Chapter 7

SPOKEN DIALOGUE SYSTEMS EVALUATION

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

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

Abstract This chapter first provides a brief introduction to evaluation methods and criteria and then presents two very different spoken dialogue research prototype systems and their evaluation. The first prototype is the non-task-oriented, multimodal Hans Christian Andersen (HCA) system for edutainment, the second prototype is the task-oriented, multimodal SENECA onboard system in the car. The systems were tested with representative users in the laboratory and in the field, respectively. For both systems we describe rationale for the chosen evaluation method, evaluation process, evaluation criteria, and evaluation results.

Keywords Evaluation; Spoken dialogue systems; Methods; Criteria.

1 Introduction

Evaluation is an important part of the software life cycle and is interwoven with development in complex ways. Its function is to provide iterative feedback on the quality of each system component, as well as of the entire system throughout the development process. Evaluation is crucial to ensure, e.g., system correctness, appropriateness, and adequacy.

For spoken dialogue systems (SDSs) and their components and aspects, including speech recognition, natural language understanding, dialogue

L. Dybkjær et al. (eds.), Evaluation of Text and Speech Systems, 187–219.

management, response generation, speech synthesis, system integration, and human factors, there has been extensive work on evaluation as documented in, e.g., the EAGLES guidelines (Gibbon et al., 1997) and the DISC Best Practice Guide (http://www.disc2.dk), both of which provide useful information on technical and usability evaluation, and in the US DARPA Communicator project, which used the (Paradigm for Dialogue System Evaluation) PARADISE framework (Walker et al., 1997) for usability evaluation, see (Dybkjær et al., 2004) for an overview.

However, the SDSs field is rapidly developing in new directions. Task-oriented unimodal systems are becoming increasingly sophisticated and are now often used in mobile environments. Multimodal task-oriented SDSs have become popular, not least in research. Mobile multimodal aspects will be illustrated in Section 4. A new breed of non-task-oriented SDSs is emerging. A multimodal example is the conversational Hans Christian Andersen (HCA) system to be discussed in Section 3. These developments continue to pose new challenges to research into the evaluation of SDSs.

The aim of this chapter is to give the reader some insight into the evaluation of SDSs, which have the challenging characteristics just mentioned. Through presentation of two concrete and very different examples, we try to demonstrate how advanced SDSs could be evaluated in practice and explain what the considerations behind the selection of evaluation methods and criteria have been. Before presenting the examples, we discuss evaluation methods and criteria in order to provide a frame of reference for what follows (Section 2). The first example (Section 3) is from progress evaluation of a non-task-oriented, multimodal SDS for edutainment for use in a stationary environment. The second example (Section 4) is an evaluation of a task-oriented multimodal onboard system in the car. In presenting these examples, our aim is not to give a complete overview of existing evaluation methods and criteria for SDSs but, rather, to focus on actual evaluation practice in research prototyping. Section 5 concludes the chapter.

2 Evaluation Methods and Criteria

According to development best practice, evaluation should be done throughout the software life-cycle. A broad selection of evaluation methods is available. These methods are general and can be used largely independently of whether the SDS is unimodal or multimodal, task-oriented or non-task-oriented, for mobile or non-mobile use, etc. Some methods are mainly applied in particular life cycle phases while others are used throughout. In all cases some model of the system is required. By a system model we understand the current version of the evolving system no matter if it only exists as hand-made drawings or scribbles, is fully implemented, or is anything in between.

2.1 Some Terminology

Before describing some frequently used evaluation methods we present some central terms related to evaluation. *Technical evaluation* concerns technical properties of the entire system, as well as of each of its components. Technical evaluation should be done by developers or by a professional evaluation team through *objective evaluation*, i.e., evaluation, which is as far as possible independent of the personal opinions of the evaluators.

Usability evaluation of a system very often involves representative users. To some extent, the evaluation team may draw upon objective evaluation metrics but a substantial part of usability evaluation is done via *subjective evaluation*, i.e., by judging some property of a system or, less frequently, component, by reference to users' personal opinions.

We are aware that some people consider the term "objective evaluation" imprecise because there is always a human in the loop and therefore one might claim that there is always some amount of subjectivity involved in the evaluation even if only in the choice of the quantitative metrics (not) to apply. Nevertheless, we prefer the term objective to other terms, such as "expert evaluation" or "instrumental evaluation". These terms are more precise but also more narrow which means that we would need several terms to cover what we mean by objective evaluation. For similar reasons, we shall use the term usability evaluation rather than, e.g., "user oriented evaluation". Usability evaluation may be narrow and, e.g., concern only the naturalness of the speech synthesis, or it may be broad and concern many different aspects of the overall system as measured both objectively and subjectively. Usability evaluation need not involve test users but may be done by usability experts, cf. below.

Objective, as well as subjective evaluation can be both quantitative and qualitative. *Quantitative evaluation*, when objective, consists in counting something and producing a meaningful number, percentage, etc. In subjective quantitative evaluation, there are at least two possibilities: non-expert personal opinions are expressed as quantitative scores of some kind, or such opinions are expressed as numbers or percentages, e.g., regarding the perceived number of times the user requested help. *Qualitative evaluation* consists in estimating or judging some property by reference to expert standards and rules or to one's personal opinion.

2.2 Evaluation Methods

When constructing software there is a number of software tests one may consider to use to ensure that the software actually runs robustly and has the specified functionality. The list of such test methods is long and we don't have space to go into details, so we just briefly describe a few frequently used methods, including unit test, integration test, and function tests in terms of blackbox

and glassbox. A *unit test* is applied to a, typically small, system component called a unit, and is written and carried out by the code developers. It is based on test suites, which may be prepared even before the code is written. The test suites are meant to be run again and again (regression test) as the code develops in order to check the functionality of the unit. An *integration test* is a test of whether two or more modules actually work together when integrated. The purpose of a *function test* is to systematically test if system functions work. It is normally carried out on an integrated system, which may be the entire system or some large module, e.g., a natural language understanding module. While a *glassbox test* focuses on the code and tests the internal logic of the system, a *blackbox test* focuses on the system's input/output behaviour to see if it is in accordance with specifications and descriptions in manuals. The code itself is not considered but is viewed as a black box.

All methods just mentioned are closely related to the code development process. Assuming an iterative life-cycle model, their primary use is therefore in the construction phase, and they mainly help evaluate certain technical aspects. There is another set of evaluation methods, which focus on interaction and, which are mainly applicable to other life-cycle phases. There are also many of these. Figure 1 shows some of the most frequently used methods, cf. (Bernsen and Dybkjær, 2007). Data collected with these methods may serve both as a basis for technical evaluation and for usability evaluation.

The methods in Figure 1 apart from walkthrough and guideline-based evaluation – typically involve representative users. Most evaluations with representative users are carried out in the lab and are often *controlled tests* in the sense that users are given precise tasks (scenarios) to carry out. Lab tests with users have much in common as regards how they are prepared and run, no matter which method is used and which development stage the system model is at.

To ensure data reliability, it is important that the users involved in an evaluation are representative of the target user group and that they jointly represent the diversity of properties characterising the target group.

<table>
<tr><th>Development stage</th><th>Early</th><th>Middle</th><th>Late</th></tr>
<tr><td rowspan="9">Evaluation methods</td><td>Walkthrough</td><td></td><td></td></tr>
<tr><td>Low-fidelity prototyping</td><td></td><td></td></tr>
<tr><td colspan="2">Guideline-based evaluation</td><td></td></tr>
<tr><td colspan="2">Wizard of Oz</td><td></td></tr>
<tr><td colspan="3">High-fidelity prototyping</td></tr>
<tr><td colspan="2"></td><td>Field test</td></tr>
<tr><td colspan="3">Think-aloud</td></tr>
<tr><td colspan="3">Interview</td></tr>
<tr><td colspan="3">Questionnaire</td></tr>
</table>

Figure 1. Frequently used interaction evaluation methods.

Low-fidelity prototypes (or *mock-ups*) and *walkthroughs* are mainly used early in the life cycle. These methods only require some preliminary model of the evolving system, such as a paper sketch (mock-up) or a dialogue model graph. Their results provide early input on, e.g., missing system functionality and interaction design problems. While prototype evaluation normally involves users, walkthroughs are often carried out by the development team.

Guideline-based evaluation is mainly used in the early life-cycle phases. Guideline-based evaluation does not require user involvement but does require a system model and a set of guidelines. Some weaknesses of guideline-based evaluation are that detailed and reliable sets of guidelines often do not exist (yet), especially for advanced types of system, multimodal or otherwise, and that focus during guideline application tends to be narrowly fixed on what the guidelines recommend. On the other hand, reliable guidelines reflect good practice and, if available, it makes good sense to try to follow them. Guideline-based evaluation is done by a person, often from the development team, who is able to apply the guidelines. There is a related method called *expert evaluation*, which is carried out by an (external) expert who is familiar with existing practice for the task or domain in question. *Heuristic evaluation* is sometimes taken to mean guideline-based evaluation and is sometimes used synonymously with expert evaluation.

Wizard-of-Oz simulation in which one or more persons act as the system in interaction with users, is useful primarily when a near-complete design of (central parts of) the system model is available to serve as basis for simulation. The method can provide very detailed feedback on the designed functionality and usability. As long as the system as a whole or part(s) of it has not been implemented, design revisions can be made without code (re-)writing.

A *high-fidelity prototype* is a prototype, which has an interface similar to that of the planned final system. High-fidelity prototypes can be used throughout the life cycle since there need not be much functionality behind the interface before the prototype can be used to generate useful information for the development team by letting users interact with it. High-fidelity prototypes are often used in controlled lab tests, i.e., in a controlled environment with invited users who often follow scenarios. In a *field test*, users are not controlled but use the system in, or from, their own environment, e.g., their office, whenever it suits them. A field test is typically used towards the end of development when the system is close to being finalised. It serves to reveal errors and weaknesses that were not detected in previous tests.

The *think-aloud method* is often used in combination with a controlled lab test based on some prototype (low or high-fidelity) with which users interact while they try to phrase their thoughts. Think-aloud may reveal where users have problems due to inadequate interface design.

Questionnaires and *interviews* are often used before and/or after users interact with some model of the system, i.e., questionnaires and interviews may accompany or complement use of any of the above methods involving users. While interviews may capture the users' immediate reactions to a system, questionnaires leave time for the users to organise their thoughts and think about how to express them. Interviews are more time-consuming to use because they require the presence of an interviewer and some post-processing of interview notes. Questionnaires may be made electronically available to minimize the post-processing.

All evaluation methods generate data, which may be captured and used for later analysis and evaluation. Depending on the nature of the data collected, this data may serve as basis for technical evaluation, usability evaluation, or both.

2.3 Evaluation Criteria

Evaluation criteria concerning technical, as well as usability aspects may be established and applied to the collected data with various purposes in mind, such as evaluation of system quality and conformance to specifications, comparison of the system with other systems, or evaluation of progress during system development. The difficult question is exactly which evaluation criteria to apply at any point, not least when it comes to usability evaluation. For instance, in most cases it may be rather straightforward to evaluate efficiency of interaction by measuring time to task completion but, as regards, e.g., user satisfaction there is no simple quantitative measure to rely on.

Technical evaluation, on the other hand, is well developed for several aspects of SDSs and their components. For instance, there is broad agreement on key evaluation criteria for some basic qualities of speech recognisers. These criteria include, e.g., word and sentence error rate, vocabulary coverage, perplexity, and real-time performance, cf., e.g., the DISC dialogue engineering model at http://www.disc2.dk/slds/.

There are ongoing standardisation efforts. For example, the International Telecommunication Union ITU-T SG12 (Study Group 12 on performance and quality of service, http://www.itu.int/ITU-T/studygroups/com12/index.asp) has issued recommendations on how to evaluate spoken dialogue systems and some of their components, cf. ITU-T Rec. P.85 and P.851, and the National Institute of Standards and Technology (NIST) develops evaluation protocols and benchmark tests (http://www.itl.nist.gov/iad/894.01/).

Standards facilitate use of a common set of evaluation criteria or, rather, perhaps, some subset of common criteria, in each specific case. 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.

From the point of view of usability, system differences include, e.g., the fact that the skills and preferences of the target users differ widely from one system to another. This and other parameters, such as application type, task, domain, and use environment, must be taken into account when designing for, and evaluating, usability no matter the technical sophistication of the system.

As experience is being gathered on technical solutions for spoken multimodal systems, it seems that a major part of the focus in research is currently on how to evaluate the usability of these systems, cf., e.g., (Minker et al., 2005a). One reason may be that there are more unknown usability factors than technical factors involved; another, that the novel usability and qualitative evaluation issues raised by this kind of systems are being addressed at an earlier stage than the majority of novel quantitative and technical issues.

3 Evaluation of the NICE Hans Christian Andersen Prototype

Our first evaluation example is the first research prototype of a multimodal non-task-oriented SDS, the NICE Hans Christian Andersen (HCA) SDS, which was developed as part of the European Human Language Technologies NICE project (2002–2005) on Natural Interactive Communication for Edutainment. The evaluated prototype was the first of two prototypes, which we shall call PT1 and PT2, respectively. We first briefly describe the system and then present our evaluation of PT1. Although we describe the actual test set-up, focus in what follows is not on evaluation *process* but on the *methods and criteria*, which were applied, as well as on the test *results* achieved.

3.1 Description of the First HCA Prototype

The main goal of the HCA system is to enable edutaining conversation with 10 to 18 year-old children and teenagers in museums and other public locations. There, users from many different countries are expected to have non-task-oriented conversation with HCA in English for an average duration of, say, 5–15 minutes. In generic terms, the system is a new kind of computer game, which integrates spoken conversation into a professional computer games environment. The user communicates with HCA using spontaneous speech and 2D pointing gesture. 3D animated, embodied HCA communicates with the user through speech, gesture, facial expression, and body movement. In the first prototype, communication takes the form of limited mixed-initiative spoken conversation.

The event-driven, modular architecture of the system is shown in Figure 2 and described in more detail in (Bernsen et al., 2004a). The speech recogniser is hatched in the figure because it was not integrated in the first prototype.

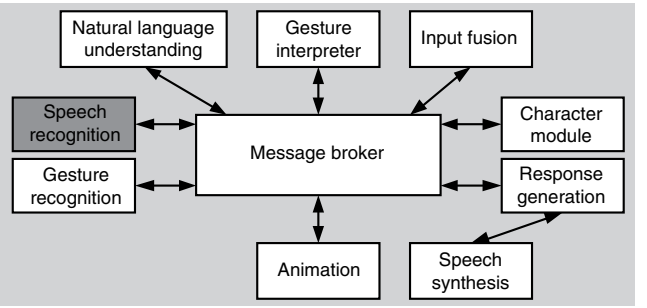

Figure 2. General HCA system architecture.

While most of the modules are self-explanatory, it may be mentioned that the character module is the system's conversation manager and that the message broker manages XML-format message passing between all system modules.

Norwegian Information Security Laboratory (NISLab) has developed HCA's natural language understanding, character module (Bernsen and Dybkjær, 2004a), and response generation (Corradini et al., 2005) components. The other components shown in Figure 2 have been developed by other project partners or are (based on) freeware (gesture recognition, message broker), or off-the-shelf software (speech synthesis from AT&T). The speech recogniser is the commercial SpeechPearl recogniser from Scansoft (now Nuance) trained on non-native English speech from children. For gesture recognition, gesture interpretation, and input fusion, see (Martin et al., 2006). For animation, see (Corradini et al., 2004). The project partners are: TeliaSonera, Sweden, Liquid Media, Sweden, Scansoft, Belgium, and LIMSI/CNRS, France.

HCA's domains of knowledge and discourse include his fairytales, his childhood in Odense, Denmark, his persona and physical presence in his study, getting information about the user, his role as "gatekeeper" for a fairytale games world developed by project partner TeliaSonera and not described here (Gustafson et al., 2004), and the "meta" domain of resolving miscommunication problems. These domains are probably among those which most users would expect anyway. Users meet HCA in his study (Figure 3), which is a rendering of his actual study on display in Copenhagen, modified so that he can walk around freely and with a pair of doors leading into the fairytale world (not visible in the figure). Pictures relating to HCA's knowledge domains have been hung on the walls. The user may point to them and ask questions about them. HCA can tell stories about the pictures and about other objects in his room, such as his travel bag.

Lacking locomotion autonomy in the first prototype, HCA's locomotion is controlled by the user who is also in control of four different virtual camera angles onto his study. In PT1, the animation engine only allowed HCA to display

Figure 3. HCA gesturing in his study.

one movement at a time, which means that he could, e.g., blink but then could not move his mouth at the same time. Basically, this means that he behaves somewhat rigidly because he is quite far from acting like a human being when speaking.

3.2 Evaluation of the First HCA Prototype

The first integrated HCA prototype (PT1) was tested in January 2004 with 18 users from the target user group of 10 to 18-year-olds. In the user test set-up, the recogniser was replaced by a wizard who typed what the user said. The rest of the system was running. The speech recogniser still needed to be trained on 40–50 hours of speech data recorded with mostly non-native English speaking children.

The test users used the system in two different conditions. In the first 20 minutes condition, they had unconstrained conversation with HCA based only on instructions on how to change the virtual camera angles, control HCA's locomotion, and input speech and gesture. This enabled them to become familiar with the system. In the following, second test condition and after a short break, the users spent 20 minutes trying to solve as many problems as possible from a hand-out list, which included 13 problems, such as "Find out if HCA has a preferred fairytale and what it is" and "Tell HCA about games you like or know". The reason for this dual-condition test protocol was to make sure that, in the second test condition, users would put a strong "initiative pressure" upon the system during conversation due to the fact that the users had an agenda of their own, even if, in the first condition, they might tend to obediently follow HCA's lead during conversation.

Each user was interviewed immediately after the second-condition interaction with HCA. The semi-structured interviews were based on a common set of questions. In total, each session had a duration of 60–70 minutes.

Users arrived in parallel, so there were two test rooms, two wizards doing recogniser simulation, and two interviewers. The wizards were trained transcribers who had also been specifically trained for the PT1 test. In one room, the user had a mouse and a touch screen for gesture input, whereas in the other room only a mouse was available as pointing device. In the room with the touch screen, the user could also watch HCA on a 42" flat-panel screen. An observer was present in this room as well. An experimenter took care of guiding and instructing the users and a technician was around for handling any technical problems arising. Everybody involved had a dedicated, printed instruction sheet to follow. In addition, system developers would come and go discretely, discuss their observations, assist the technician when needed, etc.

As we will not return to this topic below, it may be mentioned here that the different set-ups in the two test rooms did produce interesting differences in user behaviour. While we did not observe any differences related to display size, the availability of the mouse in both set-ups made the users click "like crazy" whilst talking to HCA and tended to alienate them to the touch screen alternative, when available. These effects disappeared completely in the PT2 test in which only touch screens could be used for deictic gesture input (Martin et al., 2006). All interactions with PT1 were logged, audio recorded, and video recorded. In total, approximately 11 hours of interaction were recorded on audio, video, and logfile, respectively. In addition, 18 sets of semi-structured interview notes were collected. This data served as our basis for evaluating PT1 according to criteria, which had been agreed early in the project.

3.2.1 Choice of evaluation method. The PT1 evaluation addressed a fully implemented system except that the recogniser was not yet hooked up as explained above. Thus, the evaluation method to look for should be one, which is well-suited for the first integration phase of the software lifecycle in which progress evaluation is not possible because there are no previous system versions to compare with; in which we wish to establish if the prototype performs adequately according to development plans; and, not least, in which we want to get input from representative users. We decided on a controlled laboratory test using a high-fidelity prototype bionic Wizard-of-Oz set-up in which a wizard replaced the speech recogniser.

The reason why we chose to evaluate HCA PT1 with target group users in a controlled laboratory setting rather than in the field, is the following. A field test, e.g., in a museum, is much harder to control than a laboratory test. It is difficult or impossible to: instruct users adequately in the field to ensure a strict dual test condition experimental regime, to interview the users, and to video

record the users in action with all that this entails in terms of informed consent, permission signatures, and rights to use the recorded data for purposes explicitly agreed upon. When conducting costly testing of a first system prototype, it is critically important to collect data, which is optimised for the purpose of obtaining the interaction information, which is most needed in order to judge how the system performs and is perceived by users. Analysis of this data is crucial to the processes of functional extension, re-specification, and redesign, which are normally planned to follow a first-prototype user test. To ensure full control of the corpus collection process, a controlled laboratory test seemed to be the only viable approach in our case.

3.2.2 Requirements to test users. The test plan established early in the project included a number of requirements to test users of the HCA prototypes. These requirements are listed below followed by a comment on how they were met in the evaluation of PT1.

- Each prototype should be evaluated by at least 12 test users. This was a pragmatic estimate of the minimum number of users we would need to obtain a fair amount of speech data and sufficient data to get a reasonably reliable idea of which improvements to make. PT1 was evaluated with 18 users.
- Age. At least eight users should belong to the primary target group. Since the target environment for the system is museums and the like, there may well be users present who do not belong to our primary target group. We would therefore accept up to one-third of test users not belonging to the primary target group, although we would prefer all users to be children and youngsters to collect as many relevant speech data as possible. All 18 users belonged to the primary target group of 10 to 18-year-olds with an age range of 10–18 years and an average age of 14.3 years.
- Both genders should be represented approximately equally. The test group included nine girls and nine boys.
- User background diversification. The user group shows a good spread in computer game literacy, from zero game hours per week to 20+ hours. All users were schoolchildren. We did not collect data on other diversification factors, such as general computer literacy or experience with educational computer programs.
- Language background diversification. The reason for this requirement is that the system is aimed for use in museums, which attract an English-speaking audience from across the world. Only a single user, an 18-year-old Scotsman, was not Danish and had English, the language

of conversation with HCA, as first language. However, a large-scale (approximately 500 users) Wizard-of-Oz study conducted in the field in the summer of 2003 at the HCA Museum included users of 29 different nationalities (Bernsen et al., 2004b). Thus we felt that we already had voluminous data on the differential behaviour in conversation of users with different nationalities and first languages.

3.2.3 Technical evaluation criteria. A set of technical evaluation criteria had been defined for the system as a whole (Table 1), as well as for its individual components (Table 2). The primary purpose of the technical system evaluation criteria was to test if the system has (i) the specified overall technical functionality, and (ii) the technical robustness required for users to interact with it smoothly for usability evaluation to make sense. In all cases, objective measures can be applied. Table 1 lists the technical system evaluation criteria used for both PT1 and PT2. The table indicates for each criterion whether evaluation is quantitative or qualitative and explains what we understand to be measured by the criterion. The evaluation result per criterion is shown and so is an annotated qualitative score value or comment for each PT1 result. "1" is the lowest score, "5" the highest. It should be noted that the score values allocated in Tables 1 and 2 are strongly relative to what could be expected of the first prototype given the development plans. When evaluating the second system prototype (PT2), the scores would be replaced by scores which adequately reflect the quality of the system per aspect evaluated.

The results were derived from the data collected during the user test of PT1. Mainly the test logfiles were used together with technical knowledge about the system, e.g., regarding the number of emotions, which could be expressed in principle.

Table 1. Technical evaluation criteria for the HCA system.

Technical criterion	Explanation	Evaluation	Score 1–5
Technical robustness	Quantitative; how often does the system crash; how often does it produce a bug, which prevents continued interaction (e.g., a loop)	Some crashes and a number of loops, improvement needed	3 acceptable
Handling of out-of-domain input	Qualitative; to which extent does the system react reasonably to out-of-domain input	System has only few reaction possibilities. Further improvement needed	2 acceptable

Real-time performance, spoken part	Quantitative; how long does it usually take to get reaction from the system to spoken input	OK, natural language understanding is fast; recogniser not tested	5 very good
Real-time performance, gesture part	Quantitative; how long does it usually take to get reaction from the system to gesture input	Too slow due to a designed delay of several seconds to wait for possible spoken input. Further improvement needed	3 basic
Barge-in	Is barge-in implemented	No barge-in in PT1	As planned
Number of emotions	Quantitative; how many different emotions can in principle be conveyed	4 basic emotions	4 good
Actual emotion expression	Quantitative; how many different emotions are actually conveyed verbally and non-verbally	1 basic emotion. Much improvement needed, particularly in rendering capabilities: scripts, synchronous non-verbal expressions, speed, amplitude	1 as planned
Number of input modalities	Quantitative; how many input modalities does the system allow	3, i.e., speech, 2D gesture, user key haptics for changing virtual camera angle and making HCA walk around (inconsistent with character autonomy)	As planned
Number of output modalities	Quantitative; how many output modalities does the system allow	Natural interactive speech, facial expression, gesture. More rendering capability needed	As planned
Synchronisation of output	Qualitative; is output properly synchronised	Speech/gesture/facial OK. More rendering capability needed. No lip synchronisation	As planned
Number of domains	Quantitative; how many domains can HCA talk about	6, i.e., HCA's life, fairytales, physical presence, user, gatekeeper, meta	As planned

Table 2. Technical evaluation criteria for the HCA system components.

Technical criterion	Explanation	Score 1–5
Speech recogniser		
Word error rate	No speech recognition in PT1	N/A
Vocabulary coverage	No speech recognition in PT1	N/A
Perplexity	No speech recognition in PT1	N/A
Real-time performance	No speech recognition in PT1	N/A
Gesture recogniser		
Recognition accuracy regarding gesture shape	84% of a total of 542 shapes measured on 9 hours of user test data	As planned
Natural language understanding		
Lexical coverage	66%	Ahead of plan
Parser error rate	16%	Ahead of plan
Topic spotter error rate	Not evaluated for PT1	As planned
Anaphora resolution error rate	Not in PT1	As planned
Gesture interpretation		
Selection of referenced objects error rate	30 (26%) of 117 results were erroneous, measured on 2 hours of user test data	Basic
Input fusion		
Robustness to temporal distortion between input modalities	No semantic fusion in PT1. No fusion of data structures because no waiting function for NLU input when gesture input	Behind plan
Fusion error rate	No semantic fusion in PT1	Behind plan
Cases in which events have not been merged but should have been	No semantic fusion in PT1	Behind plan
Cases in which events have been merged but should not have been	No semantic fusion in PT1	Behind plan
Recognised modality combination error rate	No semantic fusion in PT1	Behind plan

Character module		
Meta-communication facilities	Handling of user input: repeat, low confidence score, insults	As planned
Handling of initiative	Limited free user initiative; no user initiative in mini-dialogues	As planned
Performance of conversation history	Support for meta-communication and mini-dialogues	As planned
Handling of changes in emotion	HCA's emotional state updated for each user input	As planned
Response generation		
Coverage of action set (nonverbal action)	Approximately 300 spoken output templates and 100 primitive non-verbal behaviours	2 acceptable
Graphical rendering (animation)		
Synchronisation with speech output	Works for a single non-verbal element at a time. No lip synchronisation	As planned
Naturalness of animation, facial	Overlapping non-verbal elements missing. Limited number of animations	As planned
Naturalness of animation, gesture	Overlapping non-verbal elements missing. Limited number of animations	As planned
Naturalness of animation, movement	Somewhat rigid HCA walk	As planned
Text-to-speech		
Speech quality	OK	4 good
Intelligibility	Some syllables "swallowed"	4 good
Naturalness	OK	4 good
Non-speech sound		
Appropriateness in context of music/sound to set a mood	Not in PT1	N/A
Integration		
Communication among modules	PT1 is reasonably well-tested with respect to inter-module communication	4 good
Message dispatcher	OK	4/5 good
Processing time per module	Real-time overall, except for gesture modules	5/3 fine/basic

Focus in the NICE project was not on thorough and exhaustive technical evaluation. Rather, the idea has been to keep the technical evaluation of components at a limited though reasonably sufficient level. This is reflected in Table 2, which includes relatively few but important criteria per component. The evaluation results in Table 2 are mainly based on analysis of logfiles from the user test of PT1 and technical knowledge of the components. If nothing else is indicated, the results refer to the user test material as a whole. See also (Martin et al., 2004a; Martin et al., 2004b) concerning gesture recognition and interpretation. The "mini-dialogues" mentioned in the table are small hard-coded dialogue structures plugged into Andersen's conversation management structure at points where he is expected to go in more depth with some topic.

Overall, as suggested by Tables 1 and 2, PT1 conformed reasonably well to the PT1 requirements and design specification. On some points, PT1 functionality and performance was better than planned. For instance, the natural language understanding module had been integrated and could be tested ahead of plan. On other points, PT1 functionality and performance was worse than planned. For instance, input fusion had not been implemented.

3.2.4 Usability evaluation criteria. While the focus on technical evaluation is limited in the NICE project, usability evaluation plays a central role because little is known about the usability aspects of spoken computer games for edutainment. The usability evaluation criteria adopted in the evaluation plan include state-of-the-art criteria, as well as new criteria that we anticipated would be needed and had to be developed in the project itself.

We divided the usability evaluation criteria into two groups. One group includes what we call basic usability criteria (Table 3), i.e., criteria that we consider basic to usability. If one of these criteria produces a strongly negative evaluation result, this may mean that the module(s) responsible must be improved before further evaluation is worthwhile. For example, if speech recognition adequacy is very bad this means that, basically, the user is not able to communicate with the system until recognition has been improved. Technical evaluation measures of, e.g., speech recogniser performance, gesture recogniser performance, and parser performance are objective metrics, which may be compared to perceived subjective recognition and understanding adequacy.

The second group of criteria (Table 4) includes the criteria, which we consider essential to the evaluation of the NICE prototypes. Several of these are new and may need subsequent re-definition in order to serve their purposes.

Most parameters in Tables 3 and 4 must, as indicated in the second column, be evaluated using a subjective method, such as questionnaire or interview. Like Tables 1 and 2, Tables 3 and 4 list the core evaluation criteria to be applied to both PT1 and PT2. The tables explain what we understand to be measured

Table 3. Basic usability evaluation criteria for the HCA system.

Basic usability criterion	Explanation	Evaluation	Score 1–5
Speech understanding adequacy	Subjective; how well does the system understand speech input	Quite well; but fairly often HCA does not answer the user's question but says something irrelevant; vocabulary seems too small	3 acceptable
Gesture understanding adequacy	Subjective; how well does the system understand gesture input	Reaction to gesture input too slow. This perceived slowness is due to a one-second delay set in the system to allow the input fusion module to wait for possible linguistic input following the gesture. This delay will have to be reduced. It would be nice if HCA could tell about more things in his study than a few pictures	3 basic
Combined speech/ gesture understanding adequacy	Subjective; how well does the system understand combined speech/ gesture input	No semantic input fusion in PT1	Behind plan
Output voice quality	Subjective; how intelligible and natural is the system output voice	Mostly OK, intelligible, not unpleasant, modest syllable "swallowing"	4 good
Output phrasing adequacy	Subjective; how adequate are the system's output formulations	Mostly OK, no user remarks	4 good
Animation quality	Subjective; how natural is the animated output	A couple of annoying errors (HCA could walk on the ceiling and in furniture). Basically animation was OK although HCA could be more lively and he walks in a strange way	3 acceptable
Quality of graphics	Subjective; how good is the graphics	Rather good, only a (true) user remark on too dark graphics due to the study light sources	4/5 very good

Table 3 – Continued

Basic usability criterion	Explanation	Evaluation	Score 1–5
Ease of use of input devices	Subjective; how easy are the input devices to use, such as the touch screen	Microphone, mouse, touch screen, keyboard: users generally positive	4/5 very good
Frequency of interaction problems, spoken part	Quantitative; how often does a problem occur related to spoken interaction (e.g., the user is not understood or is misunderstood)	A larger number of bugs, primarily loops, found than expected. A total of 13.3% of the output was found affected by bugs. Non-bugged interaction, on the other hand, showed better performance than expected	Bugged interaction: 2 barely adequate. Non-bugged interaction: 3/4 acceptable
Frequency of interaction problems, gesture part	Quantitative; how often does a problem occur related to gesture interaction	No figures available but mouse-pointing users continued to create a stack problem due to multiple fast mouse clicks causing a number of interaction problems. By contrast, the touch screen users emulated human 3D pointing during conversation	3 basic
Frequency of interaction problems, graphics rendering part	Quantitative; how often does a problem occur related to graphics	Two serious generic bugs found: most users got lost in space outside HCA's study at least once, HCA sometimes got immersed in furniture	2 barely adequate
Sufficiency of domain coverage	Subjective; how well does the system cover the domains it announces to the user	HCA does not have enough answers to questions; there is enough about fairytales but not about his life	3/4 acceptable
Number of objects users interacted with through gesture	Quantitative; serves to check to which extent the possibilities offered by the system are also used by users	21 pointable objects in HCA's study: in general, users pointed to most of them and to many more as well	3 acceptable
Number of topics addressed in conversation	Quantitative; serves to check how well the implemented domains cover the topics addressed by users	All generic topics (approx. 30) addressed; some topic details (approx. 10) addressed but not covered	As expected

Table 4. Core usability evaluation criteria for the HCA system.

Core usability criterion	Explanation	Evaluation	Score 1–5
Conversation success	Quantitative; how often is an exchange between the user and the system successful in the discourse context	Most users pointed out that HCA's responses were sometimes irrelevant. Work on quantitative metrics in progress	3 acceptable
Naturalness of user speech and gesture	Subjective; how natural is it to communicate in the available modalities	Very positive user comments overall although some users said they had to get used to talking to a computer	4/5 very good
Output behaviour naturalness	Subjective; character believability, coordination and synchronisation of verbal and non-verbal behaviour, display of emotions, dialogue initiative and flow, etc.	Very complex criterion, hard to score. Still, users were surprisingly positive, not least about HCA's physical appearance	3/4 quite acceptable
Sufficiency of the system's reasoning capabilities	Subjective; how good is the system at reasoning about user input	No reasoning available in PT1. Needs identified for reasoning about implications of user input	As planned
Ease of use of the game	Subjective; how easy is it for the user to find out what to do and how to interact	Reasonably easy. Main difficulties due to HCA's limited understanding abilities and users' limited English, e.g., few knew the names of HCA's fairytales in English	3 acceptable
Error handling adequacy, spoken part	Subjective; how good is the system at detecting errors relating to spoken input and how well does it handle them	Limited. Main complaint is that HCA often does not answer the user's questions but keeps talking about whatever he is talking about or says something irrelevant	2 acceptable
Error handling adequacy, gesture part	Subjective; how good is the system at detecting and handling errors relating to gesture input	No error handling involving gesture	Behind plan

Table 4 – Continued

Core usability criterion	Explanation	Evaluation	Score 1–5
Scope of user modelling	Subjective; to which extent does the system exploit what it learns about the user	No user comments. User age, gender and nationality collected; age information used once in an HCA question	As planned
Entertainment value	Subjective; this measure includes game quality and originality, interest taken in the game, feeling like playing again, time spent playing, user game initiative, etc.	User test very positive	4 good
Educational value	Subjective; to which extent did the user learn from interacting with the system	User test very positive. As (user) self-assessment is occasionally misleading, we might have added some recall questions relating to what HCA told each particular user	4 good
User satisfaction	Subjective; how satisfied is the user with the system	User test very positive	4 good

by each criterion. The evaluation result per criterion is shown and so is an annotated score value or comment for each of the PT1 results. Each allocated score relates to what we believe should be expected of PT1. For instance, in PT1, one is entitled to expect a better approximation to real-time performance than to perfect handling of non-task-oriented conversation, the latter being one of the main research challenges in the HCA project. If real-time performance is a serious problem in PT1, we may have an unpleasant and unexpected problem on our hands at this stage, whereas if conversation management is not perfect in PT1, this is only what everyone would be entitled to expect. More generally speaking, and with the exception of real-time performance, we consider a score of "3" for all main challenges addressed in the HCA system clearly adequate at this stage of development. Still, we need to stress the judgmental nature of many of the scores assigned in Tables 3 and 4. The results are mainly based on our interpretation of the data collected in the interviews during the user test of

PT1. Bernsen and Dybkjær (2004b) provide details about the questions asked to users during the interviews and users' answers to each of the questions. In Table 4, "Scope of user modelling" refers to HCA's collection of information provided by the user, which he then makes use of later on during conversation.

3.2.5 Conclusion on PT1 evaluation. Despite its shortcomings, not least in its capability to conduct human-style spoken conversation, PT1 was received remarkably well by the target users in the test. Note, however, that it is common to find some amount of uncritical positive user bias in tests of new technology with exiting perspectives, especially when users have never interacted with the technology before. We would have preferred a smaller number of bugs than was actually found with respect to (a) spoken interaction and (b) the workings of the rendering when users made HCA do locomotion in his study. The system clearly does have edutainment potential, which serves to validate its underlying theory of non-task-oriented conversation for edutainment (Bernsen and Dybkjær, 2004b). In view of these evaluation results, we decided to focus work on PT2 on basic improvements in the system's spoken and non-verbal conversational abilities. The – now completed – PT2 features a completely redesigned conversation manager for full mixed-initiative non-task-oriented conversation, ontology-based natural language understanding and conversation management, speech recognition, a better speech synthesis, and multiple synchronous non-verbal output streams, including audiovisual speech and facially expressed emotion reflecting the character's current emotional state.

4 Evaluation of the SENECA Prototype

Our second evaluation example is of a task-oriented multimodal SDS for a wide range of entertainment, navigation, and communication applications in mobile environments. The research has been carried out in the EU Esprit and Human Language Technologies (HLT) project SENECA (1998–2002) on Speech control modules for Entertainment, Navigation and communication Equipment in CArs (Gärtner et al., 2001). Project partners were: The Bosch Group, Germany; Daimler-Chrysler, Germany; Daimler Benz Aerospace, Germany; Motorola Germany; Motorola Semiconductor, Israel; Centro Ricerche Fiat, Italy; and Renault Recherche Innovation, France.

The goal of the project was to integrate and further develop SDS technology for use in cars up to an almost commercial level. The usability of the SENECA prototype system has been evaluated in different development cycles by means of user tests collecting objective and subjective data. With speech input, road safety, especially for complex tasks, is significantly improved. Both objectively and as perceived by the driver, humans are less distracted from driving when using speech input for on-board devices than if using manual input as in

standard remote-controlled navigation systems (Gärtner et al., 2001; Green, 2000). In the following, focus is on description of the *evaluation set-up*, the *usability criteria*, and some evaluation *results*.

4.1 Description of the SENECA Prototype System

A large variety of electronic systems are now available in the car for comfort, ease of driving, entertainment, and communications. Some of these systems, notably for navigation and entertainment, require rather complex human-computer interaction, which increases the risk of driver distraction.

The SENECA prototype system whose architecture is shown in Figure 4, represents a step towards close-to-the-market technologies enabling drivers to interact with on-board systems and services in an easy, risk-free way. The head unit of the system – for the driver and front passenger, the COMAND head unit represents the central operating and display unit for numerous functions and devices – is linked via an optical D2B (Domestic Digital Bus) to the GSM module, the CD changer and the (Digital Signal Processing) (DSP) module. The latter contains the signal and D2B communication software. A notebook computer contains the SENECA SDS software, i.e., all the modules that are subject to research, including speech recognition, dialogue management, and access to different databases. DSP and notebook computer are connected via a serial link. In the context of a research project, such a partial software solution conveniently enables optimisation and evaluation. The SENECA SDS prototypes developed for French, German, and Italian languages allow control of entertainment (i.e., radio), navigation, and communication (i.e., telephone) equipment using command and control dialogues combined with speech and

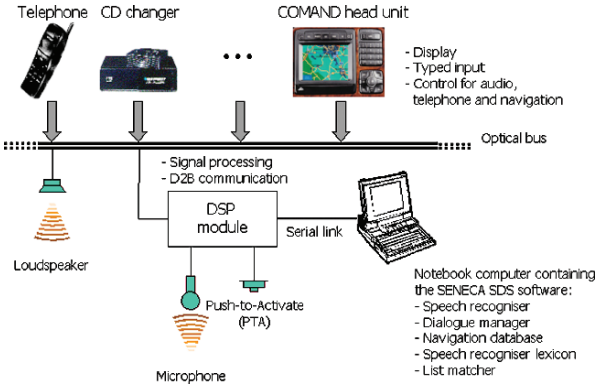

Figure 4. Architecture of the SENECA prototype system.

text output. A more detailed description of the SENECA SDS prototype system for the German language can be found in (Minker et al., 2005b). In the following, we focus on the German prototype.

4.2 Evaluation of the SENECA Prototype System

The German SENECA prototype has been evaluated early and late in the development process, namely at the *concept demonstrator* and *system demonstrator* stages. In both tests, based on an identical evaluation set-up, system operation using speech input and output was compared to functionally equivalent haptic input operation using the COMAND head unit so as to be able to demonstrate the quality of the respective input modalities and their impact on driving safety and quality.

Since industry is primarily interested in high usability and acceptance rates of their future products and since the SENECA project was intended to develop close-to-the-market prototypes, usability field tests have been carried out with mainly core criteria defined on the basis of the project's needs, including quantitative and qualitative elements. These field tests will be described in the following sections.

4.2.1 Evaluation set-up and criteria. The basic components of the experimental set-up employed for the in-field usability evaluation of the German concept and system demonstrators include a passenger car, the standard COMAND head-unit, a second set of pedals, the SENECA prototype, an additional notebook computer with event keys and clock, a video system (three cameras, multiplexer, and digital video recorder), and a set of microphones. Three seats were taken by the driver, a professional driving assessor and the principal investigator, respectively, and one seat was used for the testing equipment.

The notebook computer and the video-recording system were installed in a rack on one of the back seats (Figure 5). Three softkeys of the notebook computer were used for time event recording. A logfile was created for each test person. The notebook computer's display with the system's clock was combined with the camera recordings in order to obtain a time reference for processing the video data.

The video system was connected to three cameras, which recorded the forward traffic scene, the driver's face, and the COMAND head unit display, respectively (Figure 6). The signals from the cameras and the display image of the additional notebook computer were mixed and digitally recorded by a Digital Video Walkman. In addition to the SENECA SDS prototype microphone,

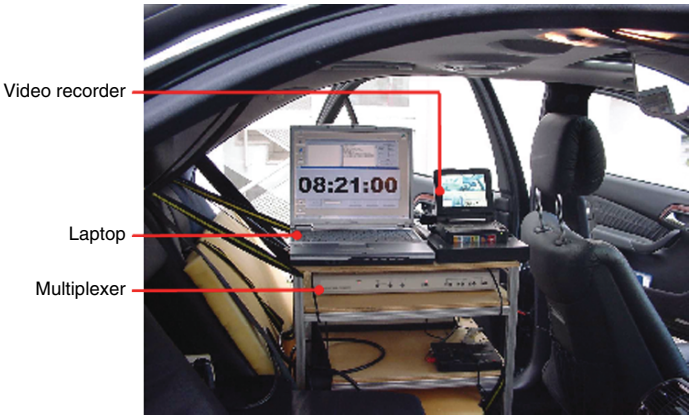

Figure 5. SENECA test car equipment (Mutschler and Baum, 2001).

Figure 6. Camera and microphone positions in the test car (Mutschler and Baum, 2001).

a microphone was installed under the car roof to capture comments made by the passengers, especially the driver. The comments were recorded synchronously with the video data.

The driving assessor noted driving errors and rated the test person's driving skills using a specially designed recording device. In dangerous situations, the driving assessor warned the test person or even intervened by using the second set of pedals or by taking hold of the steering wheel. This seemed particularly important since users had to perform the manual input tasks whilst driving, which is normally prohibited.

The principal investigator was responsible for the organisation and monitoring of the experimental set-up. He announced the tasks for the test person

to perform. He also recorded the start and end of the task, as well as special unforeseen events by pressing predefined keys on the notebook computer. Two driving assessors and two experimenters were employed.

The evaluation trials took place on a 46 km long course near a middle-sized city in Germany. The course was composed of express roadways, highways, entrances/exits, and streets across cities and villages. The test persons drove the course twice to cover both experimental conditions (speech mode/manual mode). Prior to the first experimental trial, a separate training phase was performed on a partly different course while the test person was practicing the handling of the car and several exemplary tasks in the respective input modalities. A set of nine tasks that were representative of manipulating entertainment, navigation and communication equipment was carried out in both experimental conditions, including destination entry (city and street names), activating stored destinations, telephone number dialling, activating stored numbers, etc. The tasks required typical operating actions, namely activating a main function or sub-function, selecting an item from a list, as well as spelling characters. The tasks had to be performed on pre-defined route segments. Assistance was given if necessary and noted by the experimenter. If a task was not completed at pre-defined points, it was declared aborted.

The entire trial sequence comprised written and verbal instructions, a pre-experimental questionnaire, training of SENECA/COMAND operations before the test trial, the test trial on the course, post-trial interviews, and a post-experimental questionnaire. A complete trial took about 3 hours. The Task Completion Rate (TCR) was calculated from the logfile data containing all interesting event times, as well as the speech recogniser transcripts. A task was completed when correctly and completely solved within a scheduled segment without any essential assistance. The duration of the different operations required for task completion was documented by the experimenter. The driving assessor recorded the driving errors and judged the driving quality on a six-point scale.

The driving errors were categorised into the following seven main categories: too low speed, too high speed, too low longitudinal and lateral distances to other cars, inexact lane keeping, insufficient observation of the traffic scene, no indication of change of driving direction, and sudden/late braking. The corresponding times of the driving errors made were recorded automatically. Glances to the display, speedometer/steering wheel, rear mirror, aside (including the outside mirrors) were extracted from the video data.

In a subjective evaluation, the test persons were asked about driving safety, system handling when alternatively using speech and haptic input, appropriateness of the command and control vocabulary, as well as their acceptance and preference of input modalities. The test persons had to compare the speech-based interface to other extra car devices with respect to driving safety and

driver comfort on a six-point scale. The difficulty with the performed field tests relies in the fact that the tasks were not counterbalanced in any way, which makes the interpretation of results somewhat more difficult (e.g., are variations due to task, or to sequence effects). Another difficulty is the time and resource-consuming, and hence rather expensive, evaluation set-up. Therefore, the number of test persons had to be limited. In the German system demonstrator evaluation thus only 16 test persons participated.

4.2.2 Concept demonstrator evaluation. As mentioned, the SENECA prototype has been evaluated at different stages of the project. Usability tests with the *concept demonstrator* version of the prototype allowed to identify implementation-related problems and conceptual weaknesses of the system early in the development process. Such proper diagnostics of system shortcomings can be fed back into system improvement. For instance, the dialogue was sometimes interrupted by the system itself without any traceable reason. Conceptually, test persons found it distracting to have to recall the – not always natural and intuitive – command and control vocabulary. They also criticised the sometimes short vocal prompts and the lack of spoken feedback. In terms of dialogue flow management, the system frequently required users to remember previous system manipulations in order to confirm their input. The system-driven dialogue strategy thus seemed to demand a considerable mental load of the user. Finally, the incoherent dialogue flow strategies between applications (navigation, telephone, and radio) were also judged as inconvenient by the users.

For the evaluation of the final *system demonstrator*, the project partners took the concept demonstrator evaluation results into account in order to improve the system. Also as an outcome of the concept demonstrator evaluation, a long-term user evaluation has been suggested. Analyses of the subjective evaluation data from the interviews have shown that participants would need more time to memorise commands than the 3 hours scheduled for the test trial. In a long-term evaluation, it would be possible to analyse a wider range of functions and evaluate the system, its recognition capabilities, and driver satisfaction in a more reliable way. Such a long-term user evaluation has been performed by a number of selected test persons but cannot be reported here for confidentiality reasons.

4.2.3 System demonstrator evaluation. The system demonstrator evaluation was carried out by the end of the SENECA SDS prototype development cycle. With a technically improved prototype system, quantitative results including TCR and driving error measures, as well as subjective usability evaluations were in focus.

4.2.4 Quantitative usability evaluation results. Most of the tasks showed a markedly lower TCR with speech input compared to haptic input. Averaged over all tasks, the TCR was 79% for speech input and 90% for haptic input. Incomplete speech tasks mostly occurred due to forgotten commands.

In terms of input times, speech input took 63 seconds compared to haptic input requiring 84 seconds (averaged over all tasks). In terms of input accuracy, the most important user errors with speech were vocabulary errors (e.g., wrong commands). Some input spelling errors occurred as well. Errors also occurred when users did not follow the pre-defined dialogue flow, such as when a user, instead of only confirming the ability to provide additional information with a yes/no reply, provided this information directly. Push-to-activate (PTA) errors, i.e., missing or inappropriate PTA activation, were rather sparse, which may be due to learning effects. All these user-induced errors increased the input time and reduced the TCR. However, the long-term evaluations of the SENECA SDS with selected users in France have shown the impact of learning effects: for experienced users using speech input, the TCR attained almost 100% for the navigation task. This task, requiring input of city and street names, may incite the most complex user operations for in-vehicle infotainment systems. In general, with speech input there has been a higher score for driving skills and there are fewer driving errors as compared to manual input, particularly for destination input tasks.

The evaluations in Italy have shown that the SENECA SDS reduces differences between different age groups. Comparison has shown that it is only for young users that the haptic input modality proved to be more efficient. This is probably due to the fact that youngsters have more experience in using technical devices and to their higher dexterity in interacting with them.

4.2.5 Subjective usability evaluation results. Safety and comfort of SENECA compared with other car devices were individually assessed by the test persons on a six-point scale using questionnaires. SENECA was estimated to provide almost a top level of safety and comfort.

Concerning the safety implications, speech input was judged to be below well-established devices, such as ESP (Electronic Stability Program) and the multifunction steering wheel, and to be in the range of automatic transmission. In terms of comfort, speech input was judged to be below air conditioning and in the range of automatic transmission and the multifunction steering wheel. In the Italian2 evaluation, participants were asked if the SENECA SDS was above, below, or equivalent to their expectations. Of the test persons, 47% judged the system to be equal to or above their expectations (compared to 7% for the haptic interface).

To get a general idea of user satisfaction, the test persons were asked what they like about using SENECA. Six test persons assessed speech input to be simple and comfortable. Three test persons mentioned the hands-free/eyes-free aspect. Other positive statements concerned ease of learning, little prior knowledge necessary, speech output, and the good readability of text output.

The advantage of speech input is reflected in the answers to the question about the subjective feeling of driving safety and comfort: 11 test persons mentioned "Higher distraction and less attention to traffic/higher concentration required when using haptic input". In addition to these basic disadvantages, many design aspects of the haptic system were criticised. To the question "What did you dislike about speech input?" only three test persons noted the speech misrecognitions.

Table 5 summarises the most important quantitative and subjective usability criteria and evaluation results for the SENECA prototype systems.

Table 5. Basic and core usability evaluation criteria for the SENECA prototype system.

Usability criterion	Explanation	Evaluation	Score
User satisfaction	Subjective; how satisfied is the user with the system in general	User satisfaction is quite high with the speech-based interface	Good
Task completion rate (TCR)	Quantitative; measure of the success in accomplishing the task in a given time window in either input modality (speech and haptic)	Most of the tasks showed a markedly lower TCR with speech input compared to haptic input	79% averaged over all tasks for speech input, 90% for haptic input
Input times	Quantitative; the duration of completing the task in either input mode (speech and haptic) was measured by the experimenter	Speech input required less time on average than haptic input	63 seconds averaged over all tasks for speech input, 84 seconds for haptic input
Accuracy of inputs	Includes vocabulary errors, orientation errors, spelling errors, open microphone errors	Not explicitly calculated, but affects task completion rate and input times	N/A

Driving performance	Quantitative; a professional driving assessor noted the number of driving errors and rated the test person's driving skills	Significantly less driving errors occurred when using speech input, notably in the categories *inexact lane keeping* and *speed too low*	Number (frequency) of errors averaged across subjects: Inexact lane keeping: 6.6 for speech input; 13.9 for haptic input; speed too low: 2.9 for speech input, 5.9 for haptic input
Glance analysis	Quantitative; glances to the display, speedometer/steering wheel, rear mirror, and aside were extracted from the video data and counted by the experimenter	For low-complexity tasks, the total number of display glances per task for speech and manual input are equal. For complex tasks, speech input required less short and long glances than haptic input	Cannot be reported for confidentiality reasons
Acceptance and preference of input and output modes	Subjective; test persons were asked what they (dis-)liked with speech/haptic input; assessed on a six-point scale	Test persons preferred speech input. They estimated speech output to be very helpful	Cannot be reported for confidentiality reasons
Driving comfort and safety	Subjective; the feeling of the test persons of being comfortable or distracted when manipulating the device using either input modality; assessed on a six-point scale	Test persons felt more comfortable and less distracted when using SENECA compared to using the haptic system	Cannot be reported for confidentiality reasons
Comparison with other car devices	Subjective; how is the speech-based interface judged with respect to other infotainment and entertainment systems in the car in terms of comfort and safety	SENECA was estimated to provide almost top level of safety and comfort	5 on a six-point scale with 1 being the lowest score

4.3 Conclusion on the Evaluation of the SENECA Prototype System

The scientific community is well aware that speech in cars is the enabling technology for interactively and selectively bringing news and information to mobile environments without causing a safety hazard. We have presented the SENECA spoken language dialogue system demonstrator. It provides speech-based access to entertainment, navigation and communication applications in mobile environments. The SENECA SDS demonstrators have been evaluated by real users in the field. The results show that the TCR is higher with haptic input and that with speech input, road safety, especially in the case of complex tasks, is significantly improved. The SENECA project consortium consisted mainly of industry partners that aimed at developing a close-to-the-market prototype. The SENECA evaluations focused on safety, usability, user acceptance, and potential market of the future product. Therefore, technical evaluation criteria have been judged less important. The SENECA evaluations have also demonstrated the rather time and resource-consuming experimental set-up that may be required for specific application domains, such as mobile environments. Given these important constraints, only a limited number of test persons could be recruited for the evaluation, which may have influenced the statistical significance of the results.

5 Conclusion

Having initially described the large variety of established methods available to testers and evaluators of SDSs, this chapter proceeded to illustrate the "real life" on the shop floor of testers and evaluators of today's advanced multimodal SDSs. Our first example described progress evaluation of a forefront research, first prototype embodied conversational agent system, using a controlled laboratory test. The second example described selected aspects of two successive, controlled infield evaluations of a close-to-exploitation task-oriented SDS for in-car use. Depending on how one wishes to count SDS system generations, the two systems are two or three generations apart. The SENECA system uses a command-and-control vocabulary, the dialogue is thoroughly system-driven, and the system is used to solve particular, rather well-circumscribed tasks. The HCA system uses fully spontaneous spoken input, aims – however imperfectly – to enable fully mixed-initiative conversation, and does not enable the user to solve a task at all. The SENECA system was tested in its real environment albeit in a controlled setting, whereas, for the HCA system, developers had to use laboratory testing for testing their first, even incomplete, prototype. Even some of the SENECA findings were predictable from the literature, such as that users were likely to have difficulty learning the command keywords and phrases required for operating the system (Bernsen et al., 1998).

Still, the SENECA results reported are not entirely unambiguous. Speech input is still more error-prone than typing and it is perfectly understandable that the SENECA consortium chose to subject the system to a further, long-term user trial in which they have investigated the long-term learning effects on users who had to keep using the command keywords and phrases until they, perhaps, did manage to control the system without problems of memorising what to say to it. Thus, even with relatively familiar technology and carefully prepared user trials, it is quite possible to have to conclude that the trials made were inconclusive due to their design and that new trials are required.

The differences in the challenges addressed by the two systems are also apparent in the evaluation criteria applied. For SENECA, it was possible to apply, from the start, a series of familiar evaluation criteria, including those applied in order to evaluate various aspects of comparison between using remote-control typed input spelling for navigation and using spoken input, respectively. For the HCA system, on the other hand, we were partly groping in the dark. For instance, how does one grade conformance to specification of a first prototype when the specification was always focused on the *second* prototype and when too little was known in advance about what the first prototype could be made to do? Conversely, not being able to grade with relative exactitude conformance to specification of the first prototype runs the risk of failing to develop the second prototype to specification. Maybe we should have tried to specify both prototypes equally precisely. In another example, how can we evaluate one of the key properties of the HCA system, i.e., conversation success, when too little is known about what conversation success is (Traum et al., 2004)? The closest analogy to conversation success in the SENECA system is TCR but this is of little help since there is no task in the HCA system.

By way of conclusion, we may claim to have illustrated, at least, that (i) even if the choice of evaluation methodologies is often reasonably straightforward even for the evaluation of advanced state-of-the-art systems and components, art, craft skills, and even luck are still required for optimising the choice of evaluation criteria needed for assessment in the light of the data gathered; (ii) new generations of SDSs are likely to keep us occupied for a long time to come in order to invent, apply, discard, revise, and iteratively tune new evaluation criteria in order to optimise our – always costly – evaluations.

Acknowledgements The NICE project was supported by the European Commission's HLT under Grant IST-2001-35293. The SENECA project was supported by the European Commission's 4th framework ESPRIT programme "System Integration and Applications" under HLT (Contract number ESPRIT 26–981). The support is gratefully acknowledged. The authors want to thank all partners and colleagues involved in the projects.

References

Bernsen, N. O., Charfuelán, M., Corradini, A., Dybkjær, L., Hansen, T., Kiilerich, S., Kolodnytsky, M., Kupkin, D., and Mehta, M. (2004a). First Prototype of Conversational H.C. Andersen. In Costabile, M., editor, *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI)*, pages 458–461, Association for Computing Machinery (ACM), New York, USA.

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

Bernsen, N. O. and Dybkjær, L. (2004a). Domain-Oriented Conversation with H. C. Andersen. In *Proceedings of the Tutorial and Research Workshop on Affective Dialogue Systems (ADS)*, volume 3068 of *Lecture Notes in Artificial Intelligence*, pages 142–153, Springer Verlag, Heidelberg, Germany.

Bernsen, N. O. and Dybkjær, L. (2004b). Evaluation of Spoken Multimodal Conversation. In *Proceedings of the Sixth International Conference on Multimodal Interfaces (ICMI)*, pages 38–45, Association for Computing Machinery (ACM), New York, USA.

Bernsen, N. O. and Dybkjær, L. (2007). *Multimodal Usability*. To appear.

Bernsen, N. O., Dybkjær, L., and Kiilerich, S. (2004b). Evaluating Conversation with Hans Christian Andersen. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, volume 3, pages 1011–1014, European Language Resources Association (ELRA), Paris, France.

Corradini, A., Bernsen, N. O., Fredriksson, M., Johanneson, L., Königsmann, J., and Mehta, M. (2004). Towards Believable Behavior Generation for Embodied Conversational Agents. In *International Conference on Computational Science (ICCS), Workshop on Interactive Visualisation and Interaction Technologies (IV&IT)*, volume 3038 of *Lecture Notes in Computer Science*, pages 946–953, Springer Verlag, Heidelberg, Germany.

Corradini, A., Mehta, M., Bernsen, N. O., and Charfuelán, M. (2005). Animating an Interactive Conversational Character for an Educational Game System. In Riedl, J., Jameson, A., Billsus, D., and Lau, T., editors, *Proceedings of the 2005 International Conference on Intelligent User Interfaces (IUI)*, pages 183–190, ACM Press, New York, USA.

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.

Gärtner, U., König, W., and Wittig, T. (2001). Evaluation of Manual vs. Speech Input When Using a Driver Information System in Real Traffic. In *Online Proceedings of International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle*

Design, Aspen, USA. http://ppc.uiowa.edu/Driving-Assessment/2001/Summaries/Downloads/download.html

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

Green, P. (2000). Crashes Induced by Driver Information Systems and What Can Be Done to Reduce Them. In *Proceedings of the Convergence Conference*, pages 26–36, Society of Automotive Engineers, Warrendale, USA.

Gustafson, J., Bell, L., Boye, J., Lindström, A., and Wiren, M. (2004). The NICE Fairy-tale Game System. In *Proceedings of the Fifth SIGdial Workshop on Discourse and Dialogue*, pages 23–26, Association for Computational Linguistics, Boston, USA.

Martin, J.-C., Buisine, S., and Abrilian, S. (2004a). Requirements and Design Specification for Gesture and Input Fusion in PT2 HCA Study. NICE Project Deliverable D1.1-2a Part 2, LIMSI-CNRS, Paris, France.

Martin, J.-C., Buisine, S., Pitel, G., and Bernsen, N. O. (2006). Fusion of Children's Speech and 2D Gestures when Conversing with 3D Characters. *Multimodal Human-Computer Interfaces. Special Issue of Signal Processing*, 86(12):3596–3624.

Martin, J.-C., Pitel, G., Buisine, S., and Bernsen, N. O. (2004b). Gesture Interpretation Module. EU HLT NICE Project Deliverable D3.4-2, LIMSI-CNRS, Paris, France.

Minker, W., Bühler, D., and Dybkjær, L., editors (2005a). *Spoken Multimodal Human-Computer Dialogue in Mobile Environments*, volume 28 of *Text, Speech and Language Technology*, Springer.

Minker, W., Haiber, U., Heisterkamp, P., and Scheible, S. (2005b). 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*, volume 28 of *Text, Speech and Language Technology*, pages 287–310, Springer.

Mutschler, H. and Baum, W. (2001). Evaluation of the System Demonstrator - German Results. Final SENECA Project Deliverable, Robert Bosch, Hildesheim.

Traum, D., Robinson, S., and Stephan, J. (2004). Evaluation of Multi-party Virtual Reality Dialogue Interaction. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, pages 1699–1702, Lisbon, Portugal.

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, Madrid, Spain.