Usability Evaluation of Spoken Dialogue Systems

Lars Bo Larsen

SMC - Speech and Multimedia Communication,
Department of Communication Technology[1]
Aalborg University, Denmark.
Email: lbl@kom.aau.dk

Abstract

This paper concerns the methodologies currently applied to evaluation of spoken and multi modal dialogue systems. Usability is concerned with the effectiveness, efficiency and satisfaction by which users achieves their goals when using a system. The methods developed and applied by the speech community over the past decade to measure the usability of Spoken Dialogue Systems (SDS) are discussed and critically reviewed. The paper starts by giving a general review and discussion of the current issues and problems of establishing SDS usability. This is supported by the presentation and analysis of a case, where it is shown that e.g. learnability can be derived from simple performance measures, such as duration and turn-taking. Results of applying the PARADISE method are presented and discussed.

1. Formerly the Center for PersonKommunikation - CPK. CPK is now fully integrated into the Department of Communication Technology URL: http://cpk.auc.dk/staff/staff.html?lbl

1. Introduction

Speech technology has more than once been predicted to be on the threshold of a "major commercial breakthrough" and many analysts and professionals have believed speech to be the basis for the "next generation" human computer interface. For example, nearly twenty years ago, Jakob Nielsen conducted a study:

> *"Voice interfaces do have a way of capturing the imagination, however. In 1986, I asked a group of 57 computer professionals to predict the biggest change in user interfaces by the year 2000. The top answer was speech I/O, which got twice as many votes as graphical user interfaces."* (Nielsen 2003)

So why hasn't speech technology achieved this status?

One obvious answer has for many years been that, in particular speech recognition, turned out to be a much harder problem than imagined. However, speech recognition has reached an impressive level during the last decade, but the overall system performance is apparently still not sufficiently high for speech driven systems to be generally accepted. Another plausible explanation is that spoken interaction simply isn't competitive in terms of functionality, speed, convenience, privacy, etc. Hugh Cameron (Cameron 2000) analysed the success and failure of a large number of commercial speech systems deployed in the U.S. over the last decade and concluded that people will use speech when:

- *they are offered no choice*
- *it corresponds to the privacy of their surroundings*
- *their hands or eyes are busy on another task*
- *it's quicker than any alternative*

The first three reasons relate in varying degrees to external constraints on the user. The last reason is obviously "the best one", seen from a speech service developer's viewpoint. Unfortunately, Cameron concludes that it has rarely been used (so far). One possible explanation is that the **usability** of speech based systems was not taken (sufficiently) into account when designing the systems. Instead, the designers might very likely have focused more on the performance of the individual components (such as the speech recogniser and the integration into the existing services).

Indeed, when investigating the Best Practises of Spoken Language Dialogue Systems (*SLDS*) in the DISC projects (Dybkjær and Bernsen 2000), Dybkjær and Bernsen observe that:

> *"Far less resources have been invested in human factors for SLDSs than in SLDS component technologies. There has been surprisingly little research in important user-related issues, such as user reactions to SLDSs in the field, users' linguistic behavior, or the main factors which determine overall user satisfaction."*

Despite a growing attention to the importance of systematic methods for the evaluation of the usability of (SDS), general methods have so far been little investigated and few schemes have so far been proposed. One reason for this could well be the fact pointed out by Bernsen and Dybkjær above, namely that the usability of voice-driven services is still poorly understood due to the fact that it has been relatively little researched compared to the component technologies such as speech recognition and -synthesis.

The aim of this paper is therefore to analyse how the usability of speech systems currently is evaluated in order to set focus on the applied methods' strengths and weaknesses. Recent evaluation schemes, such as PARADISE proposed by Walker and colleagues from AT&T (Walker et al. 1998)will be in discussed. PARADISE has been used in a number of evaluations, e.g. the recent DARPA Communicator project (Walker 2004) and is an undertaking to create a standardised paradigm for SDS evaluation, which can be used to compare the performance of dialogues across different domains.

However, before discussing how to obtain and analyse measures of usability it is necessary to define more precisely what usability is and how it is measured.

2. Definition(s) of Usability

There are many different definitions of usability. However, almost all refer to the three key concepts defined in the ISO 9241 Standard (ISO 1998).

2.1. ISO and ETSI definitions of Usability

> *Usability: The effectiveness, efficiency, and satisfaction with which specified users achieve specified goals in particular environments.*

Effectiveness is the accuracy and completeness with which users can obtain their goals. Efficiency can be defined as the costs of obtaining these goals. Satisfaction relates to the comfort and acceptability of the users. So, in relation to the discussion about objective and subjective measures, effectiveness and efficiency are clearly related to objective (often referred to as performance measures), whereas satisfaction is a subjective measure. This definition is supported by ETSI (the European Telecommunications Standards Institute) adopts this view and also points out that usability, together with the costs and benefits for the user, form the concept of utility:

> *"**Usability** is considered as a pure ergonomic concept not depending on costs of providing the system. Usability together with the balance between the benefit for the user and the financial costs form the concept of **Utility**."* (ETSI 1993)

ETSI elaborates on what is termed "measures of usability". These are sharply divided into performance, or objective measures and attitude, or subjective measures. ETSI claims that this distinction is orthogonal, i.e. independent of each other. However, ETSI acknowledges that a dependency through intermediate measures such as consistency and redundancy, as well as sharing a common set of physical characteristics can exist.

The complementary roles of objective and subjective measures also leads to the fact that usability can only be established through the simultaneous measurement of both aspects.

The definitions adopted by ISO and ETSI infers that us ability can only be measured for a specific combination of users, environment and task, and cannot later be generalised. If one of these parameters are changed, the measured usability will also change and must be evaluated again. For example, given this definition, the usability of some system and user combination will change over time as the user becomes more experienced. Therefore, the concept of the learnability of a given interface is considered a separate, or external characteristic to usability. According to ETSI, the same is true for the flexibility (or adaptability) of a system.

2.2. Usability of Speech Based Interfaces

The definitions mentioned above are general for all types of interfaces. However, there are some significant differences between more traditional interfaces incorporating a visual display and speech based interfaces, that must be kept in mind. Most notably, due to the transience of speech, the user can only observe (hear) the system's output information at the exact time it is provided, otherwise s/he will miss it. It also means that the user has no chance of getting an overview of the interface prior to using it. In comparison to a graphical interface, where the user may spend as long as s/he feels necessary to visually inspect the interface to e.g. search for some specific command and in general become familiar with the interface, this is an important difference. Furthermore, the input processing in a SDS (speech recognition and -understanding) is much more complicated and error-prone than most others.

This has some important implications, which must be taken into account when evaluating the usability of speech based interfaces.

Returning to Cameron, he points to the aspects he believes to be the deciding factors for the users' preferences:

- *"users' own time;*
- *their ability to control the pace of their transactions;*
- *their trust in the other party's competence;"* ((Cameron 2000))

He argues that implicitly, people place more value on their time than they are prepared to admit explicitly and continues:

> *"..it is the avoidance of overheads and incidental complexity such as system training, configuration management and error recovery which best respects the high value to users of their own time."*

Very interestingly, this is in direct contrast to the often stated goal of "naturalness" as the overall goal for SDS, but in fact directly related to the usability of the system. This view is also supported by Heisterkamp (Heisterkamp 2003), who argues that "ease of use" is not synonymous with naturalness and may indeed be more important to users than naturalness and that the attention of speech researchers and -developers should be turned more towards this issue.

Cameron clearly identifies a number of criteria with the users' time as the most important factor for success of speech driven services. By time is meant both the actual time to complete a given task, but just as important, the time to learn to use the service, often referred to as the system learnability. The control of the dialogue pace is also found to be an important issue. Due to the inherent transience of speech, transparence and memorability also becomes highly important usability attributes.

Therefore, these attributes can be assumed to be of major importance for SDS's and special attention must be paid to ensure that these are part of the design specification for SDS, as well as the evaluation scheme.

To sum up this discussion, the following usability attributes have been identified as of special importance to SDS: Learnability, Memorability, feeling in control, transparence and help / efficient error recovery

3. Usability Measurements

As mentioned above, usability is measured as a combination of objective and subjective measures.

3.1. Objective Measures- Efficiency and Effectiveness

Many different performance measures for SDS have been suggested and used over time. This section briefly

presents and discusses the most widely accepted. The short-list shown below gives an impression of the nature of the measures:

Communication efficiency, speed

- *Duration of system and user turns*
- *System and user response delays*
- *Dialogue and subtask duration*
- *Time-out prompts*
- *Implicit recovery*
- *Number of turns in dialogue and subtasks*

Task Efficiency

- *Dialogue and task success rates*

3.2. A Case Study: The OVID home banking Application

To illustrate how objective measurements can be used to investigate the learnability and user control issues a number of concrete measurements on the OVID home banking[1] experiment are presented in this section.

To analyse the learnability of the application, the number of turns, the task completion times and the task success rates are recorded for 300 users carrying out two dialogue scenarios (A and B). Improvements in the user's time and performance are interpreted as indications of system learnability.

Task	Duration of First Dialogue (secs)		Duration of Second Dialogue (secs)		Δ
Id number	20.8	22%	17.4 secs	20%	16%
Access code	11.3	12%	9.8 secs	12%	13%
Id + Access code	32.1	33%	27.2 secs	32%	16%
Total	93.0	100%	85.0 secs	100%	8%

Table 1 Duration of user the authentication procedure. The last column is the reduction from the first to the second dialogues

Table 1 shows the amount of time spent in the two user authentication ("Id" and "Access") sub tasks of the home banking service. The table shows a reduction in time of approximately 10-15% from the first to the second time the user carries out the task. This is a clear indication of system learnability.

Similarly, the task success rates are compared for the first and second dialogues (see Table 2 below). As before, there is a significant (almost 50%) reduction in the proportion of failed dialogues

Dialogue	Total number of Dialogues	Succeeded		Failed	
		Dialogues	%	Dialogues	%
First	310	225	73	85	27
Second	303	259	85	44	15

Table 2 The proportion of users who succeeded or failed to complete the scenario of their first and second dialogues.

The question of whether the user is in control is investigated by analysing to which degree users actually do take the initiative in the dialogue. Figure 1 shows that users actually do take the initiative at various points during the dialogue. The dialogue scenarios have been constructed to include one obvious opportunity for the user to take the initiative in scenario A and two in scenario B. This is illustrated in more detail in Figure 1 . The figure also demonstrates that users tend to take the initiative more often, when they become more experienced in interacting with the system. An unpaired two-tailed t-test shows a significant ($p = 0.02$) increase in the number of user initiatives relative to the total number of turns for scenario B2 compared to B1.

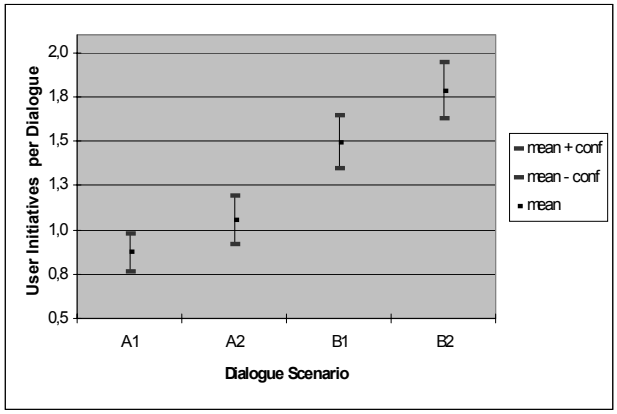

Figure 1 Average number of user initiatives per dialogue with 95% confidence Intervals

The OVID experiment was not suitable for an analysis of the memorability of the service, since the users were not required to repeat the scenarios after some interval in time.

3.3. Subjective Measures - User Satisfaction

One issue that has received little attention is the methods for recording the user's attitudes (Larsen 2003b). Since peoples' attitudes cannot be observed directly, the only way to obtain information about them is to ask the users after they have been exposed to the system. This can be done in a number of ways, such as interviews and questionnaires.

Most often, a set of Likert-like statements are put together on an ad-hoc basis and the mean is computed and interpreted as a measure of "the overall user satisfaction". A problem with this method is how **valid** and **reliable** the answers really are. In most cases the user satisfaction measure is extracted from a questionnaire, where the users are required to respond to a number of issues related to their perception of interacting with the system by ticking off their "agreement" to a number of statements (a Likert scale). The result is obviously highly dependent on the

1. The OVID project addressed the domain of home banking, and involved usability field trials in Denmark and the U.K in close collaboration with three banks. The OVID project has previously been reported in reports and articles, see ((Larsen 1999),(Larsen 2003a),(Larsen 2003b)).

nature of the questions, and this method does not in any way ensure that the outcome is a truly valid representation of the user's attitudes towards the system. Like any other measuring instrument, a questionnaire must be carefully validated before it is used, otherwise the results and conclusions can easily be misleading.

Developing a usability questionnaire is a time-consuming and difficult process and therefore it is often done on an ad-hoc basis, as mentioned above. Validation can be done by comparing to similar, previously validated questionnaires and by a careful analysis of the relationships between the individual Likert statements through the application of Factoring. The reliability can e.g. be established by computing the internal consistency of the questionnaire, expressed by Cronbach's Alpha. See e.g. (Larsen 2003b) for a general discussion of this method, and the application to the OVID questionnaire.

3.4. Combining Objective and Subjective Measures in PARADISE

An important question is of course how to combine the objective and subjective measures to derive a generalised model, e.g. capable of predicting e.g. the speech recognition rates' impact on user satisfaction. The PARADISE model was proposed by Walker and colleagues in 1998 for this purpose. It uses Multivariate Linear Regression (MLR) to derive a model that is capable of predicting user satisfaction from a number of performance measures. The

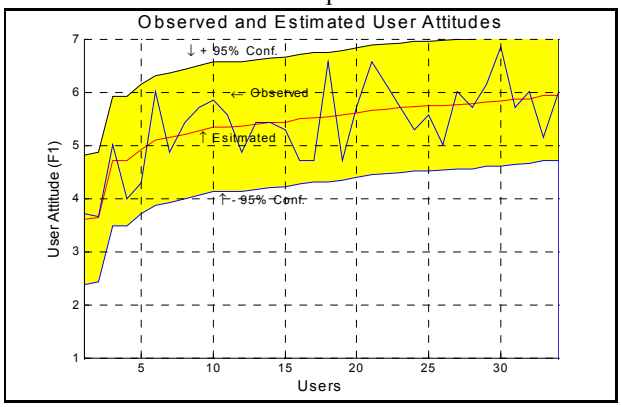

Figure 2 Observed and estimated user attitudes, using PARADISE. The red line represents the estimated user attitudes, and the blue line the observed values. The 95% confidence band is shown in yellow.

PARADIRS MLR for the OVID experiment produced the relationship shown below.

User Satisfaction = 0.41*Task Success+ 0.47*Recognition performance

Figure 2 shows the result of using the model to predict the user attitude for a subset of the OVID experiment.

This model only captures about 50% of the variability of the user satisfaction, which is also evident from the graph. This result is comparable to the original results reported in (Walker et al. 1998) and it raises the question whether PARADISE is really usable for practical purposes. Furthermore, looking closely on the individual statements in the usability questionnaires it turns out that the statements the models predict best are all related to the performance measures - which is not surprising, but indicates that maybe the model gives a biased view on the usability. The fact that only half of the variability of user satisfaction can be captured by the model also supports this notion.

4. Conclusion and Future Work

The work presented here argues for the necessity of systematically evaluating the usability of SDS. It argues that, while the definition of usability is general and of course also applicable to the case of SDS, certain attributes, such as system learnability and transparance become more prominent for SDS due to the transient nature of speech. It is demonstrated that information of learnability can be obtained from quite straight-forward measures, such as number of turns and time spent in (sub)tasks. Similarly, by computing the proportion of user-initiated tasks an indication of who's in control of the interaction can be obtained. PARADISE is applied to the case presented here, but is only able to predict roughly half of the variability of usability.

The "naturalness vs. ease-of-use" discussion is important and must continue beyond the current context. Likewise, the work towards systematic procedures and measures of SDS is of great importance to ensure the success of future speech-driven services.

5. References

Dybkjær and Bernsen 2000: Laila Dybkjær and Niels Ole Bernsen: "Usability issues in spoken dialogue systems", in Natural Language Engineering 6 (3{4): pp. 243-271. 2000

Cameron 2000: Hugh Cameron: "Speech at the Interface", in Proc. of the COST 249 workshop: Voice Operated Telecom Services - do they have a bright future?", Ghent, May 2000

ETSI 1993: European Telecommunications Standards Institute (ETSI): "Human Factors (HF); Guide for usability evaluations of telecommunications systems and services" (ETR 095), Sophia-Antipolis 1993

Heisterkamp 2003: Paul Heisterkamp: "Do not attempt to light with match!: Some thoughts on progress and research goals in Spoken Dialogue Systems" in Proc of Eurospeech 2003, Geneva 2003.

ISO 1998: International Standardisation Organisation (ISO): "ISO 9241: Ergonomic requirements for office work with visual display terminals (VDTs) -- Part 11: Guidance on usability" http://www.iso.org

Larsen 1999: L.B. Larsen: "Combining Objective and Subjective Data in Evaluation of Spoken Dialogues", in Proceedings of the ESCA ETRW on Interactive Dialogue Systems, Kloster Irsee, Germany, 1999

Larsen 2003a: "L.B. Larsen: "Assessment of Spoken Dialogue System Usability - What are We really Measuring?" Proc. of Eurospeech'03, Geneva Switzerland, September 2003

Larsen 2003b: L.B. Larsen: "On the Evaluation of Spoken Dialogue Systems" Ph.D. Thesis, Aalborg University, July 2003.

Nielsen 2003: Jakob Nielsen: "Voice Interfaces: Assessing the Potential". Jakob Nielsen's Alertbox, January 27, 2003. http://useit.com/alertbox/20030127.html (accessed March 2003).

Walker et al. 1998: Marilyn. A. Walker, Diane J. Litman, Candace. A. Kamm and Alicia Abella. "Evaluating Spoken Dialogue Agents with PARADISE: Two Case Studies." In Computer Speech and Language, 12-3, 1998.

Walker 2004: The DARPA Communicator Evaluation Committee, see: http://www.dcs.shef.ac.uk/~walker/paradise.html