Assessment of Spoken Dialogue System Usability - What are We really Measuring?

Lars Bo Larsen

CPK - Center for PersonKommunikation, Dept. of Communication Technology
Aalborg University, DK-9220 Aalborg, Denmark.
Email: lbl@cpk.auc.dk

Abstract

Speech based interfaces have not experienced the breakthrough many have predicted during the last decade. This paper attempts to clarify some of the reasons why by investigating the currently applied methods of usability evaluation. Usability attributes especially important for speech based interfaces are identified and discussed. It is shown that subjective measures (even for widespread evaluation schemes, such as PARADISE) are mostly done in an ad hoc manner and are rarely validated. A comparison is made between some well-known scales, and through an example application of the CCIR usability questionnaire it is shown how validation of the subjective measures can be performed.

1. Introduction

This work attempts to clarify some of the reasons why speech based interfaces still - despite many predictions of “imminent breakthroughs” (see e.g. [1]) and substantial technological advancements - are still some way from achieving this bright future. While the performance of individual modules - such as speech recognisers - has reached an impressive level during the last decade, 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 [1] 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* [1]

The first three reasons relate in varying degrees to external constraints on the user. The last one is obviously “the best one”, seen from a speech service developer’s viewpoint. Unfortunately, Cameron concludes that it has rarely been used (so far).

It is of vital importance to the speech community to determine which (or both) of the explanations suggested above is correct. Despite a growing attention to this, no clear answer has so far been provided. One reason for this could well be the fact 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.

Investigating the Best Practises of Spoken Language Dialogue Systems (the DISC projects [2]), Dybkjær and Bernsen observe that:

> *“Far less resources have been invested in human factors for SLDSs than in SLDS[1] 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.”*[3]

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

2. Definition(s) of Usability

There are many different definitions of usability. However, almost all refers to the three key concepts defined in the ISO 9241 Standard:

> *Usability: The effectiveness, efficiency, and satisfaction with which specified users achieve specified goals in particular environments.*[4]

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 [5], who also points out that usability, together with the costs and benefits for the user, form the concept of utility.

[1] Spoken Language Dialogue System

The definition adopted by ISO and ETSI infers that usability 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.

However, these viewpoints are not shared by all researchers. For example, Jakob Nielsen [6] places usability as a node in a tree depicting the overall “acceptability” of a product, see Figure 1.

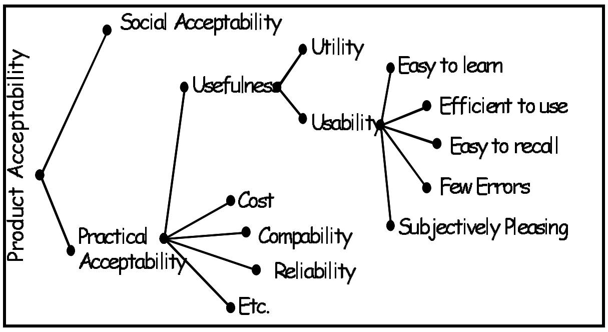

Figure 1. Jakob Nielsen’s definition of usability (redrawn from [6], p.25)

Clearly, Nielsen regards usability and utility to be components of what he denotes **usefulness**, which again is separate from e.g. cost. Contrary to ISO and ETSI he defines usability as a kind of intrinsic characteristic, without a specific user, task and environment in mind. Indeed, Nielsen states that “*Learnability is in some sense the most fundamental usability attribute*”[6]. His definition is supported by other researchers, such as Shneiderman [7], Preece et al. [8]. In particular, Preece et al. argues that **utility** is an attribute of usability and furthermore adds **safety**. The point of Figure 1 is to illustrate that the “Overall Acceptability” of a product or technology is determined by a complex interaction of may factors, all of which must eventually be understood.

2.1. The usability of speech-based interaction

The discussion above addresses the usability of HCI systems in general. Since the definitions are abstract and general, these are obviously also true for speech based interaction. However, as Dybkjær and Bernsen [3] point out, there are some significant differences between more traditional graphical interfaces and speech based interfaces, that must be kept in mind:

> *“In general terms, a usable SLDS must satisfy user needs which are similar to those which must be satisfied by other interactive systems..... However, SLDSs are very different from more traditional interactive systems whose human factors aspects have been investigated for decades,..... Perhaps the most important difference is that speech is perceptually transient rather than static.”* [3]

This has some important implications, which must be taken into account when evaluating the usability of spoken interaction. Most notably, 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 (compared to e.g. a graphical interface). Furthermore, the input processing in a SLDS (speech recognition and -understanding) is comparatively much more complicated and error-prone than most other modalities.

Therefore, it must be anticipated that attributes pertaining to these issues (i.e. learnability, error handling, user control, transparence, etc.), will have a higher impact on the overall usability of spoken interfaces compared to more traditional ones.

Unfortunately, an important consequence of this is that use of standardised methods and scales such as the well-known QUIS ([7],[9]) and SUMI ([10],[11]) questionnaires becomes problematic - as a minimum the validity of the scales must be (re-)established before being applied to speech based interfaces to avoid bias due to the increased perceptual weight of the attributes mentioned above.

3. Usability Measures

Since the early nineties, evaluation of spoken dialogue system usability has largely been based on field trials, where two distinct measures, denoted “Objective” and “Subjective” are collected and analysed.

Objective measures have been given much consideration and multiple metrics have been proposed and used, such as task completion times and -success rates, proportion of repair- and help-requests, speech understanding and -recognition rates, barge-ins and dialogue initiative.

In some cases, e.g. in the PARADISE [12] evaluation scheme, the objective measures have been divided into categories relating to either the quality of the interaction or the dialogue costs (i.e. the cost for the user to obtain some piece of information, e.g. measured in number of turns or time). Although often requiring extensive and time-consuming tagging of corpora, it is fairly straightforward to define and obtain quantitative data for objective measures.

For example, Walker and colleagues used elapsed time, system turns, prompt timeouts and the mean speech concept recognition score (SR). The Kappa coefficient is used to estimate task success (to compensate for complexity) in [12],[13]. In the OVID project [14] SR was used together with (sub)task duration, number of turns and number of user initiatives [15]. Other metrics are percentages of help requests, repair utterances, contextually correct system utterances, barge-ins, timeouts, etc.

Subjective Measures. Compared to this, subjective or attitude measures are more elusive. Since peoples' attitudes cannot be observed directly, the only way to obtain information about them is to ask the test users after they have been exposed to the system. This can be done in a number of ways, such as interviews and questionnaires.

Common to all is the problem of how **valid** and **reliable** the answers 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 nature of the questions.

Determining "the right questions", and especially establishing that the obtained results are indeed representative of the users' true attitudes are by no means a simple matter and has often been overlooked or ignored by researchers. One common problem is that researchers in speech technology do not seem to realise that a scale, like any other measuring instrument must be carefully designed, documented and validated, if the measurements are to be scientifically valid [16]. For example, even though there are numerous articles documenting the PARADISE scheme, no validation of the questionnaire used to obtain subjective measures has yet been published. [13].

Hone and Graham review a number of subjective speech system evaluations and state that: *"It can be concluded that none of the existing techniques for subjective speech interface meet the criteria for a valid psychometric instrument"* [16]. However, some efforts have been made, especially by the Center for Communication Interface Research (CCIR) at Edinburgh University in collaboration with British Telecom in the "Intelligent Dialogue Project" in the early nineties [17],[18]. Table 1 compares the development of four user attitude scales. Two (CCIR-BT and SASSI) have been developed especially for speech-based interfaces. SUMI and QUIS are included for comparison. Unfortunately, the development of the SASSI tool has only completed the first iteration and has apparently been discontinued. It is evident from the table that the development of a scale is a very time demanding process. Especially establishing the validity of a scale is difficult and requires expertise and resources.

	QUIS [9]	**SUMI** [11]	**CCIR-BT**[17]	**SASSI** [16]
Purpose	**Usability of generic GUI interfaces**	**Usability of generic GUI interfaces**	**Specifically targeted for voice based telephone interfaces**	**Specifically targeted for voice based interfaces**
Initial Version (First Iteration)	Previous research and experience, literature 90 items, 5 overall and 85 specific, divided into 20 sub groups.	Previous research and experience, literature 150 items, grouped and reduced to 75 Tested on 139 subjects	Previous research and experience, literature 22 items in core set, 3 application specific Validated by 20 experts and 20 users in control group	Previous research and experience, literature 50 items. Data collected from 226 users across 4 studies of 8 applications. 6 subscales identified, with subscale reliability in the range of 0.7-0.9 (Cronbachs Alpha)
Second Iteration	Total of 110 items. Tested for reliability: (Cronbach's alpha 0.94) by 213 users.	50 items, based on initial version. Tested on 143 users. 5 groups identified by factoring	22 core revised items, based on initial version. 5 sub groups, identified by factoring	
Third Iteration	70 items, identified by factoring. Reliability is 0.89 (Cronbach) 150 users	25 items, Reliability is 0.92 (Cronbach). Tested by more than 1100 users	Validating the questionnaire by factoring and testing for predictive power. 40 (used for factoring)+20 (test) users Have been used for numerous evaluations, see [18]	
Fourth Iteration	Version 5 and 5.5. Includes 6 general + 22 specific items. Ver. 5.5 is an online version			

Table 1 Comparison of the iterative development process for the SUMI, QUIS, CCIR-BT and SASSI questionnaires

4. Verification of a Scale - a Case Study

The CCIR-BT scale was used in a field trial within the OVID project to evaluate the usability of speech-based home banking systems [14],[15]. The statements were translated into another language (Danish), a process that potentially threatens the previously established validity of the scale. The following steps was taken to ensure the reliability and validity of the translated scale:

- The translation was done in collaboration with CCIR and cross-checked by two Danish speech experts and one banking expert
- Two iterations of a pre-test was carried out, first with 7 (speech experts) and then 20 test users, who were also asked to supply feedback on the questionnaire itself.

The main experiment involved 310 users calling the service in a field trial. All users filled out and returned the questionnaire after two scenarios had been completed. The internal consistency (reliability) was estimated by computing Cronbachs' coefficient Alpha, which was found to be satisfactory (0.92). In order to compare with previous results, the items were subjected to Factoring [19]. Five factors were identified with all item loadings above 0.4 and a difference between loadings greater that 0.2. Coefficient Alpha for the subscales were in the range (0.78-0.92) which is acceptable. A principal component analysis showed that the first five components explained 71% of the total item variance. The result of a factorisation with five subscales are shown below in Table 2 together with appropriate labels:

Sub Scales	**Alpha**	**Var[1]**
Quality of interface/efficiency/reliability	0.86	19%
Cognitive effort/stress	0.83	13%
Transparency/confusion	0.78	8%
Friendliness	0.82	8%
Voice	0.92	8%

Table 2 Identified Subscales from the OVID experiment

[1] The proportion of the explained variance of the Factors

This corresponds well with previous results obtained by CCIR. The two first subscales contains exactly the same items as found in [17], whereas some differences were found in the following.

5. Conclusions

The discussion has pointed to some problems in the process of evaluating speech based interfaces and in particular identified the inadequacy of current methods for subjective evaluation. If scales especially targeted towards speech interfaces are not systematically designed and validated, but rather composed in an ad hoc manner, there will be no guarantee that what is measured actually corresponds with the real attitudes of users.

However, user attitudes are only one attribute of system acceptability as indicated in Figure 1. As Cameron points out [1], users will not embrace a technology unless it holds a real benefit for them, compared to other alternatives, e.g. greater speed or comfort. Before all aspects are fully understood and included in end-user studies, speech service developers are in a high-risk business.

References

[1] 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

[2] The website for the DISC1 and DISC2 projects. Last Revised: 11 March, 2001: http://www.disc2.dk. Visited March 2003

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

[4] 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

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

[6] Jakob Nielsen: *“Usability Engineering”* Academic Press, Inc. San Diego, USA, 1993. ISBN 0-12-518405-0

[7] Ben Shneiderman: “Designing the User Interface”. 3rd edition, Addison-Wesley, Reading Massachusetts 1998. ISBN: 0-201-69497-2

[8] Jennifer Preece, Y. Rogers, H. Sharp: “Interaction Design”, John Wiley and Sons,. U.S, 2002. ISBN 0-471-49278-7. http://ID-Book.com

[9] Chin, J. P., Diehl, V. A. and Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. Proceedings of SIGCHI '88, (pp. 213-218), New York: ACM/SIGCHI. Also available as: http://lap.umd.edu/quis/publications/chin1988.pdf

[10]SUMI (Software Usability Measurement Inventory) Human Factors Research Group University College Cork, Ireland: http://sumi.ucc.ie/index.html. Feb. 2003

[11]Jurek Kirakowski: “Background notes on the SUMI questionnaire” Human Factors Research Group University College Cork, Ireland. Originally 1994. WWW version http://www.ucc.ie/hfrg/questionnaires/sumi/sumipapp.html. Feb. 2003

[12]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.

[13]DARPA Communicator Evaluation website: (April 2003) http://www.dcs.shef.ac.uk/~walker/paradise.html

[14]Lars Bo Larsen, “Voice Controlled Home Banking - Objectives and Experiences of the Esprit Ovid Project”, *IVTTA-96 workshop*, September, 1996.

[15]Lars Bo 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

[16]Kate Hone and R. Graham: “Towards a Tool for the Subjective Assessment of Speech System Interfaces (SASSI)”. Natural Language Engineering, 6(3-4), 287-303. 2000.

[17]S. Love, R.T. Dutton, J.C. Foster, M.A. Jack and F.W.M. Stentiford, “Identifying salient usability attributes for automated telephone services”, *Proc. International Conference on Spoken Language Processing (ICSLP-94), pp.1307-1310*, September 1994

[18]CCIR “Intelligent dialogues project” (visited April 2003) http://www.ccir.ed.ac.uk/doc/ccir_dialogues_reports.htm

[19]Richard B. Darlington: “Factor Analysis”. January 1997. http://comp9.psych.cornell.edu/Darlington/factor.htm (last visited April 2003)