g., system correctness, appropriateness, and adequacy, while *usability* is crucial to user acceptance.

Many results are available from individual development projects regarding evaluation and usability. These issues often receive some amount of attention in SDS projects although only few have their main focus on any of them. By themselves, isolated results and experience are usually neither easily generalisable nor immediately transferable to other projects. However, the results are still important. Generalisations, best practice guidelines, and, eventually, (de facto) standards are typically based on empirical evidence from many different sources and are applicable across projects within their scope. Other important approaches to evaluation and usability that are valid across projects are frameworks and theoretical approaches. A framework may be described as a general toolset with a well-defined scope which reflects some kind of principled approach. A theoretical approach is based on deeper insight into relationships among key concepts or variables. Below, we shall use the distinction between empirical generalisations, frameworks, and theory for the purpose of exposition even though it remains true that theories and frameworks are worthless without empirical evidence and empirical generalisation tends to be couched in theoretically inspired, or even derived, concepts.

This chapter surveys what we have learned and where we are today regarding SDS evaluation and usability. Section 2 presents a brief overview of the state-of-the-art in evaluation and usability. We then discuss empirical generalisations (Section 3), frameworks (Section 4), and theory and generalisations on the usability of multimodal SDSs (Section 5). Section 6 concludes the chapter. Given the fact that, in particular, the state of the art in spoken multimodal and mobile systems usability and evaluation remains uncharted to a large ex-

tent, the perspective adopted is necessarily a partial one. Moreover, the reader should be aware that this chapter focuses on spoken input and output while other modalities are only considered to the extent that they are being used together with speech.

2. State-of-the-Art

The first simple, commercial SDS appeared in 1989 (Bossemeyer and Schwab, 1991) based on many years of research, particularly in speech recognition. Increasingly complex and sophisticated technologies were introduced during the 1990s, bringing issues such as barge-in, large-vocabulary recognition in noisy conditions, robust parsing, flexible dialogue management, language generation, and easy portability to the forefront of research. The recent advent of multimodal and of mobile SDSs has compounded the challenges to establishing best practice for the development and evaluation of usable SDSs and their component technologies.

In the USA and Europe, several initiatives have addressed SDSs evaluation and usability since the late 1980s. Several of these are mentioned in (Dybkjær et al., 2004). While the focus in Europe has been on analysing various aspects of evaluation and usability, focus in the USA has been on competitive evaluation among projects addressing the same task(s). Together these initiatives have managed to, at least:

- establish various basic metrics, see examples below;
- vastly increase our knowledge of issues, such as test corpus creation, comparative component and system evaluation, portability to different languages, the need for robust parsing, and, more generally, best practice in the development and evaluation of SDSs and their components;
- introduce a range of new, difficult evaluation topics, such as, how to design informative user questionnaires, when and how to use animated interface agents, how to evaluate education value and entertainment value, how to generalise evaluation results on spoken multimodal systems, how to identify key factors influencing customer satisfaction, and when to use speech interfaces.

Today's SDSs are largely task-oriented but novel, non-task-oriented systems are emerging. Technical sophistication differs dramatically among unimodal as well as multimodal SDSs, which means that the same set of evaluation criteria cannot be applied to all. Rather, some subset of a broader set of evaluation criteria will be relevant to each particular system. As regards usability, system variability includes, e.g., the fact that the skills and preferences of the target users may differ widely. This and other parameters must be taken into account

when designing for, and evaluating, usability no matter the technical sophistication of the system.

Broadly, evaluation may be decomposed into (i) technical (including functional) evaluation of systems and their components, (ii) usability evaluation of systems, and (iii) customer evaluation of systems and components. Although (i)-(iii) are interrelated, a technically excellent system may have poor usability whilst a technically inferior system may score highly in user satisfaction questionnaires. Moreover, the customer may prefer yet another system for reasons of, say, cost and platform compatibility which have little to do with technical perfection or end-user satisfaction. Unfortunately, too little is known at present about the important topic of customer evaluation. In the following, we focus on technical evaluation (Section 2.1) and on usability and usability evaluation (Section 2.2).

2.1 Technical Evaluation

Technical evaluation concerns the entire SDS as well as each of its components. In this overview chapter we cannot give a full and detailed account of evaluation criteria for all the individual components of an SDS. This section describes some of the most important evaluation criteria for major SDS components.

Technical evaluation is usually done by developers as objective evaluation, i.e. quantitative and/or qualitative evaluation. Quantitative evaluation consists in measuring something and producing an independently meaningful number, percentage etc. Qualitative evaluation consists in estimating or judging some property by reference to expert standards and rules.

Technical evaluation is well developed for many aspects of SDSs and their components. As a minimum, as many bugs as possible should be found and repaired through diagnostic evaluation. Proper technical evaluation also includes measuring, through performance evaluation, whether the system's or component's functionality is as specified. Finally, technical evaluation may also be done in order to make comparisons with other SDSs.

There is widespread agreement on key evaluation criteria for speech recognisers and speech synthesisers. For speech recognisers, these criteria include word and sentence error rate, vocabulary coverage, perplexity, and real-time performance. Word and sentence error rate are measured by comparing the transcribed input with the recogniser's output. Other examples are metrics for speaker identification and speaker separation. For speech synthesisers, user perception continues to have a central role in evaluation. Some of the basic properties which should be evaluated are speech intelligibility, pleasantness and naturalness (Karlsson, 1999). For natural language understanding, some basic metrics are lexical coverage, grammar coverage, and real-time per-

formance. The particular grammatical properties of spoken language makes grammar coverage metrics a controversial topic. Concept accuracy, or concept error rate, see (Boros et al., 1996; Glass et al., 2000), has become increasingly popular as a measure of the extent to which the natural understanding functionality succeeds in capturing the key concepts in the user's input. It is defined on the basis of substitution, insertion, and deletion of concepts encoded in the meaning representation derived from the input utterance. Dialogue managers remain difficult to evaluate from a purely technical point of view, partly because much of their functionality is closely related with usability and are better regarded as targets for usability evaluation, cf. Section 3. One example of an important target for technical dialogue manager evaluation is reusability of task-independent parts. Like dialogue manager evaluation, response generation evaluation is to a large extent intimately related to usability evaluation. However, evaluation of the grammatical correctness of spoken output may be regarded as technical and quasi-quantitative evaluation.

At system level, several quantitative criteria have been proposed, e.g., real-time performance and robustness measured in terms of number of crashes. Two other metrics quantify how effectively a user can provide new information to a system (Glass et al., 2000): query density measures the mean number of concepts introduced per user query while concept efficiency quantifies the average number of turns for each concept to be understood by the system. In offline batch mode re-processing of user dialogues, comparison can be made between the dialogue state derived from the transcribed user utterance and the one derived from the recogniser outputs. This captures evaluation of understanding, discourse resolution and dialogue modelling as well as recognition. It is also used as a means of regression analysis/testing when system developers have made changes to the dialogue module and want to make sure that nothing has been unintentionally broken as a consequence, see (Polifroni and Seneff, 2000; Glass et al., 2000).

The emergence of spoken multimodal systems poses new challenges for the technical evaluation of SDSs and their components. For instance, at component level, we are beginning to need metrics for evaluating gesture recognisers, facial expression recognisers, gesture interpreters, facial expression and emotion interpreters, gesture and facial expression renderers, and a growing diversity of multimodal input fusion functionality. Sub-component reusability for the more complex dialogue and conversation managers needed will remain an important issue. At system level, new technical evaluation methodologies and metrics will be needed as well.

At this juncture, however, as experience is being gathered on technical solutions for spoken multimodal systems, it seems that the research focus is primarily on how to evaluate the usability of these systems, cf., e.g., the NICE project (Dybkjær et al., 2003) and the chapters in this part of the book. One

reason may be that there are more unknown usability factors than technical factors; another, that there is a common pattern involved, i.e. that usability and qualitative evaluation issues tend to become addressed at an earlier stage than do quantitative and technical issues.

2.2 Usability and Usability Evaluation

Usability remains difficult to get right even in unimodal SDSs. In general, a usable SDS must satisfy those user needs which go beyond the need for appropriate functionality, and it must be easy to understand and interact with, especially in the case of walk-up-and-use systems. Interaction should be smooth rather than bumpy and error-prone, and the user should feel in control throughout the dialogue with the system. Moreover, as SDSs begin to serve purposes other than factual information exchange, usability evaluation must be extended to address, e.g., educational value and entertainment value. To develop usable SDSs we need knowledge about issues, such as user reactions to SDSs in the field, users' linguistic, para-linguistic and non-linguistic behaviour, their comprehension of the corresponding system behaviour, and the main factors which determine overall user satisfaction.

Usability evaluation usually concerns the SDS as a whole and is typically done by developers and users. Most usability measures are qualitative or subjective but quantitative criteria also exist, such as transaction success rate, task completion time, turn correction ratio, and number of interaction problems. As a rule, usability should be factored in from the very beginning of the SDS development process. For this reason, it is recommended to have close interaction with representative users from early on. However, this does not in itself guarantee good usability. Additional help can be found in existing knowledge of important factors which affect usability, cf. Section 3.

So far, usability evaluation has been done mainly on task-oriented SDSs. However, usability evaluation is now moving into non-task oriented areas, such as conversational entertainment systems which do not assume that users perform particular tasks. This poses new demands on finding appropriate metrics and methods for usability evaluation.

Evaluation of commercial SDSs is often kept secret. Obviously, however, a crucial parameter for commercial systems is user satisfaction which is related to factors such as call statistics, transaction success, and users' opinion of the system. Contracts often include requirements to the minimum transaction success rate which must be met when users interact with the final system. However, it is well-known that a high transaction success rate does not necessarily imply happy users. It is also known that test subjects, even if representative of the target user group, may behave and judge differently from real users, for instance judging the system more positively than real users in the field. Eval-

uation metrics which could improve usability prediction would therefore seem highly desirable.

3. Empirical Generalisations

There are many individual projects which have produced, or aim to produce, generalisations based on empirical results. Among those which have had or have a major focus on evaluation and usability and thereby distinguish themselves from most other projects are ATIS (Air Travel Information Systems), Evalda, EAGLES (Expert Advisory Group on Language Engineering Standards), DISC (Spoken Language Dialogue Systems and Components: Best practice in development and evaluation), and the Danish Dialogue Project. Their contributions are presented in the following.

3.1 ATIS

The evaluation methodology for natural language understanding used in the ATIS project (1989-93) is objective response evaluation. The system's ability to understand the spoken input and respond appropriately is measured in terms of the information returned to the user. Thus, only the content of an answer retrieved from the database is assessed. Human annotation is required to identify the correct reference answers and to decide whether the query is ambiguous and/or answerable. This was considered to be easier to agree upon than to specify and evaluate a standard semantic representation.

The evaluation method (Pallett et al., 1995) is to automatically compare an annotated minimal/maximal reference answer pair with the system-generated answer. An answer is considered correct if it contains at least the set of fields in the minimal reference answer to the information explicitly requested by the subject. It should also contain no more than the set of fields described in the corresponding maximal reference answer. The maximal references use supplementary fields that can be reasonably included in an answer to the query. In other words, the system's response must be within a pre-defined range so as to avoid response overgeneration, i.e., the generation of correct answers by including all possible facts, rather than by understanding the requested information. The minimal reference answer was generated using NLPARSE whereas the maximal reference answer had to be defined manually. NLPARSE is a Texas Instruments proprietary system made available to the ARPA research community for the ATIS application. In a Wizard-of-Oz (WOZ) setup, the wizard input, an NL-parse paraphrase, is provided to NLPARSE which then simulates a system response to the user. The *Principles of Interpretation* document accompanying the data provides guidelines for annotators and system developers. Answers may contain scalars, booleans or tables. An automatic *Spoken Language System Answer Comparator* provided by NIST compares

the answer generated by the system with the minimal/maximal reference answer pair (Ramshaw and Boisen, 1990). Scalar answers are compared by comparing values. For table-based responses, the comparator explores each possible mapping from the required columns found in the specification to the actual columns found in the answer.

A strong feature of the ATIS evaluation method is that it supports regression analysis, i.e., developers can make changes to, e.g., the dialogue module and ensure that the system still behaves the same way at the dialogue level in responding to a broad range of different user queries. However, there are also several drawbacks of the method, including:

- the dialogue interaction must be entirely user-initiated since the assumption is that context resolution can be performed knowing only the user half of the conversation;
- it requires a static frozen database which is unrealistic for real dialogue systems;
- there is enormous overhead involved in the acquisition of, and adherence to, rigid standards of correctness;
- research in planning and language generation is stifled since the response is only evaluated as a tabular entry.

It follows that there is an additional burden involved in evaluating systems using real databases and mixed-initiative dialogue. If multimodal interaction is added, it becomes difficult to implement mechanisms to assure system reliability, which is a problem for system development. These problems notwithstanding, it may be concluded that the ATIS evaluation methodology succeeded in providing the community some unquestionable benchmarks.

3.2 Evalda

The French Evalda project[1]was launched in 2002 as part of the Technolangue programme (Mariani, 2002) and is coordinated by ELRA (Evaluation and Language Resources Agency)[2]. The goal of the project is within a 3-year period to establish a permanent evaluation infrastructure for the language engineering sector in France and for the French language. A first aim is to collect reusable knowledge and technology in terms of organisation, logistics, language resources, evaluation protocols, methodologies and metrics, as well as major industrial and academic actors in the field. A second aim is to run evaluation campaigns involving linguistic technologies in written and spoken media and covering various aspects of language processing and human-computer interaction. The campaigns are largely based on black-box evaluation protocols

and quantitative methods, drawing on and expanding previous evaluation campaigns, such as ARC-AUPELF[3], GRACE[4], and TREC[5]. To enable comparison of performance and benchmarking of language engineering tools, it is considered crucial that the evaluations envisaged are reproducible by third parties, using the resources assembled in the project. Evalda will make available all its evaluation resources by the end of the project in the form of an evaluation package. Eight evaluation campaigns will be run, cf. (Dybkjær et al., 2004). Contrary to the USA, SDS evaluation campaigns are not common in Europe. France is, in fact, the only European country which from early on has conducted larger-scale open evaluation campaigns in language engineering using quantitative black-box evaluation protocols.

3.3 EAGLES

EAGLES[6] (1993-1998) (King et al., 1996) aimed to develop commonly agreed specifications and guidelines for various aspects of language engineering, including evaluation issues. The approach was to collect and unify existing information and provide up-to-date reference documentation for use by researchers and developers as well as in standardisation initiatives. To reach a large audience, the EAGLES evaluation working group used the ISO 9000 norm series and proposed a strongly user-oriented methodology for application in adequacy evaluation and progress evaluation. The idea was to work in terms of classes of typical users, much in the same way that consumer associations target typical users of cars or washing machines when drawing up their product reports. User profiling can help determine the attributes of products which particular classes of users are interested in. Attribute values may then be determined for specific products.

An important point of departure for part of the work was the ISO 9126 (1991) standard on quality characteristics of software. With close contact to ISO, EAGLES looked into the modifications and extensions that would be necessary in order to apply the standard to the evaluation of language engineering systems in general, aiming to produce a formal quality model. ISO 9126 was later revised and split into ISO 9126-1 (1998) and ISO 14598 (2000). ISO 9126-1 (1998) focuses on the quality model which was missing in ISO 9126 (1991), whereas evaluation became the sole topic of the ISO 14598 series.

EAGLES also worked on SDS evaluation recommendations. A problem was that only a few SDSs had been systematically evaluated by the mid-1990s, most of which performed relatively simple tasks. Thus (Gibbon et al., 1997) clearly states that the EAGLES recommendations on SDS evaluation are provisional. EAGLES distinguishes between glass-box evaluation and black-box evaluation, see also (Simpson and Fraser, 1993). Glass-box evaluation is meant for evaluation of sub-components and their contribution to the overall

behaviour of the system. The term glass-box is used because the internals of components can be inspected. Black-box evaluation, which is a familiar term from computer science, views a component or entire system as a black box and evaluates some aspect of its overall performance. Both quantitative and qualitative measures are proposed for black-box evaluation. Quantitative measures proposed include: average number of exchanges to obtain relevant responses, task completion rate, transaction success rate, system response time, and terseness of the system's answers. Qualitative measures include 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 evaluation of systems. The proposed key comparative evaluation measures include dialogue duration, turn duration, contextual appropriateness, correction rate, and transaction success rate.

The EAGLES recommendations which are documented in the EAGLES handbook (Gibbon et al., 1997) were not only conceived at a time when evaluation experience was sparse. At the time, evaluation was often based on looking at single user-system turn pairs in isolation without regard to context, cf. ATIS. The EAGLES group was aware of the need for evaluating single dialogue turns in the larger discourse context.

EAGLES no doubt contributed to progress in SDS evaluation by articulating its underlying complexity and linking up with standardisation. Many of the proposed metrics are still useful although insufficient for evaluating the full variety of SDSs today. Two main problems with the EAGLES evaluation recommendations probably are that (i) it can be costly and cumbersome to carry out evaluation precisely as prescribed, and (ii) the methodology can be difficult to follow and may not fit equally well into different projects.

3.4 DISC

Based on its academic and industrial partners' broad collective experience in SDSs development, the ESPRIT Long-Term Research project DISC[7] (1997-1998) developed a first dialogue engineering best practice methodology (Dybkjær et al., 1998a). The idea was that, although reference methodologies exist for software engineering in general, no such reference methodology existed for the development and evaluation of SDSs and components in particular. DISC addressed both technical evaluation and usability. Evaluation from the point of view of the end-user was not fully addressed due to data scarcity. For the same reason, DISC did not systematically address the multimodal aspects of the systems analysed.

Work in DISC was grounded in detailed analyses of the properties of, and development processes for, approximately 25 SDS systems and components for different application domains and different languages. Correspondingly,

the DISC evaluation methodology is based on a best practice *grid* and *life-cycle*. A grid defines a space of aspect-specific issues which the developer may have to take into account, such as, in dialogue manager development: who should have the initiative, the system and/or the user? For each issue, the available solution *options* are laid out in the grid together with the *pros* and *cons* for choosing a particular option. A life-cycle includes recommendations on how the development process for an aspect and its options should be carried out, such as how to do a requirements analysis in the dialogue manager specification phase. The six SDS *aspects* analysed are: speech recognition, speech generation, natural language understanding and generation, dialogue management, human factors, and systems integration.

DISC produced a comprehensive set of guidelines and heuristics to help determine how a given system or component development task relates to the proposed model. For technical evaluation, DISC proposed for each of its aspects, except human factors, *what* to evaluate, i.e. the full set of properties which should be evaluated, and *how* to evaluate, i.e. the evaluation criteria to apply and how to apply them correctly at the right stages during the development life-cycle. Each evaluation criterion is described using a standard evaluation template (Bernsen and Dybkjær, 2000). As to what to evaluate, DISC defined an aspect-specific notion of evaluation completeness, i.e. that every chosen option from the aspect-specific grid must be evaluated.

The DISC evaluation template supports evaluation correctness. It is a model of what the developer needs to consider when planning to evaluate a particular property of an SDS or component. This knowledge is specified by ten template entries (Bernsen and Dybkjær, 2000): property evaluated, system part evaluated, type of evaluation, methods, symptoms, life-cycle phase(s), importance, difficulty, cost and tools. For each property to be evaluated, i.e. each selected option in the aspect-specific issue space, an empty ten-entry template must be filled by the developer. Filled templates are illustrated at `http://www.disc2.dk/slds/`. If the grid has been used during development, it is easy to generate evaluation criteria for the application by simply including the grid options selected. The harder part is to fill a template per criterion. This requires knowledge of available evaluation methods and metrics, when to apply them, and what to look for in the data. One must also be able to estimate evaluation costs and the risks involved in refraining from evaluating a particular system property.

The DISC approach to complete and correct evaluation would need updating to reflect the issue/option spaces facing today's developers as well as new evaluation metrics. Furthermore, since no developer carries out complete evaluation in practice because of the time and cost involved, developers must be able to carefully judge where to invest their evaluation efforts. How to decide

on this important issue was not addressed in DISC nor was comparative system evaluation.

3.5 Usability Guidelines

Building on results from the Danish dialogue project (Dybkjær et al., 1998b) and DISC, Dybkjær and Bernsen (2000) discuss existing knowledge of SDS usability evaluation. They propose that the general design goal of creating usable walk-up-and-use, shared-goal SDSs may be systematically pursued by addressing 13 key usability issues. These issues are aimed at carving up the complex space of SDS usability into intuitively satisfactory and complementary segments. Most issues have implications for the technical development of particular SDS components, such as speech recogniser optimisation or various aspects of interaction optimisation in which the dialogue manager has a central role. The issue of when (not) to use speech in applications highlights the fact that speech is not always the right modality choice for interactive systems, cf. Section 5.

- Input recognition accuracy: good recogniser quality is a key factor in making users confident that the system will successfully get what they say.
- Naturalness of user speech: speaking to an SDS should feel as easy and natural as possible. What is being experienced as natural input speech is also highly relative to the system's output phrasing. Thus, the system's output language should be used to control-through-priming users' input language, so that the latter is manageable for the system whilst still feeling natural to the user.
- Output voice quality: a good SDS output voice quality means that the system's speech is clear and intelligible, does not demand additional listening effort, is not particularly noise-sensitive or distorted by extraneous sounds, has natural intonation and prosody, uses an appropriate speaking rate, and is pleasant to listen to (Karlsson, 1999). Taken together, these requirements still remain difficult to meet today.
- Output phrasing adequacy: the contents of the system's output should be correct, relevant and sufficiently informative without being over informative.
- Feedback adequacy: the user must feel confident that the system has understood the information input in the way it was intended, and the user must be told which actions the system has taken and what the system is currently doing.

- Adequacy of dialogue initiative: to support natural interaction, an SDS needs a reasonable choice of dialogue initiative, depending on factors, such as the nature of the task, users' background knowledge, and frequency of use.
- Naturalness of the dialogue structure: dialogue designers may have to impose some amount of structure onto the dialogue, determining which topics (or sub-tasks) could be addressed when. The structure must be natural to the user, reflecting the user's intuitive expectations.
- Sufficiency of task and domain coverage: even if unfamiliar with SDSs, users often have rather detailed expectations to the information or service obtainable from the system. It is important that the system meets these expectations.
- Sufficiency of reasoning capabilities: contextually adequate reasoning represents a classical problem in the design of natural interaction. SDSs must incorporate both facts and inferences about the task as well as general world knowledge in order to act as adequate interlocutors.
- Sufficiency of interaction guidance: users should feel in control during interaction. Useful help mechanisms may be an implicit part of the dialogue, be available by asking for help, or be automatically enabled if the user is having problems repeatedly, e.g., in being recognised.
- Error handling adequacy: this issue may be decomposed along two dimensions. Either the system or the user initiates error handling meta-communication. When error-handling meta-communication is initiated, it is either because one party has failed to hear or understand the other, because what was heard or understood is false, or because what was heard or understood is somehow in need of clarification.
- Sufficiency of adaptation to user differences: it is useful to distinguish between system expert/domain expert, system expert/domain novice, system novice/domain expert and system novice/domain novice users. An SDS needs not support all four groups.
- Modality appropriateness: the dialogue designers should make sure that spoken input and output, possibly combined with other input/output modalities, is an appropriate modality choice for the planned application. See Section 5 for more detail.

In addition, Dybkjær and Bernsen (2000) discuss user satisfaction measures and metrics for counting the number of interaction problems, both of which provide important information on system and component usability. The work

on interaction problem metrics is based on a body of guidelines for cooperative dialogue design (Bernsen et al., 1998) which extends Gricean cooperativity theory (Grice, 1975). These guidelines may be compared with the guidelines for advanced spoken dialogue design developed in a UK project with business exploitation in mind (Gilbert et al., 1999).

This is not an exhaustive list, of course, but it probably covers a good deal of usability basics for task-oriented SDSs. An important problem is that too little is known about the differential effect on general system usability of each of the individual elements on the list. A point missing is that of cultural differences in the perception of SDS usability, such as the degree of system politeness required, which remain poorly understood.

4. Frameworks

In this section we describe the PARADISE (Paradigm for Dialogue System Evaluation) (Walker et al., 1997) framework which addresses usability evaluation of unimodal SDSs. Attempts have been made in the SmartKom project to adapt and extend the PARADISE framework to cope with the evaluation of task-oriented multimodal SDSs (Beringer et al., 2002). However, there does not seem to exist results yet on how well this extended framework works.

4.1 PARADISE

It is a well-recognised fact that too little is known about how to predict overall user satisfaction, i.e., how users will receive a particular SDS. Some would argue that, from a practical standpoint, usability boils down to what users like and prefer although user satisfaction and usability is not one and the same thing. Others would argue that, since they are not identical, the field would do better to keep them separate. One reason why usability and user satisfaction are different quantities is that the latter is very much a function of a constantly changing environment of product availability, cost, and competing technologies, whereas the former is a constant which depends on human nature.

The PARADISE framework views user satisfaction as a measure of system usability and tries to predict user satisfaction from objectively measurable performance parameters. The framework was first applied to SDSs built at AT&T and later adopted as evaluation framework for the DARPA Communicator project (Walker et al., 2002; Sanders et al., 2002). PARADISE has been used in several other projects as well, e.g., (Hjalmarson, 2002). The PARADISE model assumes that the primary objective of an SDS is to maximise user satisfaction (Walker et al., 2000). Task success and various dialogue costs relating to efficiency and quality contribute to user satisfaction. To maximise user satisfaction, one must maximise task success and minimise dialogue costs. Task success is measured as the perceived task completion by users to-

gether with the observed task completion. Efficiency is measured through, e.g., elapsed time and number of utterances. Quality is measured via, e.g., recognition score, repair and help (Walker et al., 1997). 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.

The basic claim is that a performance function can be derived by applying multivariate linear regression with user satisfaction as the dependent variable and task success, dialogue quality, and dialogue efficiency measures as independent variables (Walker et al., 2000). Modelling user satisfaction as a function of task success and dialogue cost is intended to lead to a predictive performance model of SDSs, enabling prediction of user satisfaction based on measurable parameters which can be found in log-files, and eventually avoiding costly and hard-to-interpret subjective user evaluation.

It is probably too early for any final assessment of the PARADISE framework. For the moment, there is no better proposal of its kind. Several potential weaknesses may be noted, however:

- The framework may make too tight a coupling between user satisfaction and usability. What users like is significant to usability, but what they like changes depending on what is available, cf. above.
- It is questionable if the model can be concluded to have any reliable predictive power as regards user satisfaction based on log-files alone. Clearly, the independent variables measured are not the only contributors to user satisfaction. An issue which may be difficult to handle quantitatively based on logfiles concerns users actually getting what they want. This may be relatively easy to decide in controlled environments. However, how can we decide from the logfiles from, e.g., a frequently asked questions system whether users actually feel that they got the information they needed or just did not find it in the system? In this case, apparent task completion may be high even if users are dissatisfied because they did not obtain the information they wanted.
- User questionnaires are hard to interpret and there does not seem to exist any strong theoretical foundation for the selection of questions to include. So, how do we know that PARADISE actually correlates objective metrics with "real" user satisfaction? Since the PARADISE questionnaire has not been proven to be reliable and valid for eliciting information about user satisfaction, we cannot be certain that the results obtained with it actually reflect users' real attitude (Larsen, 2003).
- For the moment, application of PARADISE is restricted to controlled experiments, which makes the framework unsuited for tests with real

users having real needs in real environments. Test subjects tend to behave differently from real users. In the Dutch part of the ARISE system, for instance, subjects were very satisfied. However, when the commercial version was launched, user satisfaction dropped dramatically. The drop might be due to in-the-field factors, such as waiting time and price which are not considered in PARADISE, but, ultimately, these factors co-determine user satisfaction.

- User satisfaction is inherently a difficult parameter to deal with. In experiments one can sometimes find dialogues which seem to be inefficient and of low quality. Nevertheless, the user seems happy about the dialogue if the questionnaire or interview data are to be believed. The opposite may also be the case, i.e., the dialogue seems smooth and efficient but the user is not overly satisfied. For some users, task completion may be what really counts while, for others, efficiency or some third parameter is the more important factor. A predictive model might straighten out these differences to some extent but we should be aware that user needs may differ more widely than assumed in the model. In education and entertainment applications, for instance, ultimate educational or entertainment value(s) may be far more important than, for instance, task efficiency, if, indeed, task efficiency is relevant at all.

5. Multimodal SDSs Usability, Generalisations and Theory

Solid empirical generalisations on the usability of multimodal SDSs are emerging. It is well-known, for instance, that system behaviour causes expectations as to what the system can do. Thus, if a human voice and fully natural language is used for spoken output, users may tend to forget that they are interacting with a limited-capability system, expecting human-like capabilities instead. This generalisation seems extensible to multimodal SDSs. For example, Bickmore and Cassell (2002) evaluated the effects on communication of a life-like embodied conversational real-estate agent. They concluded that users tend to compare such animated agent interlocutors with humans rather than machines, judging the system in an unexpected negative fashion as a result. To work with users, life-like animated agents need a high degree of naturalness and personally attractive features communicated non-verbally. This imposes a tall research agenda on spoken and non-verbal output performance, requiring conversational abilities both verbally and non-verbally, cf. also (Cassell et al., 2000). Another empirical generalisation and one supported by modality theory (see below), is that spoken and pointing input, and spoken and graphics output, go well together, see, e.g., (Oviatt, 1997; Roth et al., 1997; Cohen et al., 1997).

So far, we have not gone into much detail with the theoretical underpinnings of the approaches to usability presented, although these certainly exist in many cases. However, when addressing the usability of spoken multimodal systems, in particular, it seems important to point out that, potentially, given their huge scope and the early stage of their investigation, these systems could benefit from the application of a wide range of theoretical approaches. Moreover, several of those approaches definitely do not belong to the standard body of knowledge of the SDS community. Approaches range from film theory and theories of conversation applied to conversational animated agent SDSs for entertainment, through classical psychological theory, such as Gestalt theory, and theories of emotional behaviour, gesture, facial action, etc., to AI planning theory, modality theory, and more. Below, we limit ourselves to briefly presenting a single theoretical approach, i.e. modality theory.

A core question in developing usable spoken multimodal systems is whether or not speech is appropriate for the application to be developed. This is a complex question because the answer depends on many different factors, such as the type and purpose of the application, the application environment (Bühler et al., 2002), bandwidth and transmission channel stability, prioritised performance parameters, such as speed or efficiency versus time to reflect, learning overhead, and the intended users. Clearly, however, a basic factor is the properties of the modalities involved. These are investigated in modality theory based on an exhaustive, hierarchically organised taxonomy of unimodal input/output modalities accessible to human hearing, vision, and touch, see (Bernsen, 1994). Each modality has a number of objective modality properties, such as the property of sound that it is omnidirectional, which implies that speech is omnidirectional as well, or the property of speech in a known language that it has high saliency compared to other acoustic modalities. Modality theory has been applied to the speech functionality problem of when (not) to use speech in unimodal and multimodal applications, see (Bernsen, 2002) for more detail.

Two comprehensive studies of the literature on unimodal and multimodal SDSs written between 1992 and 1998 showed that some 95% of 273 "blind"-selected speech functionality claims made by various authors on when (not) to use speech in unimodal or multimodal contexts could be evaluated as being either true, false, or supported by modality theory. An interesting finding was that the evaluation could be based on only 25 modality properties (Bernsen, 1997; Bernsen and Dybkjær, 1999) of the kind exemplified above. Moreover, the first study looked at 120 early speech functionality claims which mostly concerned unimodal speech input and/or output, whereas the second study looked at 153 claims which were made later in the 1990s and which included a large fraction of claims about speech in multimodal combinations. Nevertheless, it was only necessary to augment the 18 modality properties used for

evaluation in the first study by seven, mostly non-speech, modality properties in order to evaluate the new data set. In other words, there is evidence that the theoretical basis needed for evaluating the use of speech in any possible modality combination may be limited and achievable.

6. Discussion and Outlook

With the technical advances and market growth in the SDS field, evaluation and usability of unimodal and multimodal SDSs are becoming crucial issues. We have discussed the state-of-the-art in evaluation and usability and reviewed a number of initiatives which have collected, or built on and contributed to the consolidation of the pool of knowledge we have on SDS evaluation and usability.

There are still important gaps in our knowledge of unimodal, task-oriented SDSs evaluation and usability, and the increasing sophistication even of these systems continues to demand new evaluation metrics. Moreover, the field is moving rapidly beyond the standard task-oriented, speech-only SDS towards multimodal SDSs, mobile systems, situation-aware systems, location-aware systems, internet access systems, educational systems, entertainment systems, etc. In fact, technology development may appear to speed further ahead of the knowledge we already have on evaluation and usability, increasing the proportion of what we do not know compared with what we do know. In the following, we discuss some issues essential to the development of more advanced SDSs.

Online user modelling. By online user modelling we understand the ability of a system to create a model of some property, or properties, of its user at run-time in order to adapt its dialogue behaviour to that property. In generic user modelling, the property is characteristic of a group of users, such as that they are novices in using the system. In individual user modelling, the system builds a model of the property of each individual user, for instance, of the user's hotel reservation preferences, and then uses the model to make it easier for the user to carry out some task. Individual user modelling is, of course, only suitable for frequently used systems. online user modelling for SDSs is receiving increasing attention today for several reasons. Mobile devices (mobile phones, PDAs, note pads, in-car devices, etc.) are usually personal (or quasi-personal) belongings used on a fairly frequent basis. The user of these devices may benefit from functionality which builds knowledge of the individual user. Secondly, many user groups could benefit from generic user modelling functionality. For instance, novice users could receive more extensive interaction guidance; users who repeatedly make particular types of error could be helped by explicit advice or by adaptation of dialogue structure, initiative distribution, and otherwise. Only a few applications of online user modelling in SDSs

have been reported in the literature so far. Bernsen and Dybkjær (2004a) describe online individual user modelling for an in-car SDS at a fairly general level while Bernsen (2003) describes an application of individual online user modelling to the hotel reservation task in the same in-car SDS. Komatani et al. (2003) describe an application of generic online user modelling which adapts the system's information level to user experience with a bus information system. Whittaker and Walker (2004) show via a Wizard of Oz experiment the benefit of individual user modelling in a restaurant application. General online user modelling is an active research area. See, for instance, the 9th International Conference on User Modelling in 2003[8] (Brusilovsky et al., 2003). Some key questions to be considered by developers of online user modelling are: (i) is the user modelling functionality feasible and (ii) will it 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, the user may experience that adaptation fails because of overly primitive update algorithms or insufficient information about when the user model has been used.

Emotions and personality. Recognition of the emotional states of users followed by appropriate system reactions may contribute to perceived system naturalness. Ongoing research addresses the recognition of facial expressions of emotion, cf. (Ekman and Friesen, 1975; Cohen et al., 2003), and the recognition of prosodic cues to emotion (Batliner et al., 2000; Hirschberg et al., 2001). The ERMIS project[9] on Emotionally Rich Man-Machine Interaction Systems, 2002-2004, analyses speech and face input signals in order to equip systems with the ability to recognise emotions and interact with users in a more natural and user-friendly way. Emotion interpretation could be used to, e.g., change dialogue strategy if the user appears upset. Also, output expression of emotion is an active research topic, see, e.g., (André et al., 2004), and some speech synthesisers are beginning to accept emotion tags. For example, emotion-dependent prosody in synthetic speech is strongly needed in several multimodal entertainment SDSs in current development. System output may among other things be used to communicate certain personality features and thereby influence the user's spoken input, cf. (Oviatt et al., 2004) who investigated the influence of introvert versus extrovert voices on children's vocal behaviour in a spoken multimodal SDS with animated marine animals.

Non-task-oriented dialogue. So far, almost all SDSs have been task-oriented applications. However, research has started in non-task-oriented dialogue, cf., e.g., (Gustafson et al., 1999; Bernsen and Dybkjær, 2004b). In the absence of task constraints, the dialogue may have to follow principles entirely different from task-oriented dialogue. Little is known at this point about the novel usability issues arising in this kind of dialogue. Some of the usability

issues discussed in Section 3 will clearly become irrelevant, such as sufficiency of task coverage, and others may suffer the same fate, such as informativeness. Instead, other issues may move into focus, such as conversational naturalness, turn-taking adequacy, and others which will depend on the type of application involved.

Mobile versus static environments. Speech may be a good choice in many mobile environments because of its modality properties of being hands-free and eyes-free. On the other hand, speech is not very private in public spaces because it is omnidirectional, it is potentially disturbing to others because it is highly salient, and speech recognisers remain sensitive to noise. Graphics (including text) output and, e.g., pen-based input may be useful additions because these are not sensitive to noise, do not disturb others, and are usually sufficiently private. Mobile SDSs raise a number of other evaluation issues which have not been fully solved yet, including how (not) to use, and when (not) to use, small and very small screens (Almeida et al., 2002; Sturm et al., 2004), 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 (Bühler et al., 2002; Gärtner et al., 2001; Minker et al., 2004; Bernsen and Dybkjær, 2004).

User preferences and priorities. One thing which can really make life hard for developers are user preferences. For instance, users do not necessarily prefer what is empirically the most efficient modality combination. Thus, some users may prefer pen-based or keypad-based input to spoken input simply because they feel more familiar with GUI-style interfaces, cf. (Sturm et al., 2004) who analysed the behaviour and satisfaction of subjects interacting with a multimodal SDS offering speech input/output, pointing input and graphics output, and (Jameson and Klöckner, 2004) who 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. In other words, depending on the target user group(s), alternative modalities may have to be enabled 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 recommendable and why online user modelling appears attractive.

Another aspect to user preferences is what is perceived as an adequate presentation of information within a given modality, cf. (Geldof and Dale, 2004) who compared two ways of textual presentation of route descriptions.

In fact, one reason why different users may score the same system very differently in terms of usability could be that they have different preferences and priorities. Some preferences we can design for, such as modality preferences

and different presentation preferences. Others, however, are hard to cope with. For example, some users may prioritise speed (no queues on the line) or economical benefit (queues but cheap or free calls), while others prioritise human contact which, by definition, cannot be satisfied by a system. The question here is if we can create systems with a usability profile that will make these users change their priorities, and exactly which usability issues must be resolved to do so.

Concluding remarks. The issues discussed in this section are probably just a few of those which should be considered in a systematic approach to evaluation and usability of multimodal, mobile, and domain-oriented SDSs. This approach could lead to a best practice, pre-standard guide for usability and evaluation. EAGLES and DISC took major steps in this direction for unimodal, task-oriented SDSs. Arguably, the expanding SDSs field could benefit from an extension of that work to include multimodal, mobile and domain-oriented SDSs. The foundations for such an approach is just about at hand in the form of a large and growing body of results from very different projects which have built and evaluated next-generation SDSs.

Notes

1. `http://www.elda.fr/rubrique25.html`
2. `http://www.elda.fr/index.html`
3. `http://www.limsi.fr/tlp/aupelf.html`
4. `http://www.limsi.fr/tlp/grace/`
5. `http://trec.nist.gov/`
6. `http://lingue.ilc.pi.cnr.it/EAGLES96/home.html`
7. `http://www.disc2.dk/`
8. `http://www2.sis.pitt.edu/~um2003/`
9. `http://www.image.ntua.gr/ermis/`

References

Almeida, L., Amdal, I., Beires, N., Boualem, M., Boves, L., den Os, L., Filoche, P., Gomes, R., Knudsen, J. E., Kvale, K., Rugelbak, J., Tallec, C., and Warakagoda, N. (2002). Implementing and evaluating a multimodal tourist guide. In *Proceedings of International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems*, pages 1–7, Copenhagen, Denmark.

André, E., Dybkjær, L., Minker, W., and Heisterkamp, P., editors (2004). *Affective Dialogue Systems*, volume 3068 of *LNCS/LNAI Lecture Notes*. Springer-Verlag, Berlin/Heidelberg, Germany.

Batliner, A., Fischer, K., Huber, R., Spilker, J., and Nöth, E. (2000). Desperately seeking emotions: Actors, wizards, and human beings. In *Proceedings*

of ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research, pages 195–200, Belfast, United Kingdom.

Beringer, N., Kartal, U., Louka, K., Schiel, F., and Türk, U. (2002). Promise - a procedure for multimodal interactive system evaluation. In *Proceedings of LREC Workshop on Multimodal Resources and Multimodal Systems Evaluation*, pages 77–80, Las Palmas, Gran Canaria, Spain.

Bernsen, N. O. (1994). Foundations of multimodal representations. a taxonomy of representational modalities. *Interacting with Computers*, 6(4):347–371.

Bernsen, N. O. (1997). Towards a tool for predicting speech functionality. *Speech Communication*, 23:181–210.

Bernsen, N. O. (2002). Report on user clusters and characteristics. Technical VICO Report D10, NISLab, University of Southern Denmark.

Bernsen, N. O. (2003). On-line user modelling in a mobile spoken dialogue system. In *Proceedings of European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 737–740, Geneva, Switzerland.

Bernsen, N. O., Dybkjær, H., and Dybkjær, L. (1998). *Designing Interactive Speech Systems. From First Ideas to User Testing*. Springer-Verlag, Berlin/-Heidelberg, Germany.

Bernsen, N. O. and Dybkjær, L. (1999). Working paper on speech functionality. Technical Report Esprit Long-Term Research Project DISC Year 2 Deliverable D2.10., NISLab, University of Southern Denmark.

Bernsen, N. O. and Dybkjær, L. (2000). A methodology for evaluating spoken language dialogue systems and their components. In *Proceedings of International Conference on Language Resources and Evaluation (LREC)*, pages 183–188, Athens, Greece.

Bernsen, N. O. and Dybkjær, L. (2004a). Enhancing the usability of multimodal virtual co-drivers. In Minker, W., Bühler, D., and Dybkjær, L., editors, *Spoken Multimodal Human-Computer Dialogue in Mobile Environments*. Kluwer Academic Publishers, Dordrecht, The Netherlands. (this volume).

Bernsen, N. O. and Dybkjær, L. (2004b). Evaluation of spoken multimodal conversation. In *Proceedings of International Conference on Multimodal Interfaces (ICMI)*, Pennsylvania State University, Pennsylvania, USA.

Bickmore, T. and Cassell, J. (2002). Phone vs. face-to-face with virtual persons. In *Proceedings of International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems*, pages 15–22, Copenhagen, Denmark.

Boros, M., Eckert, W., Gallwitz, F., Görz, G., Hanrieder, G., and Niemann, H. (1996). Towards understanding spontaneous speech: Word accuracy vs. concept accuracy. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, volume 2, pages 1009–1012, Philadelphia, Pennsylvania, USA.

Bossemeyer, R. W. and Schwab, E. C. (1991). Automated alternate billing services at Ameritech: Speech recognition and the human interface. *Speech Technology Magazine*, 5:24–30.

Brusilovsky, P., Corbett, A., and de Rosis, F., editors (2003). *User Modeling 2003*, volume 2702 of *LNCS/LNAI Lecture Notes*. Springer-Verlag, Berlin/-Heidelberg, Germany.

Bühler, D., Minker, W., Häussler, J., and Krüger, S. (2002). Flexible multimodal human-machine interaction in mobile environments. In *ECAI Workshop on Artificial Intelligence in Mobile System (AIMS)*, pages 66–70, Lyon, France.

Cassell, J., Sullivan, J., Prevost, S., and Churchill, E., editors (2000). *Embodied Conversational Agents*. MIT Press, Cambridge, Massachusetts, USA.

Cohen, I., Sebe, N., Chen, L., Garg, A., and Huang, T. (2003). Facial expression recognition from video sequences: Temporal and static modeling. *Computer Vision and Image Understanding, Special Issue on Face Recognition*, 91(1-2):160–187.

Cohen, P., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L., and Clow, J. (1997). QuickSet: Multimodal interaction for distributed applications. In *Proceedings of ACM International Conference on Multimedia*, pages 31–40, Seattle, Washington, USA.

Dybkjær, L. and Bernsen, N. O. (2000). Usability issues in spoken language dialogue systems. *Natural Language Engineering, Special Issue on Best Practice in Spoken Language Dialogue System Engineering*, 6:243–272.

Dybkjær, L., Bernsen, N. O., Blasig, R., Buisine, S., Fredriksson, M., Gustafson, Martin, J. C., and Wirén, M. (2003). Evaluation criteria and evaluation plan. Technical Report NICE Deliverable D7.1, NISLab, University of Southern Denmark.

Dybkjær, L., Bernsen, N. O., Carlson, R., Chase, L., Dahlbäck, N., Failenschmid, K., Heid, U., Heisterkamp, P., Jönsson, A., Kamp, H., Karlsson, I., v. Kuppevelt, J., Lamel, L., Paroubek, P., D., and Williams (1998a). The DISC approach to spoken language systems development and evaluation. In *Proceedings of International Conference on Language Resources and Evaluation (LREC)*, pages 185–189, Granada, Spain.

Dybkjær, L., Bernsen, N. O., and Dybkjær, H. (1998b). A methodology for diagnostic evaluation of spoken human-machine dialogue. *International Journal of Human Computer Studies (special issue on Miscommunication)*, 48:605–625.

Dybkjær, L., Bernsen, N. O., and Minker, W. (2004). Evaluation and usability of multimodal spoken language dialogue systems. *Speech Communication*, 43(1-2):33–54.

Ekman, P. and Friesen, W. (1975). *Unmasking the Face: A guide to recognize emotions from facial clues*. Prentice Hall Trade.

Gärtner, U., König, W., and Wittig, T. (2001). Evaluation of manual vs. speech input when using a driver information system in real traffic. In *Proceedings of International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, pages 7–13, Aspen, Colorado, USA.

Geldof, S. and Dale, R. (2004). Segmenting route directions for mobile devices. In Minker, W., Bühler, D., and Dybkjær, L., editors, *Spoken Multimodal Human-Computer Dialogue in Mobile Environments*. Kluwer Academic Publishers, Dordrecht, The Netherlands. (this volume).

Gibbon, D., Moore, R., and Winski, R. (1997). *Handbook of Standards and Resources for Spoken Language Systems*. Walter de Gruyter.

Gilbert, N., Cheepen, C., Failenschmid, K., and Williams, D. (1999). Guidelines for advanced spoken dialogue design. http://www.soc.surrey.ac.uk/research/guidelines.

Glass, J., Polifroni, J., Seneff, S., and Zue, V. (2000). Data collection and performance evaluation of spoken dialogue systems: The MIT experience. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, volume 4, pages 1–4, Beijing, China.

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

Gustafson, J., Bell, L., Beskow, J., Boye, J., Carlson, R., Edlund, J., Granström, B., House, D., and Wirén, M. (2000). AdApt - a multimodal conversational dialogue system in an apartment domain. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, pages 134–137, Beijing, China.

Gustafson, J., Lindberg, N., and Lundeberg, M. (1999). The August spoken dialogue system. In *Proceedings of European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 1151–1154, Budapest, Hungary.

Hirschberg, J., Swerts, M., and Litman, D. (2001). Labeling corrections and aware sites in spoken dialogue systems. In *Proceedings of ACL SIGdial Workshop on Discourse and Dialogue*, pages 72–79, Aalborg, Denmark.

Hjalmarson, A. (2002). *Evaluating AdApt, A Multi-modal Conversational Dialogue System using PARADISE*. PhD thesis, KTH.

Jameson, A. and Klöckner, K. (2004). User multitasking with mobile multimodal systems. In Minker, W., Bühler, D., and Dybkjær, L., editors, *Spoken Multimodal Human-Computer Dialogue in Mobile Environments*. Kluwer Academic Publishers, Dordrecht, The Netherlands. (this volume).

Karlsson, I. (1999). A survey of existing methods and tools for development and evaluation of speech synthesis and speech synthesis quality in SLDSs. Technical Report DISC Deliverable D2.3.

King, M., Maegard, B., Schutz, J., and des Tombes, L. (1996). Eagles - Evaluation of natural language processing systems. Technical Report EAG-EWG-PR.2.

Komatani, K., Ueno, S., Kawahara, T., and Okuno, H. (2003). User modeling in spoken dialogue systems for flexible guidance generation. In *Proceedings of European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 745–748, Geneva, Switzerland.

Larsen, L. B. (2003). Assessment of spoken dialogue system usability - What are we really measuring? In *Proceedings of European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 1945–1948, Geneva, Switzerland.

Mariani, J. (2002). Technolangue: Language technology. In *Proceedings of International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Gran Canaria, Spain.

Minker, W., Haiber, U., Heisterkamp, P., and Scheible, S. (2004). Design, implementation and evaluation of the SENECA spoken language dialogue system. In Minker, W., Bühler, D., and Dybkjær, L., editors, *Spoken Multimodal Human-Computer Dialogue in Mobile Environments*. Kluwer Academic Publishers, Dordrecht, The Netherlands. (this volume).

Oviatt, S., Darves, C., Coulston, R., and Wesson, M. (2004). Speech convergence with animated personas. In Minker, W., Bühler, D., and Dybkjær, L., editors, *Spoken Multimodal Human-Computer Dialogue in Mobile Environments*. Kluwer Academic Publishers, Dordrecht, The Netherlands. (this volume).

Oviatt, S. L. (1997). Multimodal interactive maps: Designing for human performance. *Human-Computer Interaction, special issue on Multimodal interfaces*, 12:93–129.

Pallett, D., Fiscus, J. G., Fisher, W. M., Garofolo, J., Lund, B. A., Martin, A., and Przybocki, M. A. (1995). 1994 benchmark tests for the ARPA spoken language program. In *Proceedings of ARPA Workshop on Spoken Language Technology*, pages 5–36. Morgan Kaufmann, San Francisco, California, USA.

Polifroni, J. and Seneff, S. (2000). Galaxy-ii as an architecture for spoken dialogue evaluation. In *Proceedings of International Conference on Language Resources and Evaluation (LREC)*, pages 725–730, Athens, Greece.

Ramshaw, L. A. and Boisen, S. (1990). An SLS answer comparator. Technical report, BBN Systems and Technologies Corporation. SLS Note 7.

Roth, S. F., Chuah, M. C., Kerpedjiev, S., Kolojejchick, J., and Lucas, P. (1997). Towards an information visualization workspace: Combining multiple means of expression . *Human-Computer Interaction*, 12:131–185.

Sanders, G. A., Le, A. N., and Garofolo, J. S. (2002). Effects of word error rate in the DARPA Communicator data during 2000 and 2001. In *Proceedings of*

International Conference on Spoken Language Processing (ICSLP), pages 277–280, Denver, Colorado, USA.

Simpson, A. and Fraser, N. (1993). Blackbox and glassbox evaluation of the SUNDIAL system. In *Proceedings of European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 1423–1426, Berlin, Germany.

Sturm, J., Cranen, B., Terken, J., and Bakx, I. (2004). Effects of prolonged use on the usability of a multimodal form-filling interface. In Minker, W., Bühler, D., and Dybkjær, L., editors, *Spoken Multimodal Human-Computer Dialogue in Mobile Environments*. Kluwer Academic Publishers, Dordrecht, The Netherlands. (this volume).

Wahlster, W., Reithinger, N., and Blocher, A. (2001). SmartKom: Multimodal communication with a life-like character. In *Proceedings of European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 1547–1550, Aalborg, Denmark.

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. (2002). DARPA Communicator: Cross-system results for the 2001 evaluation. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, pages 269–272, Denver, Colorado, USA.

Walker, M. A., Kamm, C. A., and Litman, D. J. (2000). Towards developing general models of usability with PARADISE. *Natural Language Engineering, Special Issues on Spoken Dialogue Systems*, 1(1):1–16.

Walker, M. A., Litman, D., Kamm, C. A., and Abella, A. (1997). PARADISE: A general framework for evaluating spoken dialogue agents. In *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-EACL)*, pages 271–280.

Whittaker, S. and Walker, M. (2004). Evaluating dialogue strategies in multimodal dialogue systems. In Minker, W., Bühler, D., and Dybkjær, L., editors, *Spoken Multimodal Human-Computer Dialogue in Mobile Environments*. Kluwer Academic Publishers, Dordrecht, The Netherlands. (this volume).