On the Usability of Spoken Dialogue Systems

Larsen, Lars Bo

Publication date:
2003

Document Version
Publisher's PDF, also known as Version of record

Link to publication from Aalborg University

Citation for published version (APA):
Larsen, L. B. (2003). *On the Usability of Spoken Dialogue Systems*. Aalborg University, Department of Communication Technology.

On the Usability of Spoken Dialogue Systems

Lars Bo Larsen

SMC - Speech and Multimedia Communication
Dept. of Communication Technology

Ph.D. Thesis submitted to
The Faculty of Engineering and Science
Aalborg University
Denmark

in partial fulfilment of the requirements for obtaining the
degree of
Doctor of Philosophy
in Electrical and Electronic Engineering
June, 2003

Report No. R03-1011 ISBN 87-90834-32-1 ISSN 0908-1224

Summary

This work addresses usability evaluation of Spoken Dialogue Systems through field trials. It reports on the results of the Danish experiments within the ESPRIT OVID project. The experiments were carried out in 1997 at the Center for PersonKommunikation (CPK), Aalborg University. The experiments address the domain of home banking, i.e. telephone access to a speech controlled service, enabling the user to obtain information about his/her bank accounts.

The OVID experiments and the obtained results are mainly documented in a number of articles and reports, included in the current report. These have been published in the period 1996 to 1999, except for two more recent papers in 2003. In these, the requirements for the OVID service are defined and documented, as well as methods for design and implementation of the service. The results are presented and evaluated, based on more than 700 transcribed dialogues by 310 users. The system's learnability is analysed through the turn-taking strategies and it is shown that users are capable of taking the initiative in the interaction after only two dialogues with the service. User satisfaction is found to be generally high, and no significant demographic differences have been identified.

Usability evaluation of speech based services and the theoretical background for the elicitation of user attitudes through the use of questionnaires are introduced and discussed. This is traditionally a branch of the social sciences denoted psychometrics, and it is a field that has received comparatively little attention from the speech community. As a consequence the field is not fully mature and methods and tools have only to a limited degree been developed and applied properly. The user attitude questionnaire used in the OVID experiments is analysed and shown to be valid and reliable, and results comparable to those obtained by other researchers are demonstrated.

The PARADISE scheme for the evaluation of Spoken Dialogue Systems is described and applied to a subset of the OVID dialogue corpus and the results are analysed and discussed in detail. PARADISE attempts to estimate a performance function through Multiple Linear Regression (MLR). Thus user satisfaction is directly expressed as a function of various measures of Dialogue Costs and Dialogue Quality. It is found that speech recognition- and task success rates are significant predictors of user satisfaction, and a model is derived which predicts user attitudes from these two measures. The model explains 51% of the variability of the user's attitudes towards the system, a result comparable to those obtained from similar experiments in the U.S.

The strenghts and weaknesses of the applied theories and methods are identified and discussed throughout the report.

Danish Summary

Denne rapport omhandler evaluering af brugbarheden af Talestyrede Dialogsystemer gennem feltforsøg. Resultaterne fra de danske eksperimenter i Esprit OVID projektet beskrives. Eksperimenterne blev udført i 1997 på Center for PersonKommunikation (CPK), Aalborg Universitet. Eksperimentet omhandler en talestyret service, der giver brugeren telefonisk tilgang til hans/hendes bankkonti.

OVID eksperimenterne og de opnåede resultater er hovedsageligt dokumenteret i et antal artikler og rapporter, inkluderet i et appendiks til denne rapport. Disse har været publiceret i perioden 1996-1999, på nær to artikler i 2003. Heri specificeres og dokumenteres OVID telefonbanken, såvel som metoder for design og implementering. Resultaterne, der præsenteres og analyseres er baserede på mere end 700 transskriberede dialoger udført af 310 brugere. Graden af systemets læringsevne belyses, og det påvises at brugernes evne til at tage initiativet i dialogen øges allerede efter to opkald til systemet. Brugernes tilfredshed er generelt høj, og der blev ikke påvist demografiske forskelle.

Evaluering af brugbarheden af talestyrede systemer og den teoretiske baggrund for udtrække brugerenes holdninger via spørgeskemaer introduceres og diskuteres. Dette er et område der har tiltrukket relativt lille interesse indenfor taleteknologiforskningen. Som en konsekvens heraf, er området ikke fuldt udviklet og metoder og værktøjer er kun i begrænset blevet udviklet og korrekt anvendt. Det spørgeskema, der anvendes i OVID eksperimentet påvises at være validt og pålideligt, og de opnåede resultater er sammenlignelige med resultater opnået andetsteds.

PARADISE paradigmet til evaluering af Talestyrede Dialogsystemer anvendes på OVID korpuset og resultaterne analyseres og diskuteres i detaljer. Paradise forsøger at sammenkoble objektive og subjektive evalueringsresultater og opstille en performans funktion vha. Multivariabel Lineær Regression (MLR). Derved udtrykkes bruger tilfredshed direkte som en funktion af mål for dialogens performans og kvalitet. Det påvises at talegenkendelses- og et mål for dialog succes rater er signifikante prediktorer for bruger tilfredshed og en model opstilles, der forklarer 51% af variabiliteten af brugertilfredsheden, hvilket er sammenligneligt med resultater opnået i USA.

De anvendte teorier og metoders styrker og svagheder beskrives og evalueres løbende igennem rapporten.

Preface

This thesis report is submitted in partial fulfilment of the requirements of the Faculty of Engineering and Science at Aalborg University for the degree of Doctor of Philosophy (Ph.D.). It is based on the authors' more than ten years of experience of Spoken Dialogue Systems, acquired through the participation in various Danish and International (EU) research projects. The main focus is on the work carried out within the Danish part of the Esprit OVID[1] project, and the articles and reports published from that project in the period 1996-1999. These are included here and form the basis of the present report. Reprints of the complete articles have been placed in Appendix C, page 111 ff. To supplement the articles a more recent part has been added, containing an in-depth presentation and discussion of the theoretical background of the methods applied in the articles. The results of more recent work carried out during 2002-03 are also presented here.

Rather than structuring the present report as a presentation and elaboration of the methods and results of each of the articles one by one, I have chosen to structure the first part as a stand-alone text in its own right to increase the readability and smoothly integrate the results of more recent analyses. The intention is that the present report should appear as a complete work, structured by the addressed subjects and concepts, rather than a chronologically oriented presentation of a list of articles.

A brief introduction to the layout of the report

The report consists of two main parts. The first part starts with an account and discussion of the theoretical background for the methods associated with usability testing in general and in particular for speech based systems. The next chapters focuses on how to obtain and analyse objective and subjective measures. This is followed by a presentation of the most important results of previous and recent analyses of the OVID experimental data. Next, results of applying the PARADISE evaluation scheme, which has emerged in the period after the original work on the OVID project ended, are presented and discussed. The first part concludes with a general discussion and critique of the applied theories and methods and a summary of the implications of the achieved results. Some perspectives for continuation of the research are also given.

[1] The Esprit OVID experiments were carried out in Great Britain by the Centre for Communication Interface Research (CCIR) at Edinburgh University, together with the Royal Bank of Scotland and Barclays Bank. The Danish experiments were carried out by CPK at Aalborg University together with the Lån & Spar Bank during the spring of 1997.

The second part consists of the previously published articles and reports. The included articles are thus referenced in the standard manner throughout the report. The titles are listed below, and a brief resume of each article is given in Appendix C, on page 105 ff.

Articles included in Appendix C[1]

“Voice Controlled Home Banking - Objectives and Experiences of the Esprit Ovid Project”, *IVTTA-96 workshop*, September, 1996. See page 111 ff.

“A Strategy for Mixed-initiative Dialogue Control”, *Proceedings of Eurospeech '97*, September, 1997. See page 119 ff.

“Investigating a Mixed-Initiative Dialogue Management Strategy”, *Proceedings of IEEE Workshop on Speech Recognition and Understanding, ASRU 1997*, December, 1997. See page 129ff.

“The OVID Project Objectives and Results” Technical Report 98-0201 CPK Aalborg University, March 1998. See page 137 ff.

“Combining Objective and Subjective Data in Evaluation of Spoken Dialogues“, in *Proceedings of the ESCA ETRW on Interactive Dialogue Systems,* Cloister Irsee, Germany, 1999. See page 187 ff.

The two final papers included in Appendix C present more recent work (2002-03) done on the corpus. The subjects are covered in the main part of the report in more details and the articles are included in order to give the reader a more condensed version of the recent findings.

“Assessment of Spoken Dialogue System Usability - What are We really Measuring?” *To appear in the proceedings of Eurospeech’03*, Geneva Switzerland, September 2003. See page 197 ff.

“Applying The PARADISE Evaluation Scheme to an Existing Dialogue Corpus”. *Submitted to ASRU’03*, St. Thomas, U.S., December 2003. See page 205 ff.

The work carried out within the OVID project - which constitutes the bulk of the experimental work reported here - was not originally intended to be part of a Ph.D. thesis, but with goals and constraints of its own. Furthermore, there is a gap of five years between the OVID experiments (and the associated papers and reports) and the present report. Therefore, the field has moved in the meantime, and some recently developed methods obviously could not have been included in the original OVID work. Most notably, the PARADISE methodology has been developed and applied in the U.S. by AT&T and the Darpa community. Therefore, some minor inconsistencies and/or omissions might be observed while reading the original articles compared to the text in

[1] All articles are exclusively written by the author.

the first part in the report. Nevertheless, I have chosen to order the presentation of the experiments and results in a logical manner more than a strictly chronological.

The thesis deals extensively with issues relating to spoken dialogue systems. The focus is on the theories and methods associated with the evaluation of the usability of such systems and as such does not require a detailed knowledge of the processing taking place within the system. However, it might be profitable for readers unfamiliar with the concepts to turn to Appendix A on page 89 for a brief introduction of spoken dialogue systems. Readers familiar with the basic concepts will have no need to do so.

Background

It might puzzle some readers why there is a five years gap between the OVID experiments and the present thesis. In order to understand the background of the thesis, a short account of my recent activities is provided here. The OVID project ended in 1997 and immediately following this, I joined a project focusing on Multi-Modal User Interaction (the MMUI project). A part of that effort was to set up a new cross-disciplinary Masters programme. I became heavily involved in the planning of this, and I have since it’s initiation been the coordinator of the programme. Apart from being a completely new programme reaching across several research areas including computer science and electronic engineering, it was also pioneering the internationalisation of the educational programmes at the Faculty of Science and Technology at Aalborg University. Unfortunately (for me), this turned out to be a more than full-time job, and the Intelligent MultiMedia (IMM) programme has currently produced more than 60 master candidates, a fourth of those coming from abroad.

Apart from managing the IMM programme, I have authored and co-authored a number of research reports and conference and journal papers on issues of multi-modal user interaction. However, these are addressing a number of different topics and are not suitable as a basis for a thesis. A list of these recent publications are included in Appendix C, page 107 ff., but not the articles themselves.

When given the opportunity to return to full-time research again for a period, I therefore chose to take up some of the research issues from the evaluation of the OVID experiments that had intrigued me at the time, but which I had to leave behind for the tasks described above. The present thesis is the result of this.

Acknowledgements

I wish to thank the members of the OVID project team in Denmark and the U.K. for their support and professionalism. In the U.K., especially Mervyn Jack and Keith Edwards from CCIR, Peter Dalziel from the Royal Bank of Scotland, and Lois Parkins and Kevin Quinn from Barclays Bank. In Denmark, Lars Rud from Lån & Spar Bank and Paul Dalsgaard from CPK. Not only did I learn a lot from you, we also had great fun.

Many of my colleagues at CPK have put their time and expertise at my disposal when I worked at this report, and I can not mention you all. However, in particular I'm grateful to Børge Lindberg for his work on the speech recogniser we used for the OVID experiments, I recall some late nights then. Also thanks to Anders Bækgaard for his work on the dialogue platform and to Louise Dalsgaard, who transcribed all the dialogues in the OVID corpus. More recently Børge and Tom Brøndsted took over a lot of my teaching and other tasks, although they both have plenty of work to do of their own, and to Lisbeth Schiønning, who practically has been running the IMM for the past year.

Without Paul Dalsgaard's continuing support and encouragement (and patience) I would not have been able to put this work together, for which I'm deeply grateful. Also thanks to Paul for spending his time on reading various versions of this report and his many valuable comments and suggestions.

Last, but not least, I'm grateful for the support from my family Anne, Anders and Jens, and especially my wife Lone for bearing with the many long evenings and weekends spent with me sitting in front of the computer.

Although I have received much help and advice from colleagues, any mistakes, omissions or errors in the present work are solely the responsibility of myself and cannot in any way be ascribed to them.

June, 2003

Lars Bo Larsen

Speech and Multi Media Division, Dept. of Communication Technology,
Institute of Electronic Systems, Aalborg University

Notation and References

Quotations. Quotes are either clearly indicated in a separate indented paragraph, or shown inline for very short quotes. Quotes are always shown in *“italics”*. A string of dots *“....”* within a quote indicates that a sentence or paragraph has been omitted for clarity or brevity. The reference is included in the same paragraph as the quotation, and is specified with page number where applicable. This is also the case when citing figures. If a figure has been redrawn or modified instead of directly copied from the original source, this is indicated in the figure caption along with the reference.

References. A reference index is included at the back of the report for easy access in the section “References” on page 217 ff. A reference includes authors, title, and the original source (book, journal, proceedings, etc.). If the reference is available on-line, a Web URL is included together with a date. However, due to the dynamic nature of the Web, the correctness of the URLs cannot be guaranteed. Regardless, the links are considered a valuable asset, for those wishing to obtain the referred articles. Citations are shown in brackets, with the surname(s) of the author(s) together with the year of publication, like e.g. (Jurafsky and Martin 2000). This is preferred to the standard ieee citation style (e.g. [21]) for better readability.

Abbreviations and Acronyms. A list of abbreviations and acronyms is placed at page 225.

Figures and Tables. Figure and tables are numbered consecutively throughout the first part of the report and individually for each Appendix.

Contents

Chapter 1 Introduction

During the past ten to fifteen years speech technology has evolved from a few commercial applications of limited size in terms of vocabulary and complexity (often with isolated word recognition) to more elaborate systems, capable of handling much more complex task domains, larger sized vocabularies, more complex linguistic constructions and multi modality. The success of the Web has spurred standardisation work as for example VoiceXML[1] and more recently SALT[2], in an attempt to boost the deployment of voice-driven services.

In research laboratories, there have been a similar development. Ten years ago prototypes of applications that has now evolved into the current commercial state-of-the-art were developed. For example, the EU SUNDIAL[3] (see e.g. Simpson and Fraser 1993) and SUNSTAR (see e.g. Dalsgaard and Bækgaard 1994) and the U.S. DARPA ATIS[4] programme (see e.g. Polifroni et al 1992, Price et al 1992). Speech recognition and -synthesis performance have made dramatic performance improvements. Large databases are now available for more than twenty five European languages and dialects with more than one hundred thousand Europeans participating in the recordings (the SpeechDat projects[5]). By now, the focus has expanded to issues like multi modality, robustness and architectures for distributed wireless speech processing (see e.g. SmartKom[6], AURORA[7]; FACE[8]).

The technological development has increased the computing power of common desktop PCs (and soon even of PDAs) to be capable of executing realtime continuous, large vocabulary speech recognition.

Notwithstanding, by 2003 it is still not a common-day experience for the vast majority of the population to speak to a machine.

Speech technology has more than once been predicted to be on the threshold of a “major commercial breakthrough” and many analysts and profession-

[1] http://voicexml.com/
[2] http://saltforum.org/
[3] http://www.sics.se/~scott/sundial.html
[4] DARPA - (U.S.) Defence Advanced Research Projects Agency. ATIS - Air Traffic Information System
[5] http://www.speechdat.org/
[6] http://www.smartkom.org/start_en.html
[7] http://www.etsi.org/frameset/home.htm?/technicalactiv/dsr/dsr.htm
[8] http://cpk.auc.dk/FACE/

als have believed speech to be the basis for the “next generation” human computer interface. For example:

> *“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?

Several explanations come to mind.

- The field simply isn’t mature yet. It has turned out that speech technology is orders of magnitude more difficult to handle than expected, and for many applications, it still hasn’t reached an acceptable level of accuracy and reliability yet.
- The premises under which the systems were developed did not match the real-life conditions, e.g. for robustness against noise, speaker variability, willingness of users to spend time on configuration and training, ease of use, etc.
- New technologies, most notably the Web, have been able to successfully provide many of the services that was envisioned to be “killer applications” for speech. In other words, speech has turned out not to be competitive.
- Speech might simply not be the best modality for many applications. Speech is often portrayed as “the most natural way for humans to communicate”. While this might be true for human-human communication, there is certainly no evidence that this statement necessarily is true in the general case of human-computer interaction. Research in the most recent years might turn out to show that speech in combination with other modalities (e.g. pen-gestures and graphics) will provide the long-awaited and -predicted breakthrough for speech interaction.

Which of these explanations - or combination of them - turn out be the cause, can only be determined through careful evaluations of how and why the systems are actually used and perceived by their intended users. Or, in other words, by investigating the **usability** of the systems. This issue is most important. During the last 20 years vast amounts of man-power and money have been spent in the EU, U.S and Japan on improving the basic speech technologies and this can not be expected to continue unless a breakthrough is soon achieved. In fact, most countries have already reduced their funding for speech recognition research efforts. Investigating the Best Practises of Spoken Language Dialogue Systems - SLDS (DISC 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 behaviour, or the main factors which determine overall user satisfaction."* (Dybkjær and Bernsen 2000, p. 245)

Hugh Cameron analysed the success and failure of a large number of commercially available 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"* (Cameron 2000, p.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 the case (so far).

This work is concerned with usability evaluation of Spoken Dialogue Systems (SDS) in an attempt to clarify some of these issues. In particular, further understanding of the issues related to usability of SDS is sought by the study of field trials.

The experimental work reported here originates from the OVID[1] project, carried out in 1997 at CPK, Aalborg University. In the OVID project, a preliminary (simulated) trial was first conducted, followed by a full-scale field trial.

The goal of the OVID project was to investigate whether the (then) state-of-the-art of spoken language technology was acceptable to users of a home banking system (OVID 1995, Larsen 1996). This is essentially a usability issue and to establish this, field trials were specified and carried out. User attitudes were recorded using a usability questionnaire, devised by the Centre for Communication Interface Research (CCIR) at Edinburgh University, see (Love et al 1994). The trials were carried out in the U.K. by CCIR in collaboration with Barclays Bank and the Royal Bank of Scotland, and in Denmark by CPK in collaboration with Lån & Spar Bank

[1] The OVID I trial was originally carried out within the Esprit OVID project (OVID 1995) in 1996-1997. See also the Preface on page iii

The OVID field trial opened a number of interesting research questions regarding the methodologies employed in the experiment. Since then, new methods for evaluation of SDSs have emerged, most notably the PARADISE (paradigm for dialogue system evaluation) framework, proposed by AT&T (Walker et al 1998) and used for dialogue strategy evaluation in a number of field experiments, among them the DARPA Communicator Project (Walker et al 2001b, Walker et al 2001c).

Chapter 5 introduces the theoretic background for the elicitation of user attitudes through the use of questionnaires. This is traditionally a branch of the social sciences denoted psychometrics. It is a field with a number of pitfalls, both regarding the psychological aspects as well as the statistical methods applied in the analysis of the results.

The PARADISE scheme for the evaluation of Spoken Dialogue Systems is introduced in Chapter 7 and is applied to the OVID corpus. PARADISE is unique in the sense that it attempts to estimate a performance function through Multiple Linear Regression (MLR). Thus “User Satisfaction” is directly expressed as a function of various metrics for “Dialogue Costs” and “Dialogue Quality”. The scheme has quickly gained support and have been applied to a number of domains, especially in the U.S. However, the approach also has limitations and weaknesses, which are also treated in Chapter 7.

1.1 Summary

To recapitulate, the focus in the thesis is on the application of field trials to evaluate the usability of spoken dialogue systems. This is achieved through the design and test of the OVID home banking service, and the subsequent analysis of the results. The theories underlying the applied methods are introduced and discussed.

A number of related issues, while significant, have been left out of the scope of the present thesis. Most notably, the increasingly important issue of multi modality is not addressed. Partially because the OVID experiment did not include multi modal user interaction and partially to keep a narrow focus in the work.

Chapter 2 Dimensions of Usability

There are many different definitions of usability. However, almost all refer to the three key concepts as defined by ISO (the International Standardisation Organisation) in the 9241 - 11 standard (ISO 1998b) as the core concepts of usability:

> ***"Usability:*** *extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use."* (ISO 1998b, p. 1)

Furthermore:

> ***"Effectiveness:*** *Accuracy and completeness with which which users achieve specified goals.*
> ***Efficiency:*** *Resources expended in relation to the accuracy and completeness with which users achieve goals.*
> ***Satisfaction:*** *Freedom from discomfort, and positive attitudes towards the use of the product."*
> (ISO 1998b, P.1)

The ISO standard also notes that effectiveness and efficiency are often referred to as performance measures.

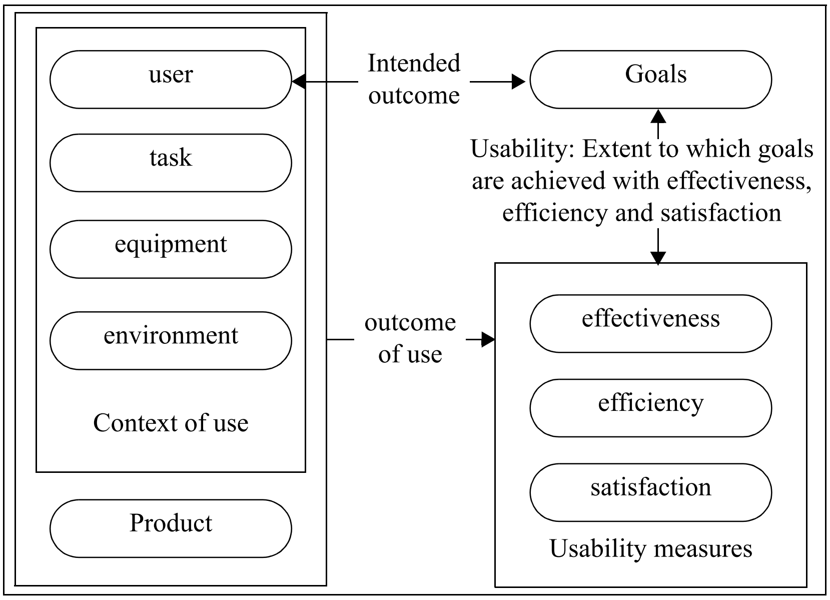


Figure 1. .Usability framework according to ISO/DIS 9241-11.2 (ISO 1998b, p.3)

Figure 1 illustrates the relationships between the “Context of use”, the “Usability measures” and “goals” of usability as it is seen by ISO. ETSI (the European Telecommunications Standards Institute) adopts this view and points out that:

> *“**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 p.13)

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

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. Learnability can be defined in terms of usability, namely the change of usability over time, as a particular user becomes more experienced in using the system.

Likewise, the **flexibility** of a given system is defined as the degree to which the system will adapt to changes in users, task and environment. As with learnability, the flexibility of a system may be established by recording the changes in usability when varying these variables independently (ETSI 1993).

These definitions are a direct consequence of the ISO and ETSI usability standard’s view of usability purely as an aspect of the **users**’ experience, by placing the user as the central point around whom everything revolves. This viewpoint is not without controversy, however. For example, Jakob Nielsen

(Nielsen 1993) regards learnability as an aspect of usability. This is shown in Figure 2

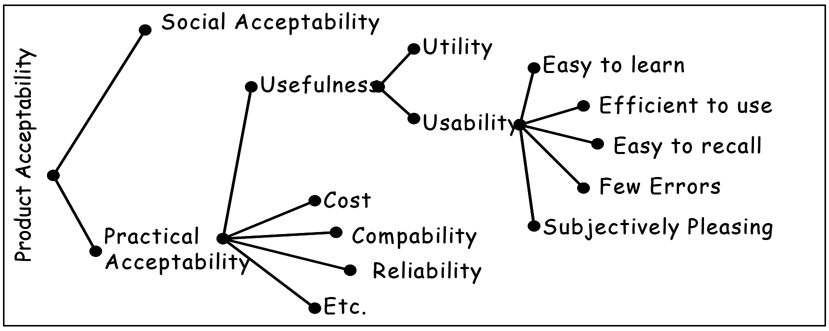


Figure 2. Jakob Nielsen's definition of usability (Nielsen 1993 p. 25)

This viewpoint indicates that he - and others - place the product (or system) at the centre and then defines usability as a kind of intrinsic characteristic, without specifying a user, task and environment. Indeed, Nielsen states that "*Learnability is in some sense the most fundamental usability attribute*" (Nielsen 1993 p 27).

Another divergence to the ISO and ETSI standards' definition is that Nielsen does not regard effectiveness as an attribute of usability. Instead he seems to place it as an attribute of Utility: *"... where utility is the question of whether the functionality of the system in principle can do what is needed, and usability is the question of how well users can use that functionality"* (Nielsen 1993, p. 25). ISO and ETSI defines usability as an attribute of utility along with the product costs and benefits for the user, where Nielsen places both usability and utility as attributes of usefulness.

Ben Shneiderman (Shneiderman 1998) places himself somewhere in-between the two viewpoints. He agrees with Nielsen that usability can be measured along the five dimensions:

- Time to learn (Learnability)
- Speed of Performance (Efficiency)
- Rate of Errors
- Retention over time (Memorability)
- subjective satisfaction (pleasing)

In their book “Interaction Design” Preece, Rogers and Sharp (Preece et al 2002) support this view, but go further and place utility as an attribute of usability. They even suggest the additional attribute “safety” should be included:

- Utility
- Safety

Shneiderman also maintains that the designer must consider these dimensions *“for each user and for each task”* (Shneiderman 1998 p.15) and thus concurs with the ISO definition. He seems to take a very mechanistic view in his definition, as he stresses the importance of being able to precisely measure the usability attributes.

For practical purposes, however, this digression is not overly important. The utility and learnability of a system are important attributes independent of how they are viewed. Learnability must be measured and optimised regardless of whether it is perceived as an aspect external to usability or an attribute of usability itself. Likewise, the system must cater for novice and expert users, be effective and efficient for different tasks and environments, etc.

The definition of usability supported by Nielsen and the other researchers mentioned above will be adopted in this work. The reasons for this are that the ISO/ETSI definition is less operative in the sense that there is a large gap from the definition shown on page 5 to a practical application of the principles. Another reason is that the usability attribute “learnability” turns out to be of particular importance to speech based interfaces. This is further described in section 2.2. Note also, that this implies that the issue of utility will not be considered further.

2.1 Objective and subjective measures

It is important to note the duality of usability measures: The attributes of effectiveness and efficiency can be measured and quantified through observation of users performing various tasks with the system at hand. Concrete measures can be defined and extracted for numbers and types of errors, time spent on performing tasks, number of times help was consulted, etc. These may in turn be analysed and measures for learnability, errors, efficiency, memorability, etc. can be computed for various user profiles, tasks and environments. These measures are often referred to as **objective** or **performance** measures.

In contrast to this, the attributes of user satisfaction can not be directly observed and quantification is more problematic. The only way to establish user satisfaction is to ask users their opinion, after (or during) they have interacted with the system. This is traditionally done either in interviews, questionnaires or a combination of both (see e.g. Nielsen 1993, Rubin 1994, Shneiderman 1998). This information is most often designated user satisfaction, -preferences or -attitudes. ETSI uses the term **subjective** measures as opposed to objective measures. Like the objective measures, user preference data is subjected to quantification and statistical methods are applied to analyse and systemize the information.

Since user preferences are only indirectly observable, the problem of ensuring and proving the **validity** and **reliability** of the data and the methods and tools applied to obtain it becomes primary factors. This problem is one of the principal subjects of this work and is addressed in detail in Chapter 5.

So far, the analysis has only addressed usability in very general terms and the definitions above are not in any way specifically targeted towards interfaces involving spoken interaction. Indeed, the ISO 9411 particularly addresses the usability "*....for office work with visual display terminals*"(ISO 1998b) and the ETSI is concerned with the "*usability evaluations of telecommunications systems*"(ETSI 1993).

The following section focuses the discussion of usability towards SDS and introduces and analyses the currently applied objective and subjective measures of usability for this domain.

2.2 Usability Measures of Speech based interfaces

According to the general definitions of usability discussed in the previous section, two complementary types of measures must be obtained to estimate and evaluate the usability of a specific system: Objective (or performance) measures and subjective (or user preference/attitude) measures. Since the definitions are abstract and general, this is obviously also true for speech based interaction. However, as Dybkjær and Bernsen point out, there are some significant differences between more traditional interfaces incorporating a visual display and speech based interfaces, that must be kept in mind:

> *"In general terms, a usable SLDS*[1] *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."*
> (Dybkjær and Bernsen 2000, p.245)

[1] SLDS: Spoken Language Dialogue Systems

This is also noted in the ISO9241-11 Standard:

> *"Care should be taken in generalization the results of any measurement of usability to another context which may have significantly different types of users, tasks or environments"(ISO 1998b, p.5)*

This has some important implications, which must be taken into account when evaluating the usability of speech based interfaces. 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, where the user visually may inspect the interface to e.g. search for some specific command and in general become familiar with the interface). Furthermore, the input processing in a SDS (speech recognition and -understanding) is comparatively much more complicated and error-prone than most others (Dybkjær and Bernsen 2000).

Therefore, special attention must be paid to the usability attributes pertaining to these issues (i.e. transparency, learnability, error handling, user control, etc.), when evaluating the usability of spoken interfaces. One consequence is that the use of standardised scales such as the Questionnaire for User Interaction Satisfaction (QUIS, see Chin et al 1988) and Software Usability Measurement Inventory (SUMI - see Kirakowski 2003a) becomes questionable - at least 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. The matter will be even more complicated when taking multi-modal and/or Web interfaces into consideration. This is discussed in more detail in section 5.2.

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, p.7)

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."* (ibid)

2.3 Spoken Dialogue Systems issues

This section contains a short-list of the most important factors when evaluating the usability of SDS. It is rather brief, partly because the main focus of the current work is on field trials and partly because discussions about the selection of users, definitions of tasks to be measured, the environment, etc. are discussed in more detail in Chapter 3 and in (Larsen 1996 and Larsen 1998), also appearing in Appendix C, page 111 ff.

A very large body of literature with standards, recommendations and guidelines about how to attack the problem of establishing the usability of a given product exists. Both for the general case and for specialised applications, users, environments, etc. Some examples are Nielsen 1993, ETSI 1993, Rubin 1994, Shneiderman 1998, ISO 1998b, Preece et al 2002. However, instead of going into a discussion about the generic case, the attention will be focused directly at the relevant issues for the case of SDSs. Many of these are derived from the generic case anyway. As Dybkjær and Bernsen note:

> *“In general terms, a usable SLDS must satisfy user needs which are similar to those which must be satisfied by other interactive systems.“*(Dybkjær and Bernsen 2000, p.245)

2.3.1 Methods for Usability Evaluation of Spoken Dialogues Systems

In 1997, the EAGLES (Expert Advisory Group on Language Engineering Standards) group published a handbook of standards and resources for spoken language systems (Gibbon et al 1997). In the section addressing evaluation of interactive systems a list is drawn up, built on literature published to that date. More recently, (Dybkjær and Bernsen 2000) review and update the list based on the EU DISC project (DISC 2000).

EAGLES asserts that the type of the evaluation must depend on the test environment (e.g. laboratory vs. field tests) and whether the evaluation is a “glass box” (assessment of individual components) or a “black box” (overall systems behaviour as experienced by the user) test. From a user point of view, the black box evaluation is naturally the only concern, whereas the system designer will of course be very interested in the performance of individual components, such as the speech recogniser.

Likewise, Dybkjær and Bernsen distinguish between “Diagnostic” (i.e. glass box) and “Performance” (i.e. black box) evaluation. They also introduce the term “Adequacy evaluation” - how well the system fits its purpose and meets actual user needs and expectations (Dybkjær and Bernsen 2000). This definition of adequacy resembles Nilsens and ETSI’s “Utility”. They claim that the adequacy evaluation is the most important from a usability point of view. However, according to Nielsen's taxonomy, (see Figure 2 on page 7) adequacy is not really an aspect of usability.

A review of the literature on usability testing quickly reveals a very high degree of agreement about a common short-list of requirements. The focus must be on early and continuous evaluation. This means that evaluation must also be performed in the requirements gathering/specification phase, as well as throughout the design process. The goal is ultimately to ensure that the specifications are likely to satisfy the user needs, and hence to avoid design errors (and subsequent redesign at additional cost and time). Evaluation at this stage can be difficult to carry out in a formalised manner. See e.g. (Nielsen 1993 and Rubin 1994). A number of methods are often mentioned in addition to end-user testing, for example: Focus groups - 4-10 prospective users review design ideas in a common discussion. Focus groups are also known from marketing and is most often used in the early phases of design. A Cognitive Walk-through is a test used in the early stage in the development. Test persons are asked individually to go through the design and make projections about how this will work out in the final version.

Heuristic Evaluation is perhaps the most well-known representative of these methods, although it is not necessarily carried out only in the early phases of development. Proposed by Jakob Nielsen (see e.g. Nielsen 1993), it is not a user test per se, since no real users are involved. Instead domain or usability experts each review the design and draws up a common list of potential problems. It is also referred to as "discount usability evaluation", due to the avoidance of large-scale user testing.

In the case of SDS, the **Wizard-of-Oz** (or WOZ in short) method is commonly used in the early phases of the design (see e.g. Gibbon et al 1997, Larsen 1998). It is a simulation technique, where a person (the wizard) emulates the behaviour of either parts of or the full system. Most often the wizard carries out the task of the speech recognition and simply types the users speech into the system. However in the early phases of the system development, the wizard may emulate more or all parts of the system (Dybkjær et al 1996).

The purpose of WOZ is - as with the other methods briefly described above - to get a first impression of the intended design, before important and costly design decisions and implementation is begun. In particular, issues related to the expected linguistic behaviour (vocabulary, grammar, discourse phenomena, etc.) or the preferred sequence of tasks of the intended users can hardly be investigated in any other way. Furthermore, the users' goal-seeking strategies, extra-linguistic behaviour (e.g. use of intonation and pauses), preference for modalities (in the case of multi modal systems), etc. can be established by

WOZ-simulations. WOZ is often used in an iterative manner, where the dialogue design is gradually improved over a number of design-test cycles.

2.3.2 User tests

Testing with “real users” i.e. persons representative of the intended population of end users remains the most important source of information, when evaluating the usability of speech based systems. For example, Dybkjær and Bernsen, state:

> *“... representative users from the target user groups must be involved in evaluation from early on,... there is no alternative to involving the target users in all or most system evaluation phases and for most evaluation purposes”* (Dybkjær and Bernsen 2000, p.267)

This viewpoint is supported by most researchers. User tests can be conducted in a variety of ways. An important factor is how close the test situation resembles a later real-life situation in terms of test users, environment, tasks, etc. An important distinction is whether a test is conducted “in the field” i.e. in an environment intended to emulate the real world conditions in which the system will later be deployed. This type of test is denoted a **field test**, in contrast to a **laboratory test**, where the test is conducted in an artificial laboratory environment.

Both types of tests might employ users recruited to match the demographic[1] distribution of the intended target users. However, a lab test will always be a more “artificial” situation for the test users compared to a field, as the environment will be unknown to the users, and they might experience stress and perform differently than in a relaxed atmosphere (Rubin 1994). Furthermore, there is a risk that the person conducting the test (the experimenter) might greatly influence the outcome, as is shown in an experiment by Nielsen (Nielsen 1993). The advantages of lab tests are that the experimenter has full control over the test environment, data recording, etc.

In contrast, a field test is conducted in the test users’ normal environment, and thus removes these potential stress factors. Furthermore, there is a greater chance of discovering additional “external” factors (e.g. interruptions of the test user by a ringing phone, other people present, background noise, etc.), which might in fact turn out to be decisive factors for the acceptability of the system. The disadvantage of the field test situation is that the experimenter might loose (at least some) degree of control of the test situation and data col-

[1] E.g. age, gender, geographical distribution, occupation, level of education, etc.

lection. Furthermore, field tests can be expensive and difficult to set up compared to using a fully equipped usability lab. A common form of field test is known as software **Beta-Tests**. In this case the experimenter has almost no control over the test situation, and the main objective is usually restricted to uncover as many bugs as possible before the final release.

In some cases, lab tests can be made to resemble a field environment very closely. For example, the Sparekassernes Datacentral (a software development centre owned collectively by Danish banks) uses a laboratory setup resembling a bank branch, complete with cashiers desks, terminals, etc.

2.4 Summary

The preceding discussion clearly demonstrates that there are indeed quite diverse opinions about how usability should be defined. Whether this is of great practical consequence is less clear, since everyone agrees that the attributes of usability, utility, adequacy, usefulness, etc. are central and must be measured regardless of how they are placed in the general framework of the overall system acceptability, as Nielsen puts it.

Furthermore, it is evident that usability comprises a duality of subjective and objective measures, and any evaluation will be incomplete without both.

Although often requiring extensive and time-consuming transcription and annotation of corpora, it is fairly straightforward to define and obtain quantitative data for objective measures. This is further elaborated in Chapter 3. In contrast subjective evaluation pertains to the intended users' attitudes and as such cannot be observed directly and requires quite different approaches, as described in Chapter 5.

It has become evident that special attention must be paid to the fact that speech-based interfaces in some important respects are fundamentally different from most other interactive systems. Therefore, special attention must be paid to the problems of the non-persistency of speech and the greater uncertainty in the input processing, as these will greatly influence the impact of e.g. the transparency, error-handling strategies, learnability, etc. of the systems.

Consequently, existing methods must be carefully reviewed and modified, or new must be designed to cope with the usability evaluation of speech-based interfaces.

WOZ simulations are an important way of getting information about potential problems in speech based interfaces at an early stage. Likewise, laboratory- or field tests with test users representative of the target populations must be carried out in the later stages of the development.

The following chapters will discuss the concrete methods for the elicitation of objective and subjective measures for the usability of SDS, illustrated by examples and results from the OVID experiments. This is followed by a combined analysis using the PARADISE method.

Chapter 3 Objective Measures for Spoken Dialogue Systems

This chapter discusses the objective (performance) measures associated with spoken dialogue system evaluations. Less emphasis is put on the theoretical background compared to the subjective measures, because performance measures are more straightforward to understand and are more easily observed or derived from e.g. logfiles and annotation of dialogue corpora than the subjective measures discussed in Chapter 5.

3.1 Definition of Spoken Dialogue Performance Measures

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:

- percentage of correct system answers
- percentage of successful transactions
- percentage of repair utterances
- percentage of user initiated turns
- number of “help” requests
- number of user barge-in’s
- completed tasks and sub tasks
- dialogue or task completion times
- mean user and system response times
- percentage of sentences containing more than one word
- mean length of utterances

For example, Walker and colleagues used elapsed time, system turns, prompt time-outs and the mean speech concept recognition score. The Kappa coefficient was used to compensate for complexity when calculating task success rates in (Walker et al 1998). See Chapter 7 on page 61 ff. In the OVID experiment speech concept recognition rates are measured together with dialogue (sub)task duration, number of turns, (relative) number of user initiatives and task completion rates.

Historically, SDS evaluation using performance measures was first used in the early 1990ies. In Europe, (Simpson and Fraser 1993) reported on the evaluation of the Sundial project. This was followed up by (Danieli and Gerbino 1995), and EAGLES (Gibbon et al 1997). In the U.S. similar measures were developed in the DARPA ATIS programme (Price et al 1992, Polifroni et al 1992), and more recently the PARADISE scheme was used in the DARPA Communicator project for evaluation (Walker et al 2001b). However, rather than presenting a long list of individual measures, it is more relevant to categorise the measures into broader classes, according to the usability attributes of SDS they are actually measuring.

Structuring of Performance Measures. The list shown above comprises an unstructured list of measures in the sense that it does not explicitly refer to the particular aspect of dialogue performance each parameter actually measures.

A division of the measures into more general aspects of performance is desirable to achieve a better understanding of the purpose of each measure. Rather than basing the characterisation directly on the generic usability attributes defined in the previous chapter, the following discussion is based on a taxonomy for the quality of service for SDS, proposed by Sebastian Möller (Möller 2002). Möller suggested the following categories for the classification of dialogue performance measures[1]:

- **Dialogue Cooperativity.** Inspired by the Gricean Maxims (see below), this category covers aspects of informativeness, truth and evidence, manner and relevance.
- **Dialogue Symmetry.** Dialogue initiative and interaction control.
- **Speech I/O quality**. Performance of the speech input/output processing devices.
- **Communication Efficiency**. Speed, conciseness, smoothness of the interaction
- **Task and Service Efficiency**. Task success rates

Using this taxonomy it is now possible to assign each measure to one or more of the categories in a structured manner. Note that in many cases a given measure may belong to more than one category. For example, if the user interrupts the system (barges in), this can be classified as a dialogue control issue (the user takes the initiative), but at the same time it might also be categorised as an event relating to the "Relevance" cooperativity maxim (i.e. the system utterance is not relevant, hence the user interrupts it).

3.1.1 Dialogue Cooperativity Measures

The cooperativity category requires special attention, since it might not be perceived as straight forward as the remaining categories. In 1967 H.P. Grice (published in Grice 1975) defined cooperativity for (human-human) conversations by the four concepts: Quantity, Quality, Relation and Manner (clearly inspired by Kant).

[1] (Möller 2002) proposes more categories than mentioned below, but these mostly refer to subjective measures, which are treated in Chapter 5 and therefore not included here.

From these he formulated a number of maxims, that must be obeyed by the cooperative speaker. The most widely quoted Maxims[1] are (slightly paraphrased):

- Quantity: Make your contribution as informative as required (but not more).
- Quality: Be truthful. Only provide information you are certain is correct.
- Relation: Be relevant.
- Manner: Be perspicuous (clear). Avoid obscurity and ambiguity, be brief and orderly.

See also (Jurafsky and Martin 2000). (Danieli and Gerbino 1995) and later (Bernsen et al 1998) suggested that Grice's Maxims could be applied as central design and evaluation criteria for SDS. The reason that the maxims are important for SDS is that, according to Grice, the maxims govern human-human conversation and allow us to infer meaning from utterances. Humans (implicitly) expect their dialogue partner to be cooperative and adhere to the maxims, and consequently also expect an SDS to behave in a similar manner. An example of how we use the Gricean maxims is given below (adapted from Jurafsky and Martin 2000):

Question: "Are there any flights leaving for Copenhagen soon?"
Answer: "Yes, there are two flights within the next hour"[2]

The maxim of quantity allows us to conclude that there are two and only two (and not e.g. five) flights leaving Aalborg. Formally, the utterance would still be true, if there were five flights, but assuming the quantity maxim, most humans would infer that the number is no more than two[3]. Similarly, the answer "yes" would also be truthful, but not informative (the quantity of information is too low). Most humans will therefore regard the answer for uncooperative. The relevance and manner maxims dictate that only flights bound for Copenhagen leaving in the near future (here judged to be an hour) are included in the answer. Finally, the maxim of quality allows us to assume that the answer is truthful and can be relied upon.

[1] Grice suggested more maxims than mentioned here, e.g., "Be Polite".

[2] This example is actually not "good dialogue design" for human-machine dialogues, since it does not contain an implicit confirmation of the destination.

[3] Quantity does not in this example refer to the quantity (number) of flights, it is the quantity of information given that is meant.

The following list gives examples of dialogue cooperativity measures that have been used in the literature. Note that the references has been moved to footnotes to maintain readability:

- Quantity:
    - Number of words per system utterance[1]
- Quality:
    - Proportion of user questions that was correctly-, incorrectly-, failed to be-, or wrongly answered[2]
    - The DARPA ATIS score (difference between percentage of correct and wrong answers)[2]
    - Appropriateness of the system utterance[3]
- Relevance:
    - Number of user barge-ins[4]
- Manner:
    - Number of words per system utterance[1]
    - number of system turns[4]

The cooperativity measures has not yet reached widespread use for evaluation purposes, but will no doubt become more important in the future, especially if or when dialogue design will be more explicitly based on the Gricean maxims.

3.1.2 Dialogue control and -efficiency, speech input and output quality

The following categories have been more widely used than the cooperativity maxims for SDS evaluation:

- Dialogue Control and Symmetry issues:
    - Number of user and system questions in the dialogue[2]
    - Number of user and system words per utterance[1]
    - User Initiatives[5]
    - User and system correction rates[6]
    - Number of Timed-out prompts[4]
    - Number of user barge-ins[4]

[1] Used by e.g. (Dybkjær et al 1996)

[2] Used by e.g. (Polifroni et al 1992)

[3] Used by e.g. (Simpson and Fraser 1993, Gibbon et al 1997)

[4] Used by e.g. (Walker et al 1998)

[5] Used in (Larsen 1998) and Chapter 4

[6] The ratio of user and system turns concerned with the correction of some problem in the dialogue. See e.g. (Simpson and Fraser 1993 and Danieli and Gerbino 1995)

- Communication efficiency, speed
  - Duration of system and user turns [1]
  - System and user response delays[2]
  - Dialogue and subtask duration [2,3,4]
  - Time-out prompts[3]
  - Implicit recovery[5]
  - Number of turns in dialogue and subtasks[2,3,5]
- Task Efficiency
  - Dialogue and task success rates[2,3,4,5,6]

Speech input quality does strictly speaking not belong to the category of black box dialogue measures, since it is not directly observable to users. However, the effects of the speech recognition performance will be evident in inappropriate, confusing or wrong system answers. Furthermore, the speech recogniser is still the most crucial module in an SDS, and some representation of the speech input performance is always measured.

- Speech input quality
  - Word and sentence accuracy/error rates[6]
  - Speech concept or speech understanding accuracy[3,4]
  - Number of speech recogniser rejections[3]

The list shown above is not exhaustive, but gives an impression of the nature of the measures. Clearly, some of the measures are redundant or at least highly correlated, such as task completion times and number of turns per task, or word, sentence and speech concept recognition accuracy.

Two additional aspects of human-machine spoken dialogues are not included in the list above. One is the quality of the spoken output. Contrary to speech input, (and directly observable to the users) speech output quality has not to date been a main concern in the evaluation of SDS, but has been considered a task for the speech synthesis community. However, users are often asked for their subjective evaluation of the system's voice (synthetic or, more often, prerecorded human speech). This is therefore included in Chapter 5 on subjective measures. CCIR at Edinburgh University has conducted a range of experiments to determine the users attitudes to aspects of the spoken output, see e.g. (CCIR 2003a, CCIR 2003b).

[1] Used by e.g. (Dybkjær et al 1996)
[2] Used by e.g. (Polifroni et al 1992)
[3] Used by e.g. (Walker et al 1998)
[4] Used in (Larsen 1998) and Chapter 4
[5] Used by e.g. (Danieli and Gerbino 1995)
[6] Used by e.g. (Simpson and Fraser 1993, Gibbon et al 1997)

The other aspect is the measures of meta-communication (i.e. communication about the dialogue itself) in spoken dialogues. The most obvious examples are explicit help requests, “go back” and “cancel” commands, etc. In some sense, meta communication is undesirable since it does not directly contribute to the fulfilment of the task goals, but as discussed in the previous chapter, speech based interfaces encounters some special problems requiring measures of this sort.

(Bernsen et al 1998) treat meta communication as an aspect of dialogue cooperativity, which indeed it can be regarded as. Some performance measures relating to meta communication are:

- User help requests[1]
- System error messages[2]
- User cancel attempts[1]

3.2 Recording Spoken Dialogue Performance Parameters

As mentioned in the introduction to this chapter, the performance measures are often directly observable, or at least fairly straightforward to extract or infer from the system log files. This is in particular true for measures like the time spent in various sub-tasks, the number of user and system turns, barge-ins (if this is handled and recorded by the system), time-outs, etc. Speech recognition accuracy and similar measures can be obtained by transcribing the recorded dialogues. Although still a fairly simple task, this typically requires a substantial amount of manpower. Furthermore, the result will to a certain (often small) degree depend on the transcriber, since human errors are unavoidable.

Other measures require a more detailed analysis and are to some extent the result of a subjective judgement by the evaluators, or at least the evaluation scheme adopted for that particular experiment. This is for example true for measures such as “system response appropriateness”, where it is not always obvious how to evaluate the degree to which an answer is appropriate or not. Consider the following example from the OVID corpus:

User: “How much do I have in my savings and budget accounts?”
System: “The balance of your savings account is 2200 kroner.”

Clearly, the answer is not incorrect, but it is neither (fully) appropriate, since information was provided only for one account.

[1] Used by e.g. Walker et al 1998
[2] Used by e.g. Polifroni et al 1992

Experimenters have tried to reduce or eliminate the influence of human error by developing labelling schemes and building automatic or semi-automatic tools to aid the transcription and categorisation. The purpose is two-fold, since it is desirable in itself to reduce this very time-demanding task. A first step towards this is of course to gain an impression of the scale of the problem. For example, (Polifroni et al 1992) analysed the (dis)agreement between seven evaluators. The task was to evaluate the appropriateness of the system answer, given the user query in the ATIS task. The result is shown in Figure 3 below.

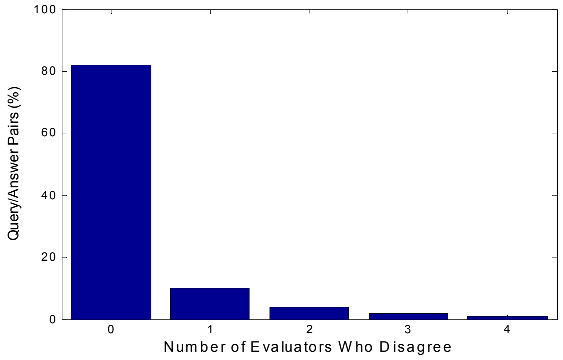


Figure 3. Consistency between seven evaluators during log file evaluation, based on 115 query/answer pairs. Redrawn from Polifroni et al 1992.

In 82% of the cases all evaluators agreed, and one or fewer disagreed in more than 90% of the time. During the analysis of the results, the team built several tools to support the process (Polifroni et al 1992).

Another problem that might occur is the question of determining the goals of the user. This is a requirement in order to derive measures of dialogue and task success. Often (as is also the case for the OVID experiment) test users are presented with one or more strictly defined scenarios, designed to test the SDS functionality in all aspects. In such a case, the user's goals are explicitly defined by the scenario, and it is a straight forward matter to verify whether the user actually reached the objectives. This situation is so common that it for example is a requirement for the PARADISE evaluation scheme, described in Chapter 7.

However, this might not always be true. Since the scenario puts the user in an artificial situation (s/he has no real desire to achieve the goals), some experimenters do not use predefined scenarios. In this case, the evaluators must again judge each dialogue individually, either manually or by some (perhaps semi-) automatic process. This is for example the case in the evaluation of the AdApt system - a multi modal dialogue system within the domain of real estate agents, developed by KTH, Sweden (Hjalmarsson 2002). This experiment did not put any constraints on the user in terms of scenarios. Instead, it was attempted to define a set of rules to identify user goals and determine whether a goal has been met. Following this, the PARADISE paradigm was then applied for evaluation.

This will of course also be the case whenever the evaluation is of users interacting with an SDS in a real-life situation. In this case, the users' goals can never be known beforehand, but must be inferred, either manually or (semi) automatically.

3.3 Summary

This chapter has introduced and discussed the most widely used performance measures for the evaluation of SDS, how they relate to each other and the various aspects of SDS behaviour, as experienced by the user.

The emphasis has been on what (Simpson and Fraser 1993) terms "black box evaluation". Black box evaluation captures the users' viewpoint rather than the developers, since the focus is on the behaviours directly experienced by the user. There are a few exceptions, most notably the measures relating to the performance of the speech recogniser. This reflects the fact that the performance of the speech recognition (and -understanding) module is by far the most crucial single parameter for the overall behaviour of the system.

Each performance measure addresses a very specific issue (like e.g. dialogue duration), which must be the related to the usability attributes (e.g. efficiency) discussed in Chapter 2 for a broader interpretation. This is done using a categorisation suggested by (Möller 2002), into categories relating to Grice's maxims of cooperativity, dialogue efficiency and -control, speech input performance, dialogue control and meta communication.

The issue of obtaining the actual values for the desired measures was discussed. In particular the problems related to the labour-intensive transcription of corpora and disagreement between evaluators were addressed, together with the potential problems of identifying the users goals.

Chapter 4 Performance Measures from the OVID Corpus

The results of the OVID experiments have previously been published, partly in numerous deliverables from the ESPRIT OVID project (a full list is included in Appendix C, page 140) and partly in articles and technical reports (Larsen 1996, Edwards et al 1997, Larsen 1997a, Larsen 1997b, Larsen 1998, Larsen 1999). Except from (Edwards et al 1997) all are included in the present report in Appendix C on page 105 ff.

The purpose of this chapter is threefold. Firstly, to introduce the objectives of the OVID project and describe the experiments that has been carried out to achieve the objectives. Secondly, to provide a quick overview of the outcome of the experiments by giving a condensed account of the results documented in the publications included in Appendix C. Thirdly, to present the results of some additional recent analyses not documented elsewhere.

The results from a more detailed analysis of the questionnaire data are reported in Chapter 6. The Paradise evaluation scheme and the application to the OVID corpus are presented in Chapter 7.

The objectives of the project are summarised in the OVID Technical Annex:

> *"The partners intend to approach the work via a series of controlled usability trials of the software*[1] *in a realistic banking service with real bank customers. The results will be an assessment of how bank customers are able to use the automated service without training in its use, to design an optimal user interface dialogue which can* ***accommodate the untrained user****. ....."The project seeks to carry out a user-centred trial of the application needs in terms of acceptability and usability of speech processing software in interactive voice-response banking systems."* (OVID 1995, p. 3)

Given these objectives, the concrete requirements for the trial service and the experiments can now be defined.

4.1 Specification of Requirements

The requirements for the experimental OVID prototype was determined through two sources: Interviews with personnel from the banks and call cen-

[1] "Software" refers to speech and spoken dialogue processing software.

tres and information from the banks' existing IVR (interactive voice response) services. They are documented in e.g. (Larsen 1996 and Larsen 1998). The purpose of the interviews was - for the Danish part of the project - to get ideas and suggestions from bank staff and to identify potential topics for the prototype. The outcome of this was a prioritised short-list of requirements. Furthermore, personnel experienced in answering customer calls were interviewed to determine how and what customers usually were asking for, when calling the bank. The results of the requirements elicitation are shown below in Table 1 and Table 2.

Rank	Lån & Spar Bank	Royal Bank of Scotland	Barclays Bank
Highest	Convenience	Convenience	Convenience
	24-hour Service[a]	24-hour Service	24-hour Service
	Speed	Speed	Speed
	Security	Operator helpful	Security
	Informative	Security	Confidentiality
	Confidentiality	Confidentiality	Informative
Lowest	Operator helpful	Informative	Operator helpful

Table 1 Rank order for service features

[a] The need for a 24 hour service cannot be documented as the present IVR service closes between 00 and 04 hours.

Rank	Lån & Spar Bank	Royal Bank of Scotland	Barclays Bank
Balance Enquiry	93%	54%	38%
Account Enquiry	42%	43%	22%
Bill Payments	n/a	28%	21%
Transfer Own acc	28%	9%	8%
Transfer 3rd party	8%	n/a	n/a
Transfer Giro	10%	n/a	n/a
Order Statement	2%	1%	2%
Direct Debit	n/a	1%	2%
Exchange Rates	3%	unknown	unknown
Change Password	2%	unknown	unknown
Standing Orders	n/a	n/a	3%
New Check-book	n/a	1%	n/a

Table 2 Transaction Densities for the three banks.

Table 1 shows a high degree of agreement among the bank customers across the three banks, when asked their preferences regarding automated banking services. Table 2 shows the relative distribution of the twelve most required transactions for the three banks' call centres (the British banks) and IVR service (the Danish bank).

Based on this, the specifications of the prototype were determined.

Requirements for the OVID prototype. The overall requirements of the OVID prototype were determined on the basis of the requirements gathering phase to be:

- The user must be able to speak naturally to the system, i.e. a natural, spontaneous speaking style is required, and no explicit vocabulary must be imposed on the user.
- The user must be in control of the communication. This means that the user should be free to request any information or give any command (within the capability of the system) that he may wish at any stage in the dialogue. The system must provide guidance in the case of mistakes or to point out the options to the user.

Based on Table 1 the concrete functionality of the service was chosen to be:

- The customer identification and verification procedure must be compatible to the one used today by Danish banks (as required by Danish legislation).
- The service will provide information of the balance and latest movements of three named user accounts (c.f. Table 2).

More details about this, e.g. specifications of prompting styles, help functionality, etc. can be found in (Larsen 1996 and Larsen 1998). The requirements led to the overall dialogue task structure shown in the (simplified) diagram in Figure 4.

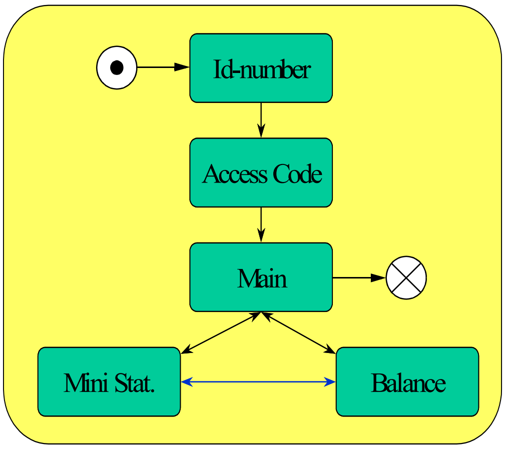


Figure 4. Overall Dialogue task structure. The white circles denote the initial and final states.

The green boxes represent the sub tasks and the arrows represent transitions between the tasks. Hence, the user has to complete the Id- and Access code tasks before progressing to the remaining ones. The dialogue is entered via the Id-number task and exited via the Main task.

4.2 The OVID Trials

The OVID experiments consisted of two trials. First a WOZ trial were carried out with a limited number of users. Based on the results from this, the main field trial was designed.

4.2.1 The WOZ Trial

The purpose of the WOZ trial was to verify the dialogue task structure shown in Figure 4 and identify the application vocabulary. Furthermore, the usability questionnaire were used in the test in order to verify the translation into Danish and collect comments from the test subjects. Two WOZ iterations were done, the first with 7 users and the second with 20. As a result of the WOZ trial the dialogue model was adjusted to achieve a more user driven interaction style (a mixed-initiative dialogue model, see Larsen 1997b). The dialogues were logged, transcribed and analysed. The users commented that they would like the possibility to barge-in (it was not possible) and to use DTMF-keys to input e.g. PIN-codes. More details can be found in (Larsen 1998) in Appendix C. Due to technological constraints, handling of barge-in could not be provided, but support of data entry using the DTMF keypad was implemented.

4.3 The Field Trial

Demographic distribution. The field trial was based on the results from the WOZ-trials. 1100 potential test users among Lån & Spar Bank's customers were contacted. They were selected to cover the demographics laid down in the requirements gathering phase. The trial test must include users from four regions of Denmark (to verify robustness against different accents), five age groups (from 18 to "above 60" years of age) and gender. There were no requirements for occupation or educational level. A surprisingly large number (close to 340) responded positively and were sent a letter containing instructions and the questionnaire. The trial was carried out in June 1997 by 310 users who completed the test and returned the questionnaire.

The demographic distribution of the distribution of test users who performed the test and returned the questionnaire are shown in Figure 5 below.

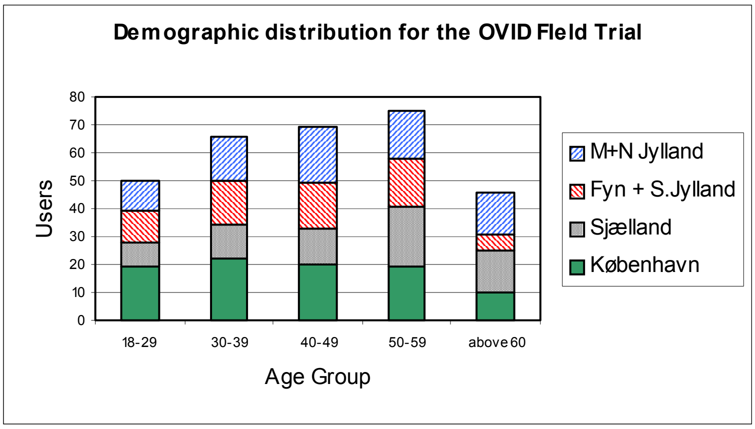


Figure 5. The distribution of the OVID field trial users according to age and region.

142 (46%) women and 168 (54%) men participated in the trial. As can be observed from Figure 5, a fairly even distribution was achieved among the age groups (from 15% to 24%). The largest group geographically was from Copenhagen (29%) and the smallest from Funen and Southern Jutland (21%). The groups are sufficiently large to calculate the statistical significance of demographic differences for the performance measures and user attitudes. However, within a 95% confidence interval, no differences in either performance or user attitudes could be detected and therefore the demographics are not considered further. Additional diagrams showing the various demographic breakdowns can be found in Appendix C, page 163 ff.

Test scenarios. Each user was given two scenarios to carry out, denoted scenario A and B. The scenarios were designed in such a way that all users must perform all subtasks at least once. Furthermore, scenario B was designed in such a way that the users could complete the required tasks faster and with less turns by taking the initiative in the dialogue. This was done in order to investigate to what degree users were actually willing to do so, even without having been given explicit information about the possibility in the test instructions. To avoid bias, half of the users were required to carry out scenario A first and then B. The other half in the opposite order.

As a result, the OVID corpus comprises more than 700 transcribed dialogues from the 310 test users.

The dialogues were transcribed and analysed with regard to the performance measures described in Chapter 3, as well as the specific OVID requirements described above. The chosen performance measures mostly relate to the dialogue efficiency (time spent and turns per (sub) task). In addition, the usability attributes 'learnability' was addressed by analysing dialogue duration and turn-taking strategies well as the task completion rates. The issue of dialogue control is investigated by recording the proportion of turns, where the user grabs the initiative. This is also an aspect of the 'symmetry' of the interaction. Furthermore, the speech recognition performance was assessed.

In short the recorded performance measures are:

- Speech recognition performance. In order to investigate what level of performance is acceptable to the users.
- Turn-taking strategy. In order to investigate to what extent the users were willing or able to take the initiative in the dialogue (the "user in control requirement"/dialogue symmetry)
- Timing. To analyse the duration of sub-tasks and to determine whether the requirement for speed had been fulfilled. (Efficiency)
- Task success rates
- The issue of user habituation (learnability) was investigated by comparing the above metrics for the users' first and second calls

All the measures in the list was broken down into the demographic categories in order to determine whether the performance measures are dependent on the demographic groups. However, as previously mentioned this was not found to be the case.

4.3.1 Impact of Speech Recognition Performance

Speech recognition was implemented using word and phrase spotting. This implies that the recogniser spots for certain key-words and -phrases instead of recognising the full user utterances[1]. These are mapped into corresponding **speech concepts,** containing the semantic information of the utterances.

[1] See Appendix A p. 89 ff. for an introduction of the basic concepts of SDS

Examples of a sentence containing speech concept (underlined) are:

“My id-number is nine two three six seven oh four”
“Please give me the balance of my cash credit account”

The first example contains one and the second two speech concepts.

Thus, the rates quoted below are for the speech concepts, not the actual speech recognition results. A special case are the digit strings used for the user identification and verification. These are treated as a single speech concept as shown in the example above[1].

Figure 6. shows the speech recognition rates divided into six intervals. The figure shows the proportion of users who experienced a given performance. For example, 8% of the users experienced a performance between 50% and 70% accuracy. On average, users uttered 9 speech concepts per dialogue. This means that a user achieving a speech recognition rate of for example 90% roughly speaking can expect to encounter one error per dialogue. The average (for all users) speech concept recognition rate was 86%.

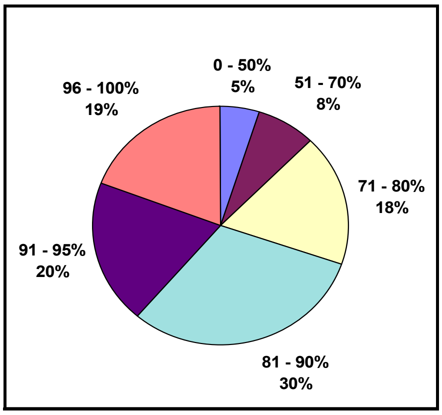


Figure 6. Recognition accuracy (top) and corresponding proportion of users who experienced this rate (bottom)

[1] Although the semantics extracted from the recognised utterances only depends on the speech concepts, recognition of the whole utterance is carried out in most cases anyway.

4.3.2 Analysis of Turn-Taking and Dialogue Initiative Strategies

The results regarding the turn-taking is shown in Table 3 below.

Scenario:	A1	A2	B1	B2
Nominal (expected) number of turns	7		9	
Minimal[a] number of turns	5		4	
Average number of turns	8.4	7.8	7.5	7.1
Average duration of dialogues (seconds)	94	86	92	84
Nominal number of user initiatives[b]	1		2	
Average number of user initiatives	0.8	1.0	1.3	1.6

Table 3 Key figures for scenario A and B

a. This requires the user taking the dialogue initiative whenever possible, and hanging up immediately after the desired information has been obtained, which none of the users did. This can be expected of more experienced users, though.

b. A hang-up is not counted as a user initiative

"A1" and "A2" refers to scenario A carried out as the first or second dialogue. The nominal number of turns is the number of turns required to complete the scenario, if the user answers each system prompt strictly according to the scenario description, and there are no mis-recognitions. The minimal number is achieved when the user fully exploits the possibility for taking the initiative. As can be observed from the table, there is a trend to reduce the number of dialogue turns (and total duration) from the first (A1,B1) dialogue to the second (A2,B2). Note also that the number of observed user initiatives increases, even though the total number of turns decreases. This clearly indicates that users, when performing their second dialogue with the system have begun to learn how to perform the dialogue more efficiently.

Task	A1,B1		A2,B2		Δ
Id number	20.8 secs	22%	17.4 secs	20%	16%
Access code	11.3 secs	12%	9.8 secs	12%	13%
Id + Access code	32.1 secs	33%	27.2 secs	32%	16%
Total dialogue	93.0 secs	100%	85.0 secs	100%	8%

Table 4 Duration of user the authentication procedure. The last column is the reduction from the first to the second dialogues

Table 4 shows the amount of time spent in the “Id” and “Access” sub tasks. The table shows a reduction in time of approximately 10-15%. A paired, two-tailed t-test only revealed significant reduction for the time spent in “Id-number” sub task (p = 0.03), so the reduction can only be regarded as a trend for the remaining figures. Table 4 also shows that the proportion of time spent in the user Id and Access sub-tasks is about a third of the total time. When comparing to the average time spent (approximately one minute) in the corresponding IVR service there is clearly some way to go yet before the requirement for speed can be met. Unfortunately, the information from Lån & Spar Bank about the timing of the IVR-service does not provide any details of the duration of individual tasks, preventing a further comparison.

User Initiatives. 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.

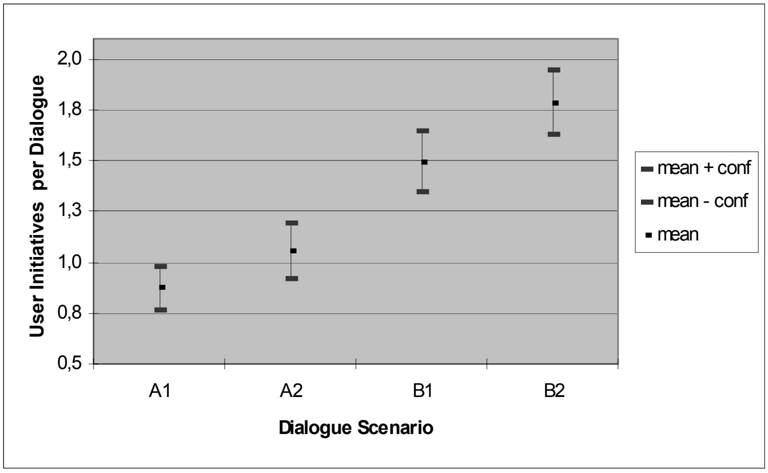


Figure 7. Average number of user initiatives per dialogue with 95% confidence Intervals

Figure 7 shows that users do take the initiative at various points during the dialogue. According to Table 3 there are 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 7. The figure also demonstrates that users tend to take the initiative more often, as 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[1] to the total number of turns for scenario B2 compared to B1.

4.3.3 Analysis of the Dialogue and Task Completion Rates

The task success rates are analysed with regard to perceived versus observed task completion rates. Furthermore, it is investigated whether differences in the task completion rates can be observed for the first and second dialogue. The results are shown in the tables below.

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

Table 5 The proportion of users who succeeded or failed to complete the scenario of their first and second dialogues. Note that a few (7) users only performed one dialogue.

Table 5 shows the result for the task completion rates. A reduction of almost 50% of the failed dialogues can be observed (from 27% to 15%), when comparing the first and second dialogues[2].

In addition, some users seemed not to be aware (or care) whether they in fact did complete the scenarios. When asked to indicate whether they completed none, one, or both of the scenarios, almost all (96%) reported that they had successfully completed both scenarios. However, an inspection of the

Task suc.	Perceived	%	Task suc.	Actual	%
Both	297	96%	Both	230	74%
Only First	3	1%	Only A	24	8%
Only Sec.	8	3%	Only B	33	11%
None	0	0%	None	23	7%
No Answe	2	1%			
Total	310	100%	Total	310	100%

Table 6 The perceived and actual task completion rates

[1] Note that Figure 7 shows the absolute number of user initiatives per dialogue, not the number of user initiatives relative to the total number of turns in the dialogue.

[2] A 25% reduction from the first to the second dialogue completion rates is significant (p= 0.02) at the 95% confidence level.

logfiles revealed that this was not the case. Table 6 shows that only 74% did complete both scenarios. This issue is further discussed in Chapter 7, where it influenced the PARADISE analysis.

A potential explanation could be that the majority of users simply misread or misunderstood the question: “Did you complete the two scenarios?” as: “Did you make the two calls?”. The problem could have been avoided by requiring the users to write down the information they obtained, but this was unfortunately not done for the Danish part of the OVID experiments.

4.4 User Behaviour

As reported in many other field tests (see e.g. Eckert et al 1995) it was found that users behaved differently than expected. Most importantly, users were expected to carry out two scenarios, A and B by making two phone calls to the service. It turned out that many users:

- Made as many calls as needed (up to six in extreme cases) to complete the scenarios
- Carried out both scenarios in one call
- Called the service again, even if both scenarios had already been completed.
- Tried to do a “Turing test” to assess the extent of the systems capability (e.g. by using complex negations or repeatedly asking the same question in different formulations).

In fact, a closer analysis of the logfiles revealed that of the 43 users who called the system 3 times, 7 of them had already completed both scenarios successfully when they did so. Most likely, they just wished to explore some aspect of the system as suggested above or demonstrate it for friends. However, there is no way to determine the cause of the extra calls.

These issues[1] made the automatic analysis of the dialogue corpus more complicated than initially expected and to some (unknown) extent contaminated the results of e.g. the turn-taking and the user attitude measurements. For example, if users had moved on to the second scenario regardless of whether they actually succeeded the first one, it would have been very simple to establish the users intended goal. However, when the user might in fact be trying the same scenario again, the goals must be inferred from the dialogue logfiles. In some cases, a user might only have partly obtained the information

[1] Furthermore, due to an unfortunate runtime error, the system crashed a number of times during the period of testing. This obviously annoyed the users who experienced it and might well have influenced their attitude towards the service. In addition, a crash would erase the corresponding log file, making it impossible to identify which users actually experienced it.

required by the scenario, realise this, and call the service again to obtain the missing piece. While this behaviour makes perfectly sense in a real-world situation, it complicates the process of determining task success rates and other measures.

4.5 Summary

The findings presented in this chapter are discussed in more details in (Larsen 1998) and the related papers included in Appendix C p. 105 ff. The focus here has been on a general presentation of the dialogue corpus supplemented with time and turn-taking analyses. The purpose was to demonstrate to which degree the requirements for user control have been met, and also to determine whether evidence of system learnability can be found, as this was identified as an attribute of particular importance for the system usability (see page 7).

This is clearly the case, since a significant reduction of the time spent in the ID_number sub task reduction of time was observed when comparing durations of the first and second dialogues. Likewise, analyses of the users' turn-taking strategy for the first and second calls reveals a significant increase in the users' tendency to take control of the interaction. Likewise, task completion rates showed a significant increase. All findings are interpreted as signs of system learnability.

As could be expected, the requirement for speed could not be met. Since the test users are recruited among Lån & Spar Bank's customers, who are all very experienced IVR users, a longer period of regular use of the OVID system should have been included as part of the experiment to assess this properly. Furthermore, barge-in capability should also be provided, as this is widely used by experienced users of IVR services, and obviously reduces the duration of IVR transactions.

Chapter 5 Subjective Measures of Spoken Dialogue Systems

This chapter describes the methodologies and problems associated with recording the users' attitudes towards a given system. This is a more complicated process than obtaining performance measures, and therefore a detailed discussion is first presented.

5.1 User Satisfaction Measures

Obtaining and analysing data on user attitudes belongs to the field of psychometrics (Aiken 1996). The term user satisfaction is used to denote the degree to which the users are satisfied or accept the system performance, most often expressed as aspects of usability, as discussed in Chapter 1. The user satisfaction measures are often extracted from a questionnaire, where the users are required to respond to a number of statements related to their experiences of interacting with the system.

In most cases the user is required to express his/her attitude towards a number of statements about the system, for example using a so-called Likert attitude questionnaire (Oppenheim 1966). The user is required to mark one of a number (most often 5, 7 or 9) of predefined check boxes, e.g. labelled from "strongly disagree" to "strongly agree". See Figure 8. on page 41 for an example of a Likert-type statement. Hereby the questionnaire becomes quantifiable and permits various statistical analyses and calculations to be carried out. The Likert and other scales, e.g. semantic opposites are described and discussed in this chapter.

If nothing else is mentioned, the sources are (Oppenheim 1966, Kline 1986, Love et al 1994, Kirakowski 2003a, SUMI 2003).

5.1.1 Obtaining information about user satisfaction

There are several ways to elicit information about the user's attitudes towards a system. The most common are interviews and questionnaires. In the case of the OVID field trial, the user interacts with the system from a remote location via the phone. Consequently, there is no direct contact between the test users and the experimenter and a usability questionnaire is the obvious choice. The following discussion is therefore focused on questionnaires. However, if an interview is strongly structured, it will closely resemble a written questionnaire and many of the characteristics of these will also apply to the interview (Preece et al 2002). Obviously, it would have been possible to e.g. call a selected subset of the test users and interview them further about the test, but the questionnaire data was considered sufficient at that time.

5.1.2 Usability Questionnaires

There exists a multitude of techniques for constructing questionnaires and as mentioned above, a very large body of research has been conducted in this area since the beginning of the 20th century. Two well-known and widely used questionnaires are the Software Usability Measurement Inventory (SUMI) (SUMI 2003) from the Human Factors Research Group, (HFRG) University College Cork in Ireland, and the QUIS questionnaire from the University of Maryland (Chin et al 1988).

A third one is the usability questionnaire used in this work, which was developed by the Centre for Communication Interface Research (CCIR), Edinburgh University together with British Telecom (BT) (Jack et al 1993, Love et al 1994) specifically for evaluation of telephone-based speech services. As mentioned in 2.1, usability questionnaires developed for generic desktop applications might not be valid for speech-based interfaces. Finally, researchers at Brunel University in Great Britain have started to develop a usability evaluation tool for speech based services, but apparently the work has been discontinued (Hone and Graham 2000).

The following section will discuss and compare the development of these in some detail, but first some basic definitions on scales, reliability and validity of questionnaires are given.

5.1.3 Survey questionnaires and checklists

The two key elements when using questionnaires to measure user attitudes is the **reliability** and the **validity** of the questionnaire. If the questionnaire has not been tested and shown to be reliable and valid, there is no guarantee that the results obtained with it actually reflects the true attitudes of the users. According to Kline (Kline 1986), validity can be divided into a number of aspects, the most important being:

- **Face validity**: Face validity simply means that the test appears to be relevant to the task at hand. It does not relate to the construct validity of the test (see below). It is important, however, that the those performing the test, perceive it to be relevant to the task to ensure their cooperativity when answering it.
- Aiken (Aiken 1996) defines **Content Validity** as an elaboration of Face Validity, involving a "*careful systematic analysis of the content of the instrument by experts who are familiar with the variables or constructs purportedly measured by it*" (Aiken 1996, p.90)
- **Construct validity**: The term was suggested by Cronbach in 1955. Construct validity can be interpreted as the extent to which a test actually measures what (i.e. the construct) it was designed to do.

The construct validity (hereafter just referred to as "validity") of a questionnaire might seem obvious at a first glance, but if the questionnaire is not carefully designed and subsequently checked against e.g. other known tests, it might lead to faulty results and conclusions. Content validity (or even face validity) is often mistakenly assumed to guarantee construct validity, but this is far from true, and can only be the case by pure chance.

A number of techniques exists to ensure construct validity. Indeed, as Aiken notes:

> *"The construct validity ..... cannot be determined by means of a single procedure. Various kinds of evidence must be sought - expert opinion, correlations of scores with other measures of the same construct, and comparing the scores of people who obviously have a high amount of the construct with the scores of people who have a low amount of the construct*[1] *".*
> (Aiken 1996, p.240)

Obviously, establishing construct validity is a complicated and difficult process. One example of errors that can lead to serious problems with validity is when test subjects interpret questions in different ways. This means that they are actually answering two (or more) different questions. These will then be treated as one (the test designers' interpretation) in the further analysis and might consequently lead to wrong conclusions (Kirakowski 2003a). The following analysis will show how to avoid this by applying factor analysis and how researchers rely heavily on content validity when validating their questionnaires, especially in the initial phase.

The **reliability** of a test is the degree to which the test will produce consistent or similar results, when performed by similar users in a similar environment (see e.g. Oppenheim 1966, Kline 1986, Kirakowski 2003a). Or, in other words; to what degree the results of the test are reproducible. In statistical terms, reliability is defined as the estimate of the ratio between the sum of observed individual item variability and the "true" variability. An often used coefficient to express reliability is **Cronbachs** α (Kline 1986, Oppenheim 1966, StatSoft 1999). It ranges between zero (unreliable) and one (reliable). See Appendix B, page 95 for the definition of Cronbachs α.

High reliabilities (0.85 or above) for Likert scales can often be achieved (Oppenheim 1966). The term "reliability" is used in the tradition of social sciences and not in terms of analysis of product reliability and lifetime. Since the

[1] This simply means that the test can be validated using people with already known attitudes towards the subject (construct) of the test.

reliability is estimated from the consistency among all statements in a questionnaire, it is also referred to as the *internal-consistency reliability* (StatSoft 1999). See section Chapter 6 p. 51 for a calculation of the reliability of the OVID questionnaire. Reliability can also refer to **test-retest** reliability, where the focus is on the reproducibility of results, e.g. from using the same test at two different points in time.

As a final note of the validity-reliability issue, it is evident that one cannot achieve a high validity without a high reliability. The opposite is not the case, though. Unfortunately, it is indeed possible to consistently produce invalid results.

5.2 Types of scales

There exists a number of popular questionnaire types (or scales). These are briefly described below, with emphasis on the Likert scale, as this is used in the OVID experiments.

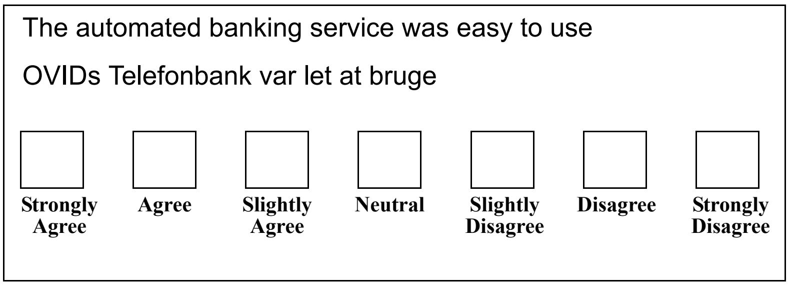
Factual questions are used to obtain concrete information from the test subjects and not to express attitudes towards the system as such. For example to capture demographic information like name, age (interval), gender, experience, occupation, test context, etc. Factual questions are most often placed in screening and pre-test questionnaires.

Closed-ended questions are questions with a strongly limited number of answers, e.g. yes/no statements. Close-ended questions are used to obtain very tightly structured answers from users, which will be comparable across all users, and are therefore easily subjected to summarisation and further statistical analyses. Close-ended questions usually comprise the main part of an attitude questionnaire.

Open-ended questions are used to collect additional information not captured by the closed-ended questions mentioned above. They are important in an exploratory phase, when close-ended questions haven't been finalised, but are more difficult to summarise. Therefore, in the final versions open-ended questions are most often used to collect additional comments from users. This can be very useful, however. For example, in the OVID questionnaire a large proportion of users commented that they did not like to speak their PIN-code in public, an issue not covered by the Likert section of the questionnaire because of its' very domain-specific nature (Larsen 1998).

Opinion scales:

Likert scale. The Likert scale is a well-known and widely used scale. It is named after Rensis Likert, who devised and described the method in 1932 (Oppenheim 1966). A Likert scale is constructed from a set of statements, where the users' attitude is expressed by indicating the extent of (dis)agreement with each statement. SUMI, SASSI and BT-CCIR (see below) are all Likert-type questionnaires. Figure 8 below shows an example from the OVID questionnaire, which consists of 25 Likert style statements.



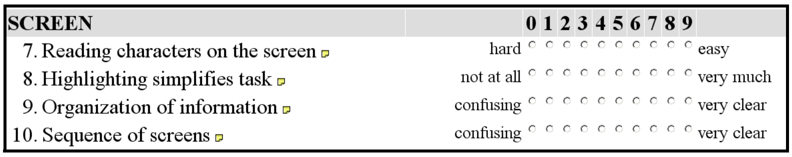
The automated banking service was easy to use

OVIDs Telefonbank var let at bruge

□	□	□	□	□	□	□
Strongly Agree	Agree	Slightly Agree	Neutral	Slightly Disagree	Disagree	Strongly Disagree

Figure 8. Example of Likert scale statement from the OVID questionnaire, the statement is shown in English and Danish (Larsen 1998))

Semantic differentials or adjective opposite pairs. Are together with the Likert scale the most used scale. The QUIS questionnaire is of this type. The user is presented with a number of "semantic opposite" adjectives (e.g. "beautiful - ugly"). Similar to the Likert scale, a number of check-boxes are placed between the two opposites and the user expresses his/her attitude by ticking one. Figure 9 below shows an example from the QUIS questionnaire.



SCREEN		0 1 2 3 4 5 6 7 8 9	
7. Reading characters on the screen ▫	hard	○ ○ ○ ○ ○ ○ ○ ○ ○ ○	easy
8. Highlighting simplifies task ▫	not at all	○ ○ ○ ○ ○ ○ ○ ○ ○ ○	very much
9. Organization of information ▫	confusing	○ ○ ○ ○ ○ ○ ○ ○ ○ ○	very clear
10. Sequence of screens ▫	confusing	○ ○ ○ ○ ○ ○ ○ ○ ○ ○	very clear

Figure 9. Example of adjective opposite pairs from the QUIS questionnaire (Chin et al 1988), from an on-line version (Perlman 2003).

Thurstone and Guttman scales. A Thurstone scale is one that has been calibrated w.r.t. equal appearing intervals. This has the advantage that the "perceptual distance" between each choice in the range of answers is equal. It has some resemblance to the Likert Scale, but requires an extra calibration step when developing it. Guttman questionnaires consist of a collection of statements which gradually get more extreme. The statement where the user begins to answer negatively rather than positively is extracted. These scales are rarely used, partly due to the fact that they require and additional step in the development. Thurstone scales have been shown to produce results comparable to the Likert scale.

A questionnaire is composed by combining one or more of the above mentioned question types. Some researchers argue that the types of questions cannot be mixed, while others (Kirakowski 2003a) are more pragmatic. Frequently, a questionnaire will contain a section of factual questions to e.g. capture the demographics, followed by a section with close-ended opinion questions. Open-ended questions are often placed at the end to capture items that the test designer has neglected or was unable to include in the other sections.

In most cases a questionnaire is composed of about 20-25 individual statements or questions. There are several reasons for this number. To ensure the full attention of the users throughout the process of filling out the questionnaire, only a limited number of questions can be included. This is particularly the case when the test design requires the questionnaire to be filled out after each scenario. In contrast, a high number of items is desirable to ensure a sufficient degree of reliability in the test, (to minimise the variability). About 20-25 questions seem to be generally agreed upon as a reasonable compromise by most researchers (Oppenheim 1966, Kline 1986, Chin et al 1988, Love et al 1994, SUMI 2003).

A number of simple measures can be employed to avoid (or at least reduce) biasing the test subjects. The sequence of questions can be randomized, so every subject answers the questions in a different order. By this procedure, cross-item bias can be avoided or at least reduced. Many researchers mix statements expressing positive and negative attitudes about the tested system in their scale. This is to ensure that the users do not tend to put all marks in one or two columns, but are forced to be more alert and careful in reading each statement. It might also serve to reduce a tendency that some have users to be “too consenting” in their responses. This is sometimes referred to as “social desirability” or “response acquiescence” (Jack et al 1993, Oppenheim 1966). However, there are exceptions to this rule, e.g. the QUIS scale by Shneiderman, Norman and colleagues (Chin et al 1988).

The user’s answers to the statements in all the opinion scales are quantified to enable statistical analysis. Each potential answer or “box” is assigned a number as shown in Figure 9. Following this, values for mean, standard deviation, correlation, etc. can be calculated.

It is common practise to sum the scores for each statement to produce a single summed score characterising this particular user’s attitude towards the system (Oppenheim 1966). Often, this sum is divided by the number of statements to yield an average score, which will then be on a scale identical to the scale for the statements. However, the notion of a single measure representing the user’s attitude towards some issue only makes sense if all the statements in the questionnaire addresses the same underlying issue (the construct). That this is the case can be ensured by applying factoring to analyse the underlying relationships between the statements.

5.2.1 Factor Analysis

A questionnaire must be carefully designed and verified according to the requirements for validity and reliability discussed above. One of the important tools for doing this is an iterative process using factoring, or explorative factor analysis (FA). FA is a combined analytical and statistical method that aims to reduce the number of variables (here statements) and also to identify the underlying relationships (denoted “Factors”) between variables. FA has been extensively used in experimental psychology in the last century. The mathematics of FA closely resembles Principal Components Analysis (PCA) (Darlington 1997, StatSoft 1999, Tabachnick and Fidell 2001). Since the underlying mathematics is close to identical for FA and PCA, there is often some confusion about the techniques and authors adopt different conventions and taxonomies when describing the methods. The basic taxonomy, attributes and parameters of FA are briefly introduced in Appendix B p. 95 ff.

In short, the main difference between FA and PCA lies in the starting point i.e. how the causality is perceived (Tabachnick and Fidell 2001):

- In FA, the factors are perceived as the cause of the observed variable scores, i.e. it is the underlying factor structure that has produced (or caused) the observed variable scores.
- In contrast, for PCA, the components are just perceived as aggregates of the observed variable scores, i.e. the variables cause or produce the component, and there is no underlying theory or need for an interpretation of how the variables are associated with the components
- Furthermore, in PCA all variance is modelled, whereas in FA only the variance the variables have in common (communalities) are considered, while the individual (specific) variance is sought to be minimised.

5.2.2 Factoring and Principal Components Analysis

The objective of explorative FA is to arrive at a small and stable set of statements in the questionnaire, where redundant and/or contradictory statements have been removed (Oppenheim 1966). The factoring process will identify and remove redundant or ambiguous statements by correlating the user’s answers. Furthermore, the underlying relationships (the factors, or principal components in the PCA terminology) are identified and interpreted, assigning meaningful labels to the factors determined in the process.

The factor labels must be determined from the statements that load (i.e. are dependent on them) on the factor in combination with the test designer’s experience and intuition (Love et al 1994, Darlington 1997). This has sometimes had the consequence that researchers analysing the same data has arrived at different numbers of factors and also labelled them differently. It should not be considered and error or a problem with the method, but rather a reflection of different objectives and viewpoints on the side of the researchers. This is the point where PCA and factoring differs. When applying PCA, the objective is to reduce the dimensionality, while retaining as much variance of the original data as desired by the experimenter. The number of required principal components are then chosen to be the minimal number satisfying this criteria. In contrast the objectives of factoring are twofold. To reduce dimensionality (the number of statements), and at the same time to assign a meaningful structure that can provide an "explanation" of the observed data.

Darlington lists four major questions which are sought in factor analysis:

- How many underlying factors are needed to explain the relationships among the items?
- What is the nature of these factors? (i.e. how do they generalise and relate to the items)
- How well do the hypothesised factors explain the observed data?
- How much random variability does each item contain (i.e. how much cannot be explained by the common factors)

Evidently, these goals are partly achieved by statistical analysis and partly by interpretation of the qualitative aspects.

5.3 The QUIS, SUMI, SASSI and BT-CCIR questionnaires

The purpose of this section is to illustrate - through the examples of well-known and widely used questionnaires - the process of developing and validating user attitude questionnaires. The section also serves to document the background of the BT-CCIR questionnaire used in the OVID experiment.

In (Kirakowski 1994/2003c) the iterative development of the SUMI questionnaire is described in great detail. From an initial pool of over 150 items (inspired from previous research), experts grouped them and reduced the number to 75 (i.e. establishing content validity). A test was then carried out with 139 users and factor analysis was used to identify five underlying groups (factors). The ten statements achieving the highest factor loading[1] for each cluster were selected and formed the second iteration questionnaire (with 50 statements). A new test with 143 new users was carried out and verified the findings from the previous iteration. At this point the identified groups of statements (denoted sub-scales) were labelled (Kirakowski 1994/2003c):

- Efficiency
- Affect (Likability)

[1] Factor loadings are the correlations between the factors and the statement scores.

- Helpfulness
- Control
- Learnability

As can be seen, these categories match well-known attributes of usability discussed in Chapter 2. A new study was now carried out that used the questionnaire to two different desktop office applications. The validity of the questionnaire was confirmed by empirically checking the findings against other sources and comparing with the ISO dialogue principles (ISO 1998a). Finally, the number of questions were reduced to the 25 with highest factor loadings. The reliability of the final version of the questionnaire was calculated using Cronbachs α and found to be 0.92. More than 1100 instances of the final version were used in the verification process. In total the development of the SUMI questionnaire took several years to complete.

Quite similar approaches and results can be seen for other usability questionnaires, e.g QUIS (Chin et al 1988, Shneiderman 1998), SASSI (Hone and Graham 2000, 2001) and the literature in general (Oppenheim 1966, Kline 1986, Aiken 1996). The SASSI (Subjective Assessment of Speech System Interfaces) questionnaire is of particular interest, as it is specifically targeted at speech-based interfaces.

The SASSI tool (Hone and Graham 2000, Hone and Graham 2001) is a recent attempt to develop a tool specifically for assessment of speech based interfaces. It uses an approach similar to that adopted in SUMI. An initial pool of 50 statements was used in four studies of eight speech based applications. In total 226 users participated. A factor analysis was carried out and six factors were identified accounting for 65% of the total variance. The six factors were labelled (appearance in order of importance):

- System Response Accuracy
- Likability
- Cognitive Demand
- Annoyance
- Hability[1]
- Speed

The internal consistency reliability was estimated for each sub-scale using Cronbach's Alpha and found to be in the range of 0.7-0.9. The SASSI tool has only completed its first iteration. Unfortunately, it seems that the development of SASSI has been discontinued.

[1] Hability corresponds to 'Visibility' and in some sense also 'Transparence'

The following section describes and discusses the development of the BT-CCIR usability questionnaire used in the OVID tests (Love et al 1994) in more detail, as a slightly different approach was taken here.

5.3.1 The BT-CCIR Usability Questionnaire for interactive telephone based services

This questionnaire was developed with a specific class of systems (namely telephone based spoken interaction) in mind, whereas the SUMI and QUIS questionnaires discussed above were developed for software systems in general. Therefore, The BT-CCIR questionnaire is very focused and includes specific questions related to telephone usage (e.g. statements about the voice and beep-tones). The development also differs in other aspects. From the initial version and onwards it comprises only 20-22 core questions. Furthermore, it accepts that a (limited) number of application specific statements can be added, bringing the total number up to around 25. Since the application specific questions have not been part of the validation process, they can not be included in the subsequent usability analysis, but can only be used to cast light on the individual topics addressed by them.

As in the development of the SUMI questionnaire, the initial set of statements were determined based on a pilot study involving observations and interviews with naive users and a literature review (Dutton et al 1993). An initial set of 22 items were identified and a group of test persons were asked to identify and rank the six statements they found most important. A control group was asked to do the same, but without a list of items to pick from. There was a strong correlation between the choices of the two groups (thus indicating that the identified items is a complete set), and the control group proposed no new items Thus content validity was established.

The initial list of statements was revised and a second experiment was carried out, this time with 154 subjects. The purpose was to obtain a ranking of the importance of the statements. Based on the results, the questionnaire was once more revised, and a final version with 22 core statements had been defined (Dutton et al 1993). The repeated process of revising, ranking and re-testing has served to ensure the validity of the questionnaire.

To further verify the validity of the questionnaire, factor analysis was applied.

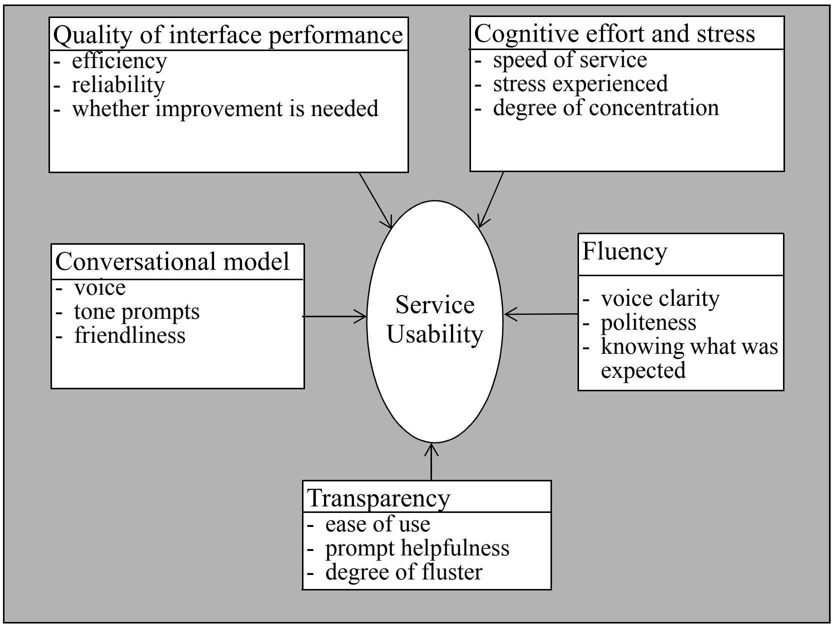


Figure 10. Factor Structure of the BT-CCIR usability questionnaire. Redrawn from (Love et al 1994)

The analysis showed that the items could be divided into a set of five factors, accounting for 74% of the variability of all 22 items. The factors were labelled (Love et al 1994):

- Quality of interface performance (21% of the variance)
- Cognitive effort and stress experienced by the user (17% of the variance)
- User's conversational model[1]
- Fluency of the experience[1]
- Transparence of the interface[1]

As expected the factor labels differ substantially from the one found in SUMI and the general attributes of usability.

5.3.2 Problems with Questionnaires

The use of opinion scales (questionnaires) and the subsequent analysis of them are not without problems. In particular, the notion of computing an average score and assuming this one, single index to represent "the user's atti-

[1] Only the percentages for the two most highly ranking factors were reported in (Love et al 1994)

tude" is problematic. One problem is of course that this single score (computed as the average of the user's answers to all statements) can be generated in a virtually infinite number of combinations of the individual answers. Thus, two users with identical average scores might possibly have very differing attitudes towards the individual statements. However, checking for internal consistency (e.g. by calculating Cronbachs α) can alleviate this problem somewhat.

Another problem has to do with the interpretation of the overall index: "Can usability be expressed as a single index?" In Chapter 2 and 3 much emphasis was put on the fact that many different attributes together constitute the concept of usability.

Furthermore, in the examples given earlier in this chapter, factoring was applied to identify the underlying relationships and these were labelled according to the usability attributes they refer to. This seems to contradict the idea of a single overall index, which has also been subject to criticism. For example, Oppenheim observes that: "*Often, for this reason, the pattern of the responses becomes more interesting than the total score."* (Oppenheim 1966, p.200).

Discussing subjective evaluation of usability, Möller states:

> *"The problem obviously turned out to be multi–dimensional. Nevertheless, many other researchers still try to estimate "overall system quality", "usability" or "user satisfaction" by simply calculating the arithmetic mean over several user ratings on topics as different as perceived TTS quality, perceived system understanding, and expected future use of the system.* (Möller 2002, p. 1)

In this case Möller was arguing for the need of a taxonomy for the quality of SDS, but his objection is nevertheless worth consideration.

In addition, there is a problem with the assumptions of the statistical methods usually employed in the analysis of questionnaire data. It comes from the fact that the quantification of the scale is usually done by mapping the categories directly to natural numbers, either starting from one or symmetric around zero. This transformation conceals the fact that the data is not continuous, but ordinal.

However, it is common practise to map the ordinal categories like "Strongly Agree" to some number, and treat them as continuous variables, which can be subjected to parametric analysis. For example, Tabachnick and Fidell observe:

> *"In practise, we often treat variables as if they are continuous when the underlying scale is thought to be continuous but the measured scale actually is ordinal, the number of categories large - say seven or more - and the data meet the other assumptions of the analysis"*(Tabachnick and Fidell 2001 p.7)

By "other assumptions" is meant e.g. normality or linearity.

5.4 Summary

Table 7 below summarises and compares the development of the usability questionnaires described above, along with the QUIS questionnaire from the University of Maryland.

	QUIS Chin et al 1988	**SUMI** Kirakowski 1994/ 2003c	**BT-CCIR** Love et al 1994	**SASSI** Hone and Graham 2000
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	Previous research and experience, literature 90 statements, 5 overall and 85 specific, divided into 20 sub groups.	Previous research and experience, literature 150 statements, grouped and reduced to 75 Tested on 139 subjects	Previous research and experience, literature 22 statements in core set, 3 application specific Validated by 20 experts and 20 users in control group	Previous research and experience, literature 50 statements. Data collected from 226 users across 4 studies of 8 applications. 6 sub-scales identified, with sub-scale reliability in the range of 0.7-0.9 (Cronbachs Alpha)
Second	Total of 110 statements. Tested for reliability: (Cronbach's alpha 0.94) by 213 users.	50 statements, based on initial version. Tested on 143 users. 5 groups identified by factoring	22 core revised statements, based on initial version. 5 sub groups, identified by factoring	
Third	70 statements, identified by factoring. Reliability is 0.89 (Cronbach) 150 users	25 statements, 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	
Fourth	Version 5 and 5.5. Includes 6 general + 22 specific statements. Version. 5.5 is an on-line version			

Table 7 Comparison of the iterative development process for the SUMI, QUIS, BT-CCIR and SASSI questionnaires

It is evident from the table that the development of a scale is a very resource demanding process. Especially establishing the validity of a scale is difficult and requires expertise and resources. (Tabachnick and Fidell 2001) recommends that 300 cases (i.e. questionnaires) should generally be used for FA, except in cases where strong reliable correlations can be ascertained.

Unfortunately, with the two exceptions discussed above, researchers in speech technology do not seem to realise that this is a necessary task that must be undertaken to obtain valid results. A scale, like any other measuring instrument must be carefully designed, documented and validated, if the measurements are to be scientifically valid. For example, even though there are numerous articles documenting the PARADISE scheme, no validation of the questionnaire used to obtain the subjective measures has yet been published (to the authors knowledge). Hone and Graham review a number of subjective speech system evaluations and conclude that:

> *"It can be concluded that none of the existing techniques for subjective speech interface meet the criteria for a valid psychometric instrument"* Hone and Graham 2000.

This is a very strong statement. As demonstrated above, documentation of the BT-CCIR scale validation has been published in (Jack et al 1993) and (Love et al 1994) and used by CCIR in a large number of usability studies of speech controlled systems apart from the OVID experiment, see (CCIR 2003).

Chapter 6 User Attitude Measures on the OVID corpus

This chapter documents how the BT-CCIR questionnaire was used for the measurements of the users' attitudes in the OVID field trial. Factoring is applied in order to verify the Danish translation of the questionnaire and for comparison with the original British version.

Given the objectives and resources of the original OVID project, the BT-CCIR usability questionnaire was used for the Danish OVID field trials 'as is' - except from a translation into Danish and the addition of five domain specific questions.

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 technology 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. See (Larsen 1998).

Based on this, adjustments were done and the questionnaire was used in the OVID field trial. The results were analysed for the demographic groups (age, gender and region), as described in Chapter 4. However, no statistically significant differences were found, so this will not be discussed further here.

6.1 Results from the User Attitude Questionnaire

The overall results of the user attitude questionnaire is shown in Figure 11 below. The diagram shows the user attitude for each statement with confidence intervals, averaged over all test users.

The following analyses of the OVID results are based only on the set of the twenty core items that have been validated, as described in the previous chapter. The core set are the 20 first questions shown on the bar diagram (counted from the top of Figure 11), and the domain-specific statements are the five last, followed by a bar illustrating the overall average. 98% confidence intervals for each statement are shown in red/yellow colours.

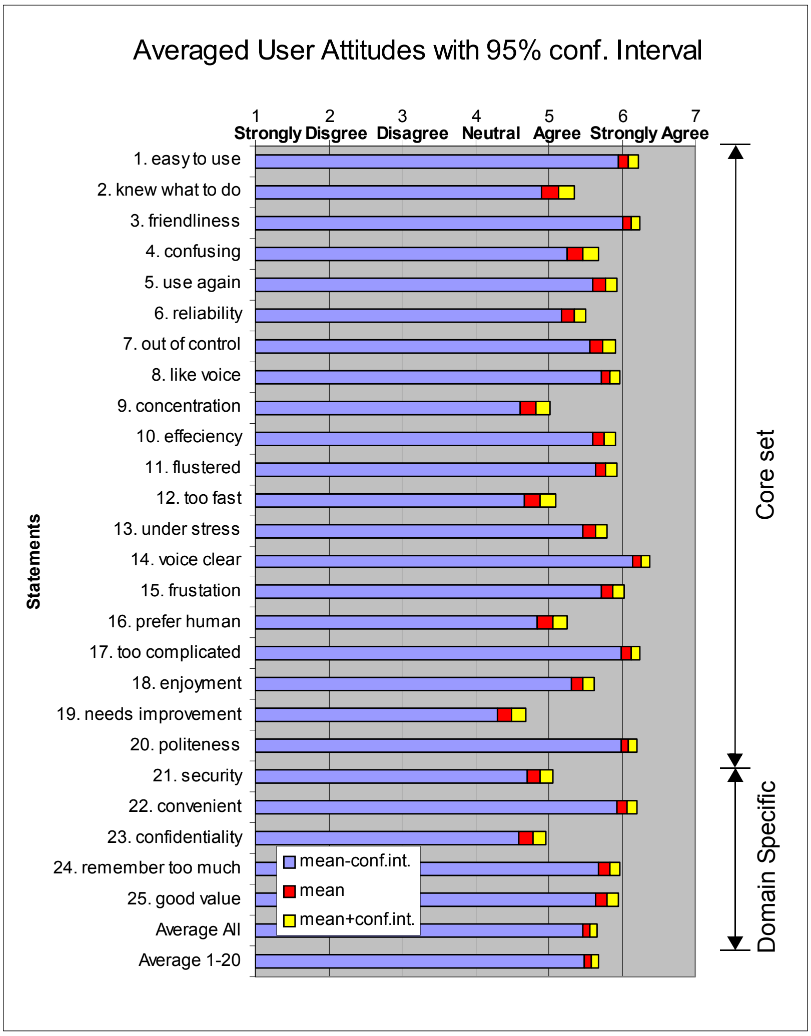


Figure 11. Overall user attitudes with confidence intervals

A seven-point Likert Scale was used in the questionnaire. The categories are assigned the numbers 1 to 7, as can be observed at the top of the diagram. In order to compare the values for the individual statements, a reversed scale was used for negative statements[1].

[1] For a more detailed discussion of this technique as well as the full wording (in Danish and English) of the Likert Statements see (Larsen 1998) in Appendix C, page 159 ff.

Therefore, a positive attitude will always be represented by a high score.

Note that the five last statements have been added specifically for the OVID field trials and are not part of the set of core statements. In general, the user responses are positive and an overall average score (for all users and all core statements) of 5.6 with a standard deviation of 0.87 was achieved. The 95% confidence bands are in the order of +/– 0.2. A number of graphs showing a detailed breakdown of the scores in Figure 11 into the demographic groups shown in Figure 5 on page 29 can be found in Appendix C, p. page 167 ff.

Unfortunately, an error in the test design was later discovered. The individual statements in a questionnaire can potentially cause the answer to one statement to influence the following one. Therefore, the order is usually randomized to insure against this sort of bias. However, this was not done in the OVID experiment, so all test users got exactly the same questionnaire, potentially causing a bias in the responses.

6.2 Validation of the OVID questionnaire using Factor Analysis

Before factor analysis was applied to the OVID corpus, the Cronbach Alpha coefficient was estimated to 0.92, which is satisfactory and indicates that there exists a strong internal consistency among the statements.

A series of factor analyses was then carried out using the MATLAB Statistics Toolbox (Mathworks 1999). The purpose of the initial analysis was to identify the number of factors best describing the data. As described in 5.2.1, factor analysis is a combination of mathematical analysis and subjective interpretation and evaluation. Factor analysis was carried out with between three and seven factors, and a factor structure with five factors was identified as the one best matching the data. This is documented in Appendix B p. 95 ff.

The optimal number of factors is determined as a compromise of the requirement to maximise the explained variance (as in PCA) and the need of obtaining a factor structure, capable of providing a qualitatively satisfactory description of the clustering of the statements. The ideal factor structure is one where each statement only belongs (loads on) one factor and the identified factors are independent (orthogonal). To achieve this, rotation is applied to the factors. If the rotation is orthogonal, the factors will be unrelated. Some heuristics[1] are usually employed to aid this:

[1] See e.g. (Hone and Graham 2000)

- As a rule, a statements' loading on a factor should be above 0.3 to 0.4 for the statement to be assigned to the cluster characterised by the factor.
- Ideally, all statements should only be assigned to one factor cluster. If a statement loads on several factors it is said to be cross-loading. A common criteria for cross-loading is that the difference between the statements' loading on two factors is less than 0.2

Initially a PCA was carried out to get an overview of the statistical properties. The result is shown in Figure 12 below. The figure shows the principal components and the cumulative variance curve.

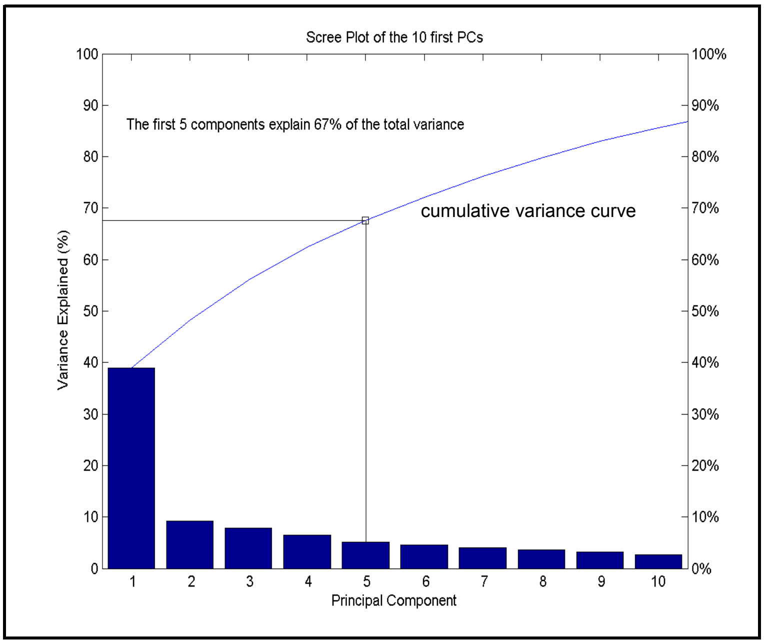


Figure 12. Scree plot (columns) and accumulated variance (curve) for a PCA of the OVID questionnaire data.

As can be observed from the Scree plot of the principal components, the first component is very dominant and the five first components together captures 67% of the total variance. The first 10 components explains close to 90% of the total variance.

Following the initial PCA, a factor analysis with the number of factors set to five produced the factors shown in Table 8

Note that the accumulated explained variance has dropped from 67% in Figure 12 to 57% in Table 8, while the variance is more evenly distributed among the individual factors (the first factor explains only 19% compared to the first principal component’s close to 40%). This difference is due to the constraints the qualitative requirements put on the factor structure. Table 8 also shows Cronbachs α for each factor sub-scale. As could be expected (since the number of items is smaller for the sub-scales), α is also lower for the sub-scales, but a high degree of consistency can still be observed.

OVID Factors	**Var[a]/α[b]**	**CCIR Factors**	**Var**
F1: Quality of interface/ performance Use Again Reliability Efficiency prefer Human Enjoyment Needs Improvement	**19%** **0.86**	**F1:Quality of interface performance** Use Again Efficiency Reliability Needs Improvement	**21%**
F2: Cognitive load Concentration Too fast Under Stress	**13%** **0.83**	**F2: Cognitive effort and stress** Speed of service Stress experienced Degree of concentration Perceived control	**17%**
F3: Control/Confusion Know what was expected Out of control Confusion Flustered Too Complicated	**9%** **0.78**	**F3: Conversational model** Voice Tone prompts Friendliness	**N/A**
F4: Friendliness Friendly Polite	**8%** **0.82**	**F4: Fluency** Voice clarity Politeness Know what was expected	**N/A**
F5: Voice Liked Voice Voice clear	**8%** **0.92**	**F5: Transparency** Ease of use Prompt helpfulness Degree of fluster	**N/A**
Total Explained Variance	**57%**		**74%**

Table 8 Comparison of the Factor Structure between OVID experiment and the original BT-CCIR experiment (Love et al 1994). The statements are shown in the order of the factor loadings.

a. **Var** is the proportion of the total item variance explained by the Factor. Only data for the two first Factors were available for BT-CCIR. The total explained variance for OVID factoring is 57%.

b. Cronbach’s α for the sub-scales.

The items loading on each factor were inspected and the five identified factors were labelled as shown in Table 8. The identified factors correspond well with the Factors identified in the original version by CCIR as shown in Table 8.

An almost perfect agreement for the statements belonging to the first two factors and a lesser agreement between the following three is found. However, some differences are to be expected, since the items differ a little between the two questionnaires. Another reason might be that the questionnaire was translated into Danish. Furthermore, the version of the BT-CCIR questionnaire reported in (Love et al 1994) contained 22 items instead of the later 20. At least one item refers to “tone prompts”, which is not the case in the OVID questionnaire. The questionnaire seems to differ slightly from later versions (see e.g. CCIR 2003a, CCIR 2003b) and the exact formulation of the statements used in (Love et al 1994) is unfortunately not available.

The subclasses produced by the factoring process are considered to be reasonable. The individual statements are shown in the order after their loadings upon each factor, with the highest loading statements first.

There are various methods to check for the statistical significance, but since factor analysis does involve an element of heuristics or choice, purely statistical tests are not necessarily the most informative and are mostly done to verify trends and check basic assumptions of e.g. covariance and linearity.

However, factor structures are often evaluated by their predictive (test-retest) power. When, as in the case of the OVID corpus, only one instance of the test is available, this can be done in a pseudo-manner way: The highest-loading statements for each factor are used as independent variables in a linear regression to predict the overall user attitude, expressed as the mean of all the core statements in the questionnaire. This is shown in Figure 13 below.

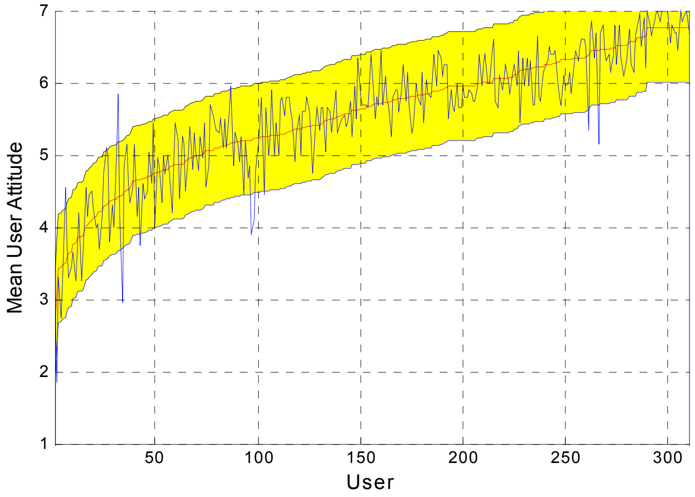


Figure 13. Predicted (the smooth curve) and observed user attitudes with 95% confidence interval band.

The estimated model accounts for 82% (R^2 = 0.823) of the variance of the full set of items. In a similar test on the original BT-CCIR questionnaire, the factor set was shown to account for 86% of the total variability (Love et al 1994).

6.3 Factor Analysis on all the OVID Statements

The preceding analysis served to validate the OVID questionnaire. Because of the comparison with previous results, only the twenty core statements could be used for this. However, it is of course interesting to investigate the factor structure for the full set of statements. The result is shown in Table 9 below.

It was found that a clustering with six factors provided the clearest structure, see Appendix B, page 99. Note that the addition of an extra factor changed the remaining ones as well, which was to be expected. However, several of the Factors (F_1: **Quality of Interface/Performance**, F_2: **Cognitive Load**, F_3: **Control/Confusion**) identified in Table 8 can still be found, but with some differences in the statements loading on them.

Factor	Label / Statements	Var.%
F_1	**Quality of Interface, Performance**	10.7
	Efficiency, Ease of Use, Frustration, Need of Improvement, Reliability	
F_2	**Control/Confusion**	10.4
	Out of Control, Too Complicated, Flustered, Remember Too Much, Knew What To Do	
F_3	**Convenience**	9.7
	Use Again, Good Value, Convenient, Enjoyment, Preference for Human	
F_4	**Personality**	9.3
	Friendliness, Like Voice, Politeness, Voice Clear	
F_5	**Confidence**	8.1
	Security, Confidentiality, Reliability	
F_6	**Cognitive Load**	7.5
	Under Stress, Too Fast, Concentration	

Table 9 Six-Factor structure with all statements included. The total explained variance is 56%.

(F_4: **Friendliness**, F_5: **Voice**) have been collapsed into a new F_4: **Personality** and F_5: **Confidence** is introduced. Note that statement 4: **Confusion** is not included. The reason is that it exhibited a low loading across a number of Factors. Some cross-loading could be observed, especially between F_1 and F_3.

The Factor Structure shown in Table 9 has a much more even distribution of variance among the factors, with almost equal contributions from each. This indicates that the users attribute an equal weight to the identified Factors.

6.4 Discussion

The measurement of user attitudes for speech interfaces is a problematic and yet unresolved issue, as demonstrated in this chapter. Although the applied methods have been in existence for a long period of time, they have not been systematically applied to the domain of speech controlled interaction. Only two tools, the BT-CCIR and SASSI questionnaires have produced any kind of documentation of the validity and reliability that is required for trustworthy scientific results.

A Danish version of the BT-CCIR questionnaire has been applied in the OVID project. The analyses and discussions in the preceding sections of this chapter and Appendix A have shown it to produce results comparable to the originally published work by CCIR on the validity and reliability of the scale.

However, there are some questions about the way the BT-CCIR scale is used, which might prove to be problematic. In the work published by CCIR it is assumed that an overall sum of all items expresses a “general user attitude index” towards the system. This is of course in accordance with the way such scales normally are used, and they are indeed often referred to as “sum scales”. However, if the goal of the test is to investigate a particular issue, say the attitude towards the system voice, the sensitivity of the summed scale will most likely be sub-optimal, compared to a detailed investigation of the items belonging to the F_4 (**Personality**) cluster (see Table 9). in the present case the F_5 items turns out to account for only 9% of the total variability.

Therefore, a number (most often from four to six) of “domain-specific” items with particular relevance to the context is added. This is e.g. the case in (CCIR 2003b), where a number of statements pertaining to the voice and prompt styles are included, and an additional semantic differential scale is added. In the present case of the OVID project five statements were added, but these statements have not been included in the validation process.

While a PCA always captures the maximum variance, factor components are rotated to produce a factor structure with items only loading on one factor to get a simpler model. Furthermore, FA, in contrast to PCA only models the common variance (communality), i.e. the variance the specific variance for each statement is left out, where PCA includes all variance.

This is the reason that the accumulated variance for the first five components shown in Figure 12 is 67%, while the sum of the variances explained by the Factors shown in Table 8 is only 57%. However, the variances are more evenly distributed among the Factors, due to the rotation of the factors.

Finally, a second FA was carried out with all the statements in the OVID questionnaire in order to investigate how the domain specific ones related to the core set. It was found that a six factor structure better explained the data, and that some adjustments to the original structure could be observed. A new factor: “Confidence” emerged, probably because users are more conscious of the aspects of security and confidentiality due to the home banking domain than for the generic case.

Chapter 7 The PARADISE Evaluation Paradigm

The PARADISE - (PAradigm for DIalogue System Evaluation) scheme was first proposed by AT&T Research in 1997 (Walker et al 1997, Kamm and Walker 1997) as a potential methodology for combining observable, quantitative metrics such as task success rates, timing information, etc. with measures of user satisfaction. The objective was to tie user satisfaction to dialogue performance and by this obtain a measure of how much each factor contributed to user satisfaction.

Furthermore, the framework attempts to decouple the dialogue management strategy from the task requirements (e.g. task complexity) by separating task success from dialogue costs, e.g. expressed as number of turns or time spent to complete a task. The rationale is to make it possible to compare (and optimise) different dialogue managers independently of the domain. Also, it became possible to calculate the cost of achieving a certain level of user satisfaction given a certain performance. To cite the authors:

> *"The PARADISE model posits that performance can be correlated with a meaningful external criterion such as usability, and thus that the overall goal of a spoken dialogue agent[1] is to maximize an objective related to usability..... The model further posits that two types of factors are potential relevant contributors to user satisfaction (namely task success and dialogue costs), and that two types of factors are potential relevant contributors to costs." from (Walker et al 1997, p.2).*

The factors relevant to cost are efficiency and dialogue quality measures. The overall structure of the PARADISE method is shown in Figure 14 below.

As their motivation for proposing PARADISE, Walker and colleagues point to a number of problems of the (then) current state-of-the-art of evaluation techniques. Firstly, although various forms of user satisfaction and dialogue systems performance measurements were often collected, there seemed no clear way to tie these together in a clear and systematic way. This was e.g. the case for the OVID experiment. Secondly, they had observed that current methods could not fully explain or would directly generate contradictive results. With reference to (Danieli and Gerbino 1995) they pointed out that although one dialogue agent had a higher performance (i.e. task success) rate compared to another one, the dialogues also tended to be much longer. No clear method existed to determine which one actually was "the best". Furthermore, they wanted a generic method that allowed to compare across domains and agents, and enabled prediction and optimisation of SDS.

[1] Agent = Dialogue Manager

Following the initial proposition of the PARADISE scheme, Walker and colleagues have used the scheme in a number of experiments, with various objectives such as comparison of dialogue management (DM) strategies across domains and tasks (Kamm et al 1999, Walker et al 2000a), algorithms for adaptive or optimised DM through reinforcement learning (Walker 2000, Litman et al 2000), prediction of DM performance (Walker et al 1998). In particular, the following introduction is based on (Walker et al 1998).

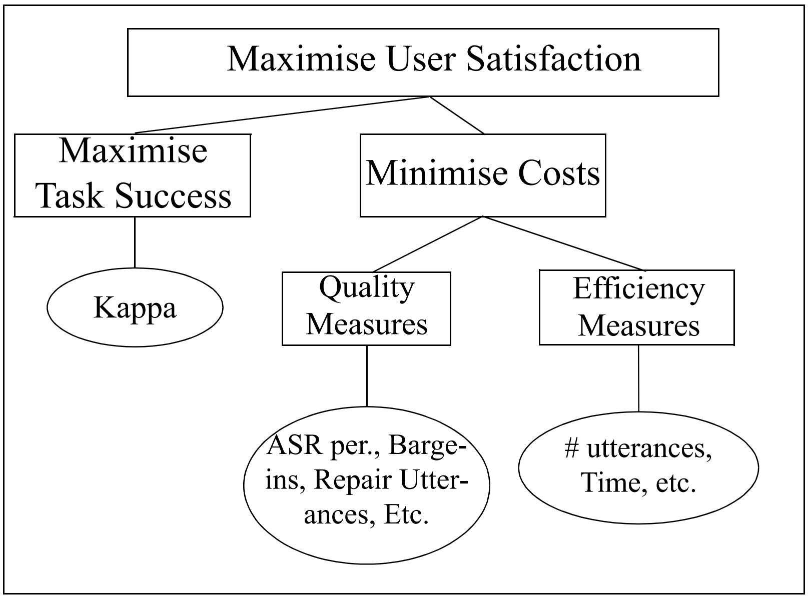


Figure 14. PARADISE structure. From (Walker et al 1997)

The elements constituting the PARADISE model shown in Figure 14 is described in detail below.

The discussion of the PARADISE methodology is illustrated with examples from the OVID Homebank experiment. The experiment was carried out shortly before PARADISE was proposed, and therefore not originally included in the analysis of the OVID results. However, the OVID corpus seems to be is well-suited for a PARADISE evaluation, as will be demonstrated in the following sections.

As described previously, qualitative and quantitative data was recorded using well-defined scenario descriptions. Thus, although the OVID experiment was not carried out with PARADISE in mind, data was collected to that can be used in a PARADISE analysis, although with a number of slight modifications compared to the measures used by Walker and colleagues.

The components of the PARADISE framework shown in Figure 14 are defined and discussed in detail in this section. Most notably, the efficiency and qualitative measures, task success and user satisfaction issues are discussed with reference to other work and the literature in general.

7.1 The Paradise Scheme

7.1.1 Dialogue Measures

As described in section 3.1 many objective dialogue measures have been proposed in various experiments and have been reported in the literature during the last decade. In PARADISE, Walker and colleagues divide these metrics into two categories. One related to “dialogue efficiency” and one related to “dialogue quality”.

The efficiency measures used in PARADISE are elapsed time (duration), system turns and user turns. These measures correspond well with the category “Communication Efficiency and Speed” identified in Chapter 3, page 21. The measures denoted “quality measures” are mean recognition score (of speech concepts), number of system prompt time-outs, speech recogniser rejections, user help requests and user barge-ins[1]. These belong to the categories “Speech input quality” and “Dialogue control and symmetry”, except for the help request, which is a meta-command.

While the Efficiency measures are well-defined and matches the taxonomy adopted in Chapter 3, the measures of Quality seem somewhat arbitrary. Most of the chosen measures relate to speech recognition performance (or quality), which can hardly be mapped directly to “dialogue quality”. None of the measures maps to the Grice maxims.

However, as stated in (Walker et al 2000a), any measure for dialogue efficiency, quality or task success can easily be incorporated into the PARADISE scheme. The corresponding parameters of the performance function are determined through a regression process and will not be influenced by which category the measure is associated with. So the distinction is more a conceptual than a computational one.

Task success is measured as the perceived task completion by the users together with the observed task completion. The distinction between perceived task completion and observed task completion is an important one, as user satisfaction can be expected to depend on the perceived rather than the

[1] plus the corresponding relative values, included to achieve a higher degree of generalisation across domains and agents

observed task completion rate, especially when in a test situation with constructed scenarios (Walker et al 1998).

7.1.2 Task Success

The Kappa (κ) coefficient represents the task success rate, as shown in Figure 14. It attempts to compensate for the task complexity by calculating the ratio between the actual task success with the "accidental" task success one would obtain by purely random answers. The compensation is necessary when PARADISE is used to compare across tasks with different complexities. A more traditional measure of task success (based on a ratio between desired and achieved goals) would favour simpler tasks for more complex, since there is a greater chance for accidentally obtaining the correct information.

κ is a well-known statistic (see e.g. Siegel & Castellan 1988) and within speech and language technology, κ has been used to assess the agreement between evaluators of e.g. segmentation boundaries (Carletta 1996, Walker et al 1998).

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

(EQ 1) The Kappa Statistic

P(A) is the proportion of correct (sub)goals achieved in a dialogue and P(E) is the number of times chance agreement is expected to occur. Thus, κ will equal one for perfect agreement (i.e. all goals are correct and P(A) becomes equal to one) and zero for only chance agreement (when P(A) becomes equal to P(E))[1]. P(E) is usually set to equal 1/(number of possible answers or goals).

This definition of task success makes it evident why PARADISE can only be applied for very well-defined tasks and goals. If this is not the case, it will become impossible to reliably compute values for P(A) and P(E) and hence obtain κ. As mentioned before, it is a basic requirement that the experiment is based on scenarios with well-defined goals.

7.1.3 User Satisfaction

As stated in the citation introducing this chapter, the goal of PARADISE is to establish a correspondence between the performance measures and "*a meaningful external criterion*" which Walker and colleagues define as the

[1] Note that κ can theoretically assume negative values. If this occurs κ is set to zero

usability of the SDS (Walker et al 1997). To obtain a measure of the usability, they employ an approach quite similar to the one described in Chapter 5 and 6 and used for the OVID experiment. (Walker et al 1998) used a questionnaire comprised by the nine statements addressing the aspects of system behavior shown below:

1. TTS Performance
2. ASR Performance
3. Task Ease
4. Interaction Pace
5. User Expertise
6. System Response
7. Expected Behaviour
8. Comparable Interface
9. Future Use

In a later version, used for the evaluation of the DARPA Communicator project (Walker et al 2000a), the number of statements have been reduced to five, (statements 2, 4, 6 and 8 were removed and a question about the users perceived task success was added).

Similar to the OVID experiment a cumulative score is calculated for each user. However, as mentioned in Chapter 5, Walker and colleagues do not provide any evidence of the validity (or even reliability) of the questionnaire. At best, it possesses content validity, although no strong arguments for this are provided either.

7.2 The PARADISE Performance Function

PARADISE seeks to establish a correspondence between the objective, observed performance of the dialogue agent with the subjective experiences of users interacting with the system as described in the previous section. This correspondence is termed the Performance Function and is estimated using Multiple Linear Regression (MLR). As the name implies, MLR models the dependent variable (the usability) as a linear combination of multiple independent variables (κ and the cost and performance measures), determined through a regression. For a general introduction of MLR, see e.g. (Anderson et al 2002 and Tabachnick and Fidell 2001). Additional results and definitions related to the MLR analysis can be found in Appendix B, page 100 ff.

7.2.1 The Attribute-Value Matrix

The Attribute-Value Matrix (AVM) is central to PARADISE and is used to assess task success rates. The idea behind the AVM is to express (sub)tasks as a combination of attribute-value pairs. The values are registered at the end of the dialogue and consequently the AVM does not record **how** a value was obtained or whether it e.g. has changed one or several times during the dialogue. For example, the result of an OVID dialogue could be the AVM shown in Figure 15. During the dialogue, the user has supplied his Id-number and obtained a statement for his cash credit account. Note that the AVM does not contain any information about the way this information was obtained, how long it took, etc.

Attribute	**Value**
Id-number	**973625**
Account	**Cash Credit**
Action	**Statement**

Figure 15. Example of AVM for OVID

This implies that the AVM can be used to e.g. compare different dialogue management strategies for the same domain. Another implication is that the AVM is defined by the goals specified in the scenario the user is required to perform, not the by the dialogue model. Therefore, multiple AVMs can be defined for a given SDS, depending on the task.

The AVM is used when computing the Kappa (κ) statistic for task success rate (see below).

7.2.2 The Performance Function

As mentioned above, the performance function is estimated by a multiple linear regression (MLR), with the various dialogue cost and performance measures as the independent variables and a usability score as the dependent.

MLR can be formulated using different mathematical notations, depending on the tradition and purpose, usually in algebraic (matrix) form. However, for reasons of convenience and to enable direct comparison, a notation similar to the one used in (Walker et al 1998) is used here,

$$Perf = \alpha \times N(\kappa) - \sum_{i=1}^{n} w_i \times N(c_i)$$

(EQ 2) Performance Function (from Walker et al 1998)

Perf is the dependent variable, representing the usability of the system. The independent variables are κ and the cost and performance measures denoted c_i. α and w_i are the weights on κ and the dialogue quality and cost measures c_i. which are to be estimated through the regression process. *N* is a Z-score normalisation function. It is included to make the identified model parameters α, w_i directly comparable in size, but does not in any other way influence the relationships.

MLR sets some constraints on the relationships between the parameters. Most obviously, there must exist a linear - or at least an approximated linear - relationship between the independent and the dependent variables. Furthermore, the parameters in the model must be normally distributed and independent (un-correlated) with each other. This is seldom the case, however. One can always assume some correlation between parameters, and a deviation from normality is also common. MLR is rather robust against minor deviances, but this must obviously be checked when applying the method.

7.3 Applying Multiple Linear Regression to the OVID corpus

There are several problems that must be overcome before PARADISE can be applied to the OVID experiment. Firstly, as mentioned before, the experiment was not designed with PARADISE in mind and therefore some adjustments must be made. Especially two factors must be handled: The questionnaire was only administered once - after both scenarios had been carried out, see section Section 4.4 on page 35. Therefore, the measured user attitudes are a result of the users' experiences from both scenarios which consequently must be treated as a single instance. This is further elaborated below. The second factor has to do with the complexity of the task. The OVID tasks are rather straightforward and the task completion rates are high. 93% of the users completed at least one scenario and 74% completed both scenarios (see Table 6 on page 34 and (Larsen 2003a)). Nearly all users (96%) reported that they had completed both scenarios successfully. This is problematic, since the users' perceived task success rate obviously must have some (unknown and unmeasurable) impact on their attitudes towards the system. For example, (Walker et al 1998) argues that the perceived task completion rate must be used instead of the real one, and in later experiments, the users are directly asked whether they were able to complete the tasks (Walker et al 2000a), similar to the OVID experiment.

In order to minimize this factor it was decided to use only a subset of the corpus, where the users had positively encountered problems that forced them to initiate an extra dialogue in order to complete the two scenarios. Therefore, they must have been fully aware that a dialogue had failed to complete the scenario. A scan through the corpus showed that this was the case for 35 users (who consequently carried out 105 dialogues). This was considered sufficient for a statistically valid analysis. Apart from this, no other criterion for selection has been applied, and the nature of the errors that led to the failed dialogue was for example not taken into account. Table 10 below summarises the sub-set. The demographic distribution are comparable to the full corpus.

Scenario	Perceived	%	Scenario	Actual	%
Both	35	100	Both	21	60
Only A	0	0	Only A	5	14
Only B	0	0	Only B	6	17
None	0	0	None	3	9
Total	35	100	Total	35	100

Table 10 Subset of the OVID corpus used for the PARADISE evaluation

Table 10 indicates that although all the users reported that they had completed both scenarios, only 60% actually did so. This is similar to the findings for the full corpus, c.f. Table 6 on page 34.

7.3.1 Constructing an AVM for the OVID task

It is necessary to construct a single AVM to cover both the scenarios that the user is asked to carry out with the system, since the usability questionnaire is only filled out once by each user. Consider for example a case, where only the second scenario was used in the PARADISE evaluation. The system might have performed perfectly in this dialogue, but the user experienced severe problems in his first dialogue. It is highly probable that this first bad experience still influences his attitude towards the system, and that s/he will answer the questionnaire accordingly (in fact s/he is expected to do so). For this reason, all dialogues must be treated simultaniously.

Because the OVID experiment was scenario-based, i.e. the users were given specific tasks to carry out it is quite straightforward formulate an AVM and thereby to identify the degree to which the goals were met and whether the user managed to extract the required information from the system. One AVM for each scenario were defined and combined into the one shown in Table 11 below.

Scenario A		Scenario B	
Balance$_{\text{Acc1}}$	<value$_{\text{Acc1}}$>	**Balance$_{\text{Acc2}}$**	<value$_{\text{Acc2}}$>
Balance$_{\text{Acc2}}$	<value$_{\text{Acc2}}$>	**MiniStat$_{\text{Acc2}}$**	<value$_{\text{Acc2}}$>
Balance$_{\text{Acc3}}$	<value$_{\text{Acc3}}$>		

Table 11 The combined AVM for the two scenarios in the OVID experiment

Note that the Id- and PIN codes are not included in the AVM. These have been omitted because the Id- and PIN codes are the only means to identify the user, and consequently dialogues where these are not present can not be ascribed to a particular user.

κ can now be calculated using equation 1 on page 64, by computing and summing P(A) and P(E) for each dialogue, selecting the appropriate AVM for each dialogue.

7.3.2 Choosing appropriate cost and quality parameters for the OVID experiment.

As described in Chapter 3 p. 17 a wide variety of performance measures have been applied for evaluation of spoken dialogue systems. However, there are a number of constraints which reduces the available choices.

- Obviously, the measures must be available (or at least easily derived from) the corpus. This is true for:
  - Speech concept recognition rates (ASR)
  - Total time and time spent in individual subtasks
  - Turns in total and individual subtasks
  - Proportion of user initiatives relative to the total number of turns
- The parameters must be “virtually uncorrelated”, meaning that only a limited interdependency can be accepted. For example, the elapsed time and the number of turns can be expected to correlate heavily and can most likely not both be included.
- There must exist a linear (or approximated linear) relationship between the measures and the recorded usability.

However, this can be checked by computing the covariance between the above mentioned parameters. The resulting covariance matrix is shown in Table 12 below. From the matrix it can be observed that “Total time” and “Total Turns” are indeed heavily correlated (0.9). It can also be seen that none of the other parameters are highly correlated. “Task Success” is a more conventional measure for task completion, roughly equivalent to the proportion of achieved (sub)goals, denoted P(A). As expected, it correlates with κ. For comparison, “User Satisfaction”, calculated as the averaged score for the 20 core statements for each user, is also included in the table and all parameters show some correlation with it. Turns and Time correlate negatively, whereas the task success and SR measures correlate positively.

Covariance Matrix	Kappa	Task Success	**ASR**	Total Turns	Total Time	**User Satisfaction**
κ (Kappa)	1.0					
Task Success	0.6	1.0				
REC	0.2	0.4	1.0			
Total Turns	0.1	0.0	-0.4	1.0		
Total Time	0.0	-0.1	-0.6	0.9	1.0	
User Satisfaction	**0.5**	**0.3**	**0.6**	**-0.3**	**-0.4**	**1.0**

Table 12 . Covariance Matrix for selected dialogue measures. REC is the speech concept recognition score. Due to the Z-normalisation the values in the diagonal (the variances) all equal 1 and the (absolute) off-diagonal values lie between 0 and 1.

Normality. The central limit theorem asserts that normality of sample means can be assumed, provided the sample is “large”. Usually this is accepted to be about 30 samples, i.e. comparable with the present case (see e.g. Anderson et al 2002 or Tabachnick and Fidell 2001). Normality of the sample often verified using graphical methods e.g a normal plot. Examples of these are given in Figure 16 below and Appendix B, page 101

Choosing the dependent variable. An individual analysis of the factor clusters identified in Table 8 p. 55 showed that prediction of the F_1 (Quality of interface/performance) cluster statements produced a slightly better fit than any other Factors or combination of them. Figure 16 shows a normal plot for the distribution of the mean user attitudes for F_1.

Although some deviation from the straight line can be observed it seems reasonable.

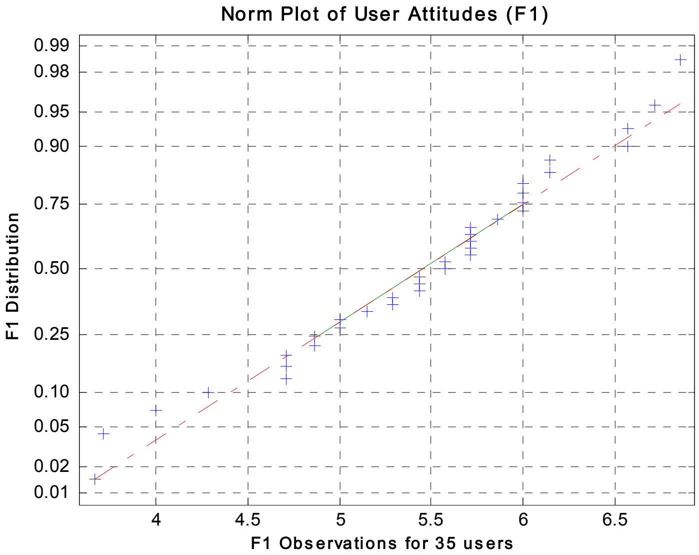


Figure 16. Normal probability plot for the distribution of the F_1 statements. The Y-axis is scaled so that a normal distribution will produce a straight line.

7.3.3 Estimation of MLR parameters for the PARADISE model.

A series of regression analyses was carried out to identify the best possible model. The requirements are that the resulting model must represent the best fit and of course be statistically significant. Details of the analysis and additional results are described in Appendix B.3 p. 100 f.f.

During the analysis various combinations of the parameters shown in Table 12 were tried out. The experiments showed that only κ and REC (speech concept recognition score) were significant predictors of user satisfaction. Furthermore, in addition to the overall summed scale, the Factor clusters identified in Chapter 6 on page 51 ff. were investigated in various combinations. In Appendix B it is shown that the MLR model could better estimate F_1 (Quality of interface/performance) than any other factor or combination of factors. The identified model parameters are shown in Table 13 below. Four examples of similar results for experiments carried out at AT&T are included for comparison.

SDS (Domain)	Performance Function	Var
OVID (Home banking)	Perf = 0.41*κ + 0.47*Rec	51%
TOOT$_1$(Train travel)[a]	Perf = 0.45*Comp + 0.35*Rec -0.42*B.I	47%
TOOT$_2$(Train travel)	Perf = 0.33*Comp + 0.45*Rec -0.14*Time	55%
Annie (Voice Dialling)	Perf = 0.25*Comp + 0.33*Rec -0.33*Helps	41%
ELVIS (Email access)	Perf = 0.21*Comp + 0.47*Rec -0.15*Time	38%

Table 13 Comparison of results from OVID and three SDS from AT&T (Kamm et al 1999). Perf is the average usability score, Rec the speech recognition score, Comp is the perceived task completion rate, B.I. is Barge-Ins, Time is the duration of the dialogue and Helps is the number of help requests. Var is the proportion of variance explained by the model

[a] The two Toot systems address the same domain, but employ different dialogue management strategies

Several interesting observations can be made from the table. As mentioned above, only κ and Rec turned out to be statistically significant predictors of user satisfaction (Perf) for the OVID system. Although "number of turns" and "elapsed time" (Time) also correlate with Perf (see Table 12) they are not significant predictors, and a model including one of these measures does not produce a better fit of the independent variable Perf.

Comparing the results from OVID with similar PARADISE analyses of the three SDS built by AT&T researchers and reported in (Kamm et al 1999), a notable correspondence between the findings is found.

Except for one case, speech recognition is found to be the most important contributor, which is to be expected. The influence of speech recognition performance for OVID is identical to those found for Toot$_2$ and Elvis and close to the ones found for Toot$_1$ and Annie. Although the AT&T experiments applied the users' perceived task success (Comp) instead of κ, close to identical results are found for Toot$_1$ and Toot$_2$. The AT&T experiments also found significant predictors, although of lesser importance, for Help, Barge-Ins and elapsed time. Except for elapsed time, these measures were not available for the OVID corpus.

A comparison of the 'goodness of fit' of the estimate (expressed as the percentage of the variance explained by the model) also shows very similar results, (between 38% and 55%) with 51% explained variance for the OVID

SDS. However, as Table 3 on page 102 in Appendix B shows, the confidence intervals are quite large, as can also be seen in Figure 17 below.

Even the best fitting model for OVID only explains 51% of the variance. This is partly due to noise in the observed values, which must always be expected. Other sources of error are non-linearity, insufficient sensitivity of the questionnaire to capture the users' attitudes and potentially effects of "social desirability", as mentioned in Chapter 5, page 42. However, the result must necessarily raise the question to what extent the model actually is useful.

Regardless, the result is comparable to those found by the AT&T research team in (Kamm et al 1999).

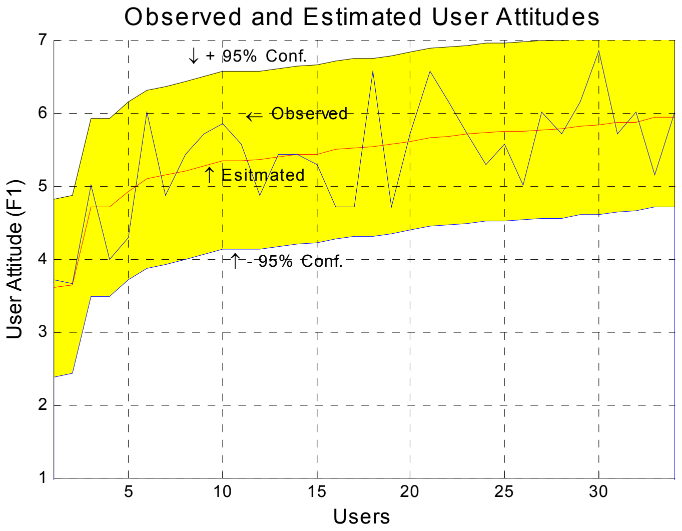


Figure 17. The red line represents the estimated user attitudes, and the blue line the observed values (F_1). The 95% confidence band is shown in yellow.

Figure 17 shows a plot of the values predicted by the model together with the recorded values and 95% confidence band. Although the observed values generally fall well within the confidence band (one outlier was identified and removed before the regression was carried out) of the predicted ones, it is obvious that a large proportion of the variability has not been captured by the model.

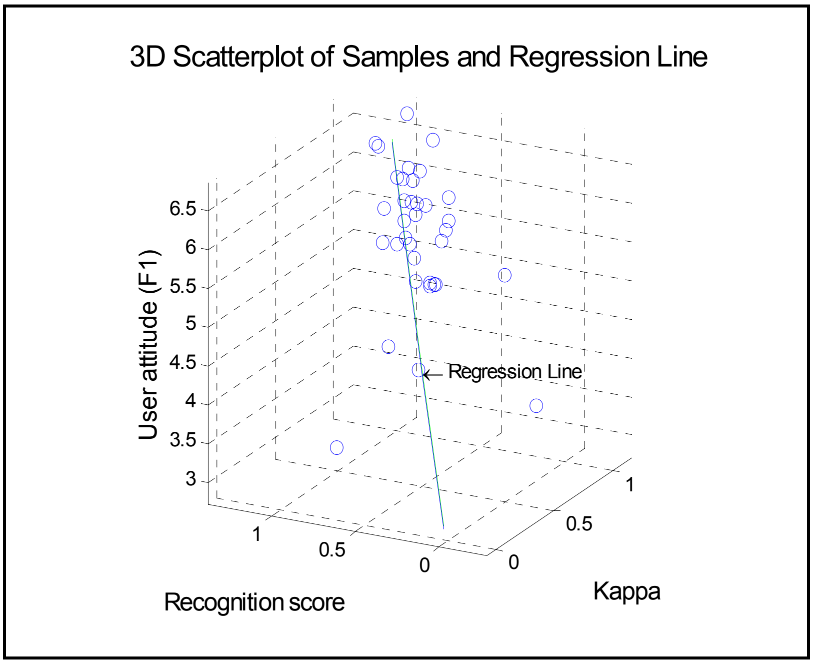


Figure 18. 3D-Scatter plot of the observed values and resulting regression line.

Figure 18 shows a 3D-scatter plot and the estimated regression line. This should be viewed with some reservations, but is quite illustrative nevertheless.

7.4 Discussion

As shown above, it was possible to apply the Paradise scheme to a subset of the OVID corpus and obtain results comparable to those published by Walker and colleagues. The important question is of course whether it revealed any new information about the corpus. It is hardly surprising that a relationship between ASR performance and user satisfaction can be observed. This is a well-known fact and has published numerous times (see e.g. (Jack et al 1993)) and also for the OVID corpus, although only a weak one[1]. Obviously there are other important factors influencing user satisfaction.

Interestingly, κ proved to be a better predictor than a more traditional measure for task completion based on a simple ratio between desired and obtained

[1] See Figure 3 on page 194, in (Larsen 1999)

goals. The main function of κ is to normalise for task complexity. This indicates that the two scenarios turned out to differ in complexity and that κ captured this fact.

The estimated model only explains approximately half of the observed variability and the coefficients are estimated with some uncertainty. One likely reason for this is that the users unfortunately did not have a very precise perception of their actual task completion. This will of course be reflected in an unclear or muddled relationship between their attitudes towards the system (as expressed in the questionnaire) and the hard facts (obtained from the logfiles). As shown in Table 6 on page 34 close to all users reported that they had carried out the assigned tasks, while in fact only three quarters of them did so.

Furthermore, the experimental design required the users to fill out a single questionnaire after completing both scenarios, obviously also plays a role. It has most likely caused the correspondence between the usability questionnaire and performance measures to become less direct, since the users were forced to "average" their attitudes over two, or in some cases even three dialogues. Put in other words, the sensitivity of the user attitude measurements has decreased.

An interesting implication of the results shown in Table 13 on page 72 concerns the questionnaires used to obtain the user attitudes. The AT&T user attitude questionnaire only comprises 9 statements (on a five-point Likert scale - included in Appendix B, page 103) and no evidence of validity has been published[1]. In contrast the BT-CCIR questionnaire consists of 20 core statements (on a seven-point scale), quite different from those of AT&T. Regardless, PARADISE produces quite similar results both regarding the combination of measures and the fit of the model. Two explanations come to mind: Either PARADISE is not particularly sensitive to the user attitude elicitation questionnaire, or both questionnaires essentially capture identical measures from the users. If the latter is correct, the PARADISE analysis can be regarded as a supplementary proof of validation of the questionnaires.

7.4.1 Further Observations about the PARADISE Scheme

Although the application of the PARADISE scheme on the OVID corpus worked out quite well, there are two matters that are problematic in a wider perspective: One concerns the requirement for specific scenarios with clearly defined goals and has been briefly touched above. This poses a serious threat to the generality and scalability of PARADISE to e.g. multi modal systems. The other concerns the calculation of some of the parameters. A number of assumptions are made, e.g. about linear relationships between performance

[1] To the authors knowledge

and subjective measures. There are really no hard evidence that this is the case. As mentioned previously, many studies have shown a relationship between speech recognition performance and user attitudes. Indeed, such curves are seldom straight lines, and sometime even have a “threshold”, where the slope changes abruptly.

Furthermore, the parameters, most notably the AVM and κ measure used to represent task success cause problems. Unless the test scenarios are very structured and well-defined the definition of these become ambiguous, as indeed a number of studies have revealed, e.g. (Bouwman and Hulstijn 1998, Hjalmarsson 2002), and also in an adaption of PARADISE for evaluation of multi modal systems in the German SmartKom project (Beringer et al 2002). Some of the problems can be solved by a more specialised and dynamic generation of AVMs, but this will in turn reduce the value of PARADISE as a powerful comparative tool across tasks and domains. The definition of κ relies on the fact that P(A) and P(E) can be unambiguously defined and computed. However, P(E), which is often approximated by $1/n$, (with n being the number of possible values a given attribute can assume in the particular context) might differ substantially across subtasks, and might become very small. Again, this can be circumvented by splitting into subtasks and calculate individual values for each, at the cost of greater complexity and less generality.

The Z-normalisation (performed by subtracting the mean and dividing by the standard deviation from all observations) allows direct comparison of the contribution (weights) of each parameter. This can be very illustrative, as shown in Table 13 on page 72, where a number of different PARADISE performance functions are directly comparable. However, if the observations that are normalised varies substantially, e.g. from one subtask to another, normalisation can be problematic. This problem can be circumvented in a similar manner as suggested above.

However, when applications of similar complexity and outlook are to be evaluated, as in the case of the DARPA Communicator project, or when e.g. successive versions of a system is tested, PARADISE is a powerful tool. Furthermore, the existence of a widely used and (although with limitations, as discussed above) standardised evaluation paradigm can only be a positive element. At least it will stimulate research and development of other, perhaps better paradigms. The efforts by the SmartKom project to extend and modify PARADISE to multi modal dialogues is a good example of this.

Chapter 8 Conclusions and Future Directions

This Chapter contain three main sections. First a summary of the results and conclusions of the topics presented in the previous chapters is given and commented. This is followed by a discussion of related research and the implications of the findings in this study. Finally, some perspectives for future research and the problems facing speech as an interface modality are outlined.

8.1 Summary of Results and Conclusions

This work is centred on the methods and problems associated with defining and measuring the usability of Spoken Dialogue Systems. The starting point is the fact that speech based interfaces has several times fallen short of the expectations and predictions. Several studies in the literature of SDS indicate that this can be ascribed to a lack of attention from the SDS research community towards the usability of such systems. It is shown that users place great value on their time and are unwilling to spend time on e.g. the configuration and training often required by speech based interfaces (Cameron 2000).

8.1.1 The Usability of Spoken Dialogue Systems

The are many different views on how usability is defined, and what attributes of usability are important. In this work the viewpoint put forward by (Nielsen 1993) and other researchers is adopted: That usability and utility together defines the usefulness of a system, and that flexibility and learnability are particularly important attributes of SDS usability and must receive special attention.

Due to a number of circumstances that set SDS apart from more traditional interfaces, these attributes require special attention in the design and evaluation phases. These are mainly related to the non-persistence of speech and issues of control of the interaction. Furthermore, speech input processing is more complex and error-prone than input processing in traditional user interfaces. Together, these factors suggest that traditional tools and methods for measuring usability are not necessarily valid for SDS. Another implication is that attributes such as error-handling, transparence, control and learnability become the main points for attention (Dybkjær and Bernsen 2000).

The usability of a given system is determined through simultaneous assessment of objective (performance) and subjective (user satisfaction) measures. These can only be established through the application of experiments with representatives of the end users. The issue of controlled (laboratory) versus uncontrolled (field) tests was discussed together with the problems associated with scenario-based testing

This leads to the, the focus of this work on the investigation of methods for establishing the usability of SDS through the application of end-user field trials. In the experiments, performance and user satisfaction measures are recorded and analysed in an integrated manner.

8.1.2 Performance Measures

A multitude of performance measures have been proposed for SDS in the literature. In order to classify the measures they are categorised into aspects of dialogue cooperativity and -symmetry, speech input quality and communication- and task efficiency. Of these, dialogue cooperativity is not addressed explicitly by most researchers. However, as shown by (Grice 1975) the maxims of cooperativity governs human behaviour and expectations of their "dialogue partner", and can therefore be assumed to also influence the usability of SDS. A given measure can belong to several categories. For example, "number of attempted user barge-ins" can be used as a measure of cooperativity (relevance) and symmetry (dialogue initiative).

The methods for obtaining values for the measures are investigated and it is found that many measures can be obtained or derived automatically from system logfiles. However, some measures (e.g. response appropriateness) rely on human transcription and interpretation, which make them costly to obtain and potentially biased.

8.1.3 Results obtained in the OVID experiments

The requirements for the OVID field test are identified and found to be heavily related to the manner of addressing the system (unconstrained natural speech) and control (the user must feel in control) (Larsen 1996). Test users were selected among Lån & Spar Banks' customers aiming to get an even demographic distribution of genders, five age groups and four geographical regions. However, no differences in performance could be detected for the groups.

In the OVID experiments, performance measures related to:

- Task Efficiency (proportion of successful goals and sub-goals)
- Communication Efficiency (elapsed time and number of turns in tasks and subtasks)
- Control and Symmetry (proportion of user initiated turns)
- Quality of Speech Input (speech concept recognition rates)

- were collected and analysed for 310 users performing 700 dialogues according to two use case scenarios.

The measures relating to user control and task- and communication efficiency were analysed in detail to determine the degree of the usability attribute "learnability". Statistically significant figures are found for the reduction of the time spent in the user identification subtask, when comparing the first dialogue to the second dialogue. Likewise, a statistically significant reduction (25%) of the number of failed dialogues was observed from the first to the second call.

Similarly, a significant increase in the user's ability (or willingness) to take the initiative in the dialogue was found for scenario B. These findings can all be interpreted as evidence of system learnability and that users are capable of taking control of the interaction.

Even though a decrease in dialogue duration has been shown, the OVID dialogues still take longer than those of the corresponding IVR service. However, as users become more experienced, further reductions in duration may be expected. In addition, experienced IVR users employ (keypad) barge-in extensively. Barge-in was not supported in the OVID experiment, and this clearly also influences the result.

8.1.4 Subjective Measures

Recording of user attitudes requires a quite different approach, since these are not directly observable, but must be obtained from the users via e.g. a questionnaire. The most important problem is to ensure the validity and reliability of the recorded measures. It is demonstrated that especially the issue of establishing the construct validity of the questionnaire used to elicit the user attitudes is problematic and has to a large degree been neglected by the speech technology community, except for a few cases.

Establishing construct validity is a difficult and time-consuming process, as is shown by comparing the development and validation process for four different cases, the SUMI, QUISS, BT_CCIR and SASSI questionnaires.

The questionnaire used in the Danish OVID experiments is a translated version of the BT_CCIR questionnaire. It is analysed with regard to internal consistency, which was found to be high (Cronbachs $\alpha = 0.92$) and comparable to similar results obtained for the original English language version. On average the user's attitudes towards the OVID home bank service was 5.6 on a 7 point Likert scale with 4 as the neural point.

Factor analysis is applied to uncover the underlying relationships (Factors) between the users responses to the statements. Through a series of experiments with varying numbers of factors, a structure consisting of a set of five underlying factors was identified. The factors are rotated to reduce cross-loading of the statements and the resulting factor set explains 57% of the total variance of the statements in the questionnaire, which is comparable to similar results obtained by CCIR.

The resulting factor structure corresponds well with the findings reported in (Love et al 1994). The statements loading on $F_1$and F_2 are nearly identical, although some differences are found for the remaining factors. However, this is to be expected, partly due to the translation of the statements and partly due to the nature of factor analysis, where the interpretation of the identified factor structure relies on the experimenter, and thus may vary on the purpose and intuition. The predictive power of the factor structure is demonstrated in a simulated test-retest experiment.

This analysis served as a validation of the OVID questionnaire. A new FA is performed, with all statements included. The purpose is to investigate how the additional statements fit into the core set. The factors are labelled according to the common topics addressed by the statements belonging to each cluster. These are (in order of explained variance):

F_1 Quality of Interface, Performance
F_2 Control/Confusion
F_3 Convenience
F_4 Personality
F_5 Confidence
F_6 Cognitive Load

As expected, some of the factor clusters are quite similar to the previous ones, while “Voice” and “Friendliness” have been collapsed into “Personality” and a new cluster “Confidence” has been added, probably caused by the home banking domain.

8.1.5 Combination of the Objective and Subjective Measures using MLR

The PARADISE paradigm, proposed by researchers from AT&T (Walker et al 1998), attempts to establish a correspondence between the objective and subjective measures obtained when evaluating SDS. Obviously, such a correspondence is highly desirable in itself, since it (in theory) enables predictions of user attitudes from a number of observable performance measures and thus eliminating - or at least greatly reducing - the need for costly and time consuming user tests. However, the PARADISE scheme also seeks to normalise for task complexity and separate what is termed cost and quality measures. The intention of this is to facilitate comparisons of different dialogue managers across different application domains.

However, there are severe limitations to the approach. The scheme requires that an Attribute Value Matrix (AVM) must be formulated, specifying the users intended goals and the extent to which these have been achieved. This information is not readily available, unless the dialogue corpus used to develop the model has been recorded using strictly controlled scenarios and the users' goals thus can be identified and a task completion rate can be computed for each dialogue.

Furthermore, the performance model is estimated using Multiple Linear Regression (MLR), that requires the model parameters to be normally distributed and express a linear relationship.

A performance function is derived for a subset of the OVID corpus and it is shown to account for 51% of the total variance of the observed user satisfaction. Speech recognition accuracy and task completion is shown to be statistically significant predictors of user satisfaction at a 95% confidence level.

While these results are comparable to those obtained in (Kamm et al 1999), it is questionable how much they contribute to the interpretation and understanding of the results obtained in the OVID experiments.

As mentioned above, the performance function is intended to promote comparison across DM strategies and domains and as the OVID experiment is a stand-alone investigation this is unfortunately not directly possible. The resulting performance equation is similar to results obtained by Walker and colleagues, but only in general terms. For example, only the end result of the analyses are publicly available. Additionally, different performance measures, and a different user attitude questionnaire have been used in the experiments, preventing a more detailed comparison.

8.2 Comparison with similar Applications

CTT-bank. In 1999-2000 the Centre for Speech Technology (CTT) at KTH in Stockholm carried out a speech controlled home banking experiment called “CTT-bank” (Melin et al 2001), quite similar in outlook to the OVID experiment. However, the important aim of the CTT-bank was to investigate speaker verification as a means of user identification. A test with 21 users performing 112 dialogues (40 speaker verification enrolment and 72 regular banking dialogues) are reported in (Melin et al 2001). Apart from the speaker verification and the use of text-to-speech synthesis, the scope of the dialogue and applied technologies are quite similar to those of OVID.

An important difference from the OVID experiment is that the users were required to call the service from three to seven times during a two-week period without fixed scenarios. This would offer a good opportunity to study system learnability, but unfortunately this is not addressed. The relationship between the users' perceived speech recognition accuracy and the observed one is investigated. some correspondence is observed, but due to the low number of users, no firm conclusions are made. An interesting result is the users attitude to the login procedure, which consists of speaking their name and a PIN code. A very high preference for spoken login (4.0 on a 5-point scale) is reported and in contrast to OVID, no users are reported to have expressed concern about the confidence of the service. User attitudes are not reported in general.

Overall (Melin et al 2001) make similar conclusions to those reached in OVID and also pointed out by (Cameron 2000): Convenience, in the form of speed and the ability to speak naturally are decisive factors for user acceptance.

Automatic Teller Machines. Hone et al. (Hone et al 1998) investigated the potential of using speech as a modality for ATMs. While the functionality is similar to that of an automatic telephone service, the environment is very different, since ATMs are placed in public locations, and you can actually withdraw cash. The experiment consisted of an initial user attitude survey with 862 persons, followed by a user trial with 23 persons who used a voice-enabled simulated ATM. Both studies revealed, as expected, that the users feared to be overheard. 80% of the users in the survey were worried that they might be overheard not only by other customers, but also potential muggers. This was verified by the trial, where only 5 out of 340 utterances were not overheard by people waiting in line by the ATM (up to a 2 m distance). Various attempts, such as placing a hood over the ATM or using it with a handset were tried, but made no difference.

When asked what they liked about the speech interface, users cited convenience and speed as factors. Focus group discussions were carried out with blind or physically impaired users. However, despite having difficulties with the existing ATMs the attitude was similar to the remaining users.

8.3 Future Perspectives

8.3.1 Multi modality

As mentioned in the introduction, the issue of multi modal user interaction is not addressed in the OVID experiments. However, there is no doubt that multi modal user interfaces, especially during the last five years have received massive attention from research and industry, and that uni modal interfaces (as e.g. speech-only SDS) in the future will be regarded as a special case of the more general multi modal user interaction.

It is therefore relevant to consider whether the methods applied in this work generalises to multi modal user interaction. There is no doubt that the general definitions of usability still holds, but the emphasis on the various attributes might be different, as it turned out to be for SDS compared to traditional desktop interfaces. The aspects of dialogue cooperativity (based on Grice's maxims) are likely to be even more relevant in the general case of multi modal user interaction.

Likewise, the methods for user attitude elicitation will need validation in much the same manner as described in the present work, although new scales must be designed and verified. For example, newer versions of the SUMI questionnaire have been developed for web based interfaces (WAMMI)[1] and is currently under development for multimedia (MUMMS)[2] interfaces at the University of Cork, using the methods described in Chapter 5.

Similarly, efforts have been made by (Beringer et al 2002) to extend PARADISE to handle multi modal interaction. They encountered the problems of defining the AVM and κ described earlier, and propose a scheme (PROMISE[3]) using 'information bits' to measure task completion. Furthermore, they relate specific user attitude statements directly to the objective measures and observe the degree of correlation between these. A graphical tool has been constructed to aid this process.

Technology providers, such as the partners of the SALT consortium[4] are pushing towards standardised formalisms for integration of speech with other modalities. Scansoft recently published information about a new platform denoted "X|mode" (Scansoft 2003).

By downloading a plugin on the client (the user's handset or a PDA) multi modal communication with the system server becomes possible as shown in Figure 19.

X|mode will support various protocols and platforms such as GPRS (General Packet Radio Service) and DSR (Distributed Speech Recognition), PDA's running PocketPC and mobile phones such as the Nokia 60 series.

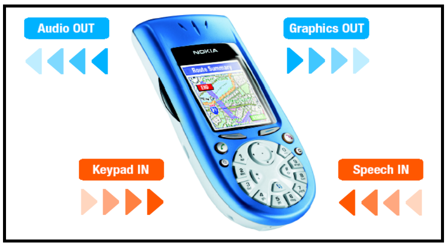


Figure 19. Multi Modal Communication on a mobile phone (Scansoft 2003)

[1] WAMMI: Website Analysis and MeasureMent Inventory. See: http://www.ucc.ie/hfrg/questionnaires/wammi/index.html

[2] MUMMS: Measuring the Usability of Multi-Media Systems. See: http://www.ucc.ie/hfrg/questionnaires/mumms/info.html

[3] Procedure for Multimodal Interactive System Evaluation

[4] http://www.saltforum.org/

To facilitate research in multi modal user interaction, the SMC group has established a setup, based on the Philips SpeechMania platform for SDS. The platform is extended with a Web interface, showing a simulated GUI of mobile phone, see Figure 20, as well as an interface to a MySQL database for storage of e.g. domain data and in particular to store dialogue logging events.

By using a simulated GUI, the interaction might become less realistic, but it provides a great flexibility, both on the development, as no specialised software has to be developed for different platforms, but also for user testing, since all that is needed is a standard (or mobile) phone and a PC with internet access.

The setup also includes a highly configurable Web based questionnaire, enabling user attitude data to be stored directly alongside the logged data for easy retrieval and further processing.

This setup provides an efficient platform for experiments, allowing students and researchers to quickly perform studies of multi modal user interaction.

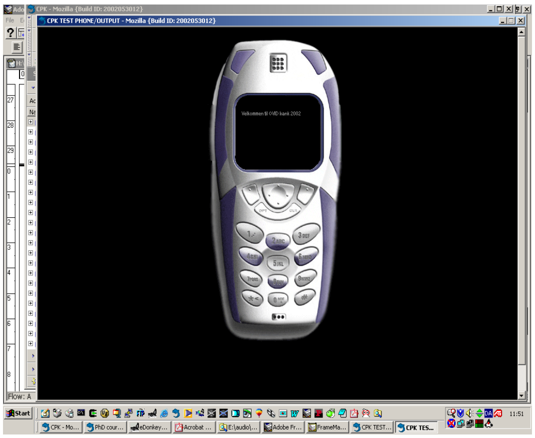

CPK - Mozilla {Build ID: 2002053012}
CPK TEST PHONE/OUTPUT - Mozilla {Build ID: 2002053012}
Velkommen til OVID bank 2002

FIGURE 20. The simulated “Fakephone” Webbased GUI used to display the textual output. The phone “display” will automatically update e.g. once per second and runs in any web browser.

Multi Modal user interaction has been in the focus of CPK research for a number of years, e.g. with the establishment of the Chameleon platform, a generic system for setting up intelligent multi media applications. One application that has been particular successful is the “Automatic Pool Trainer”, where speech, computer vision, graphics, laser and text are combined into a system for teaching pool. These applications are not subject of the present work, but of course have some relations to it. A list of publications about the research is included in Appendix C, page 107.

8.3.2 Future of Speech Based Interaction

One of the suggested causes for the problems that SDS are faced with is that many of the envisioned ‘killer apps’ for spoken interaction has turned out to be either too difficult to handle technically (e.g. operator assistance) or that another technology turned out to be preferred by users or to be more competitive from a commercial point of view.

Consider the domain of home banking systems. By the mid-nineties, the customer had the following alternatives: To visit a local branch or ATM, or to use a call-centre or an IVR service. ATMs and IVR services, while fully automated were (are) very limited in functionality. However, in 1997 the first PC-based home banking systems started to appear and five years later internet banking is the preferred mode for a large proportion of banking customers. In the same period, commercial speech technology became available for the Danish Language in 2000, but no Danish banks have to this date developed a home banking SDS. Furthermore, the Aalborg-based SparNord Bank[1] recently announced that it will close down it's WAP[2] based home bank solution due to lack of use. Clearly, telephone-based home banking, be it speech controlled or WAP has not been a great success.

In contrast, Danmarks Statistik (Statistics Denmark) found that by the end of 2002, 38% of the Danish internet users are using PC-based homebanking on a regular basis. 76% of the Danish population has access to a PC connected to the internet. (Danmarks Statistik 2003).

This is only one example, but it clearly illustrates the issues at stake: Spoken interaction must compete in a “market” with a large variety of options, both in terms of services and technologies, but also in terms of alternative modalities.

Spoken interaction will surely succeed in a number of areas, most likely in a combination with other modalities. But as Hugh Cameron points out: Users will only prefer speech, when it is more convenient and better than any other option. And speech will only become better when the usability of speech-based interfaces is carefully evaluated and optimised through application of the methods and techniques discussed here.

[1] SPARNORD: http://www.sparnord.dk/privat/netbank/om_netbank/features/artikler/wap_telefon.article

[2] WAP: Wireless Application Protocol - a protocol for wireless ‘light weigth’ internet access for mobile phones

Appendices

Appendix A Basic Concepts for Spoken Dialogue Systems

This appendix presents a brief introduction to the basic concepts of spoken dialogue systems and the associated terminology. The intention is to give an introduction to readers without prior knowledge of SDS and to define the terms and concepts used throughout the present report.

The introduction is purposefully kept brief and is only concerned with telephone-based systems. If a more thorough account of the issues presented here is necessary, please consult a good textbook, such as Daniel Jurafsky and James H. Martins "Speech and Language Processing" (Jurafsky and Martin 2000)

A.1 Spoken Dialogue Systems Architecture

Traditionally, SDS architectures are depicted as shown in Figure 1 Other configurations are of course possible, especially for multi modal or agent based systems. However, as a minimum, the functionality represented by the modules shown above must be present in an SDS.

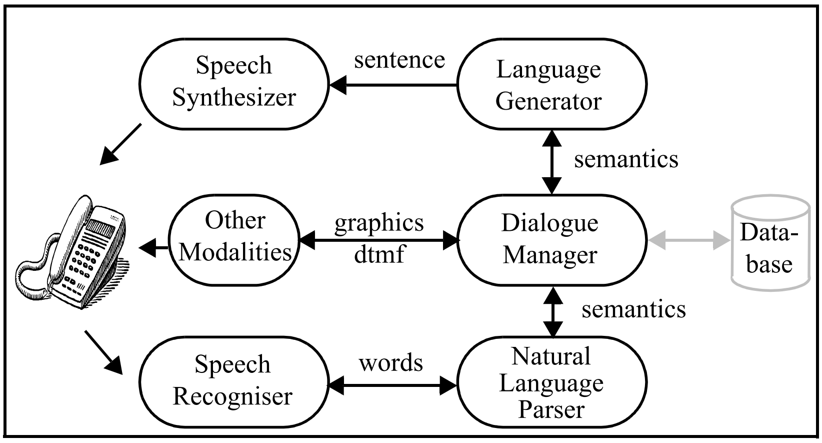


Figure 1. Generic architecture of a telephone based SDS.

The Speech Recogniser decodes the spoken utterances into a string of words, a word lattice, N-best list or similar. For simplicity Figure 1 only shows "words" as the output. Often, the speech recogniser consists for a number of modules, such as a front-end to reduce noise and detect barge-in. The speech recogniser uses a number of resources, most notably acoustic models, a lexicon and a grammar. Often the acoustic models will be statistic models of context-dependent phonemes, so-called triphones, and require extensive training. The lexicon contains information about the system application vocabulary and how it is transcribed. The grammar may explicitly contain (finite-state) syntax rules or be a statistical table of possible word combinations. In neither case is the grammar used to generate structure or extract meaning from the utterances, the purpose is only to constrain the search space of the speech recogniser.

A number of different speech recognition strategies are briefly presented here:

- I**solated words**. The user is only permitted to say one word at a time, or at least make a clear pause between words. This mode is not used in contemporary SDS any more, but is still used in very simple command-and-control applications, e.g. voice-dialling.
- **Word-spotting**. The speech recogniser only attempt to spot a reduced set of key words (typically less than fifty to a hundred words) within the users' utterance. This relaxes the constraints from isolated words considerably, but is not very robust for e.g. short or easily confused words. Likewise, the linguistic complexity must be kept low.
- **Phrase-spotting**. Similar to word spotting this mode does not attempt a full recognition of the users utterances, but spots for one or more key phrases. Examples could be credit card numbers (digit strings), dates, or destinations in a travel reservation domain. Phrase-spotting is more robust than word-spotting, since recognition of a string of word is more likely to succeed than just spotting single words. Phrase-spotting is more widely used than the previous modes and is often denoted **speech concept** recognition, since the phrases typically matches speech concepts.
- **Sentence (or full) recognition**. In this case, all the users' speech is recognised. This is the optimal mode, since in principle no information is lost. However, it is also more complex than the modes described above, both regarding the development of the domain lexicon and grammar and the computational complexity of the speech recogniser. Full recognition is obviously necessary to ensure optimal interpretation of the meaning of the utterance, especially if the domain is linguistically complex, where the partial recognition modes will be insufficient.

The Natural Language Parser. The task of this module is to interpret (parse) the result from the speech recognition process, extract the semantics of the user utterances and pass it on to the dialogue manager. For this a more powerful grammar formalism is often employed, e.g. based on unification, but it can also be based on a probabilistic approach. The parser may employ a number of robustness strategies, since spoken language is seldom completed nor well-formed. Depending on the circumstances, the parser may be more or less tightly integrated with the speech recogniser or the dialogue manager. It may be able to decode certain discourse phenomena, e.g. reference resolution. In the case of multi modal input, the parser (especially if it is unification based) can be utilised for modality fusion and e.g. generate unified semantic frames and pass them to the dialogue manager. In uncomplicated applications, the parser may degenerate to simple mapping of words in the input string to some semantic representation. The speech recogniser and natural language parser are sometimes collectively referred to as the **speech understanding** module.

The Dialogue Manager (DM). The primary task of the DM is to advance the communication towards the fulfilment of the users' intended goals. In other words, the objective is to control the interaction is such a way as to continuously elicit information from the user in order to eventually be able to provide the user with the desired information/functionality. The DM acts as the controller of the SDS and receives processed input from the user in the form of a semantic representation. Based on the information present in the system, the DM then either chooses to elicit further information from the user, or make a query to the domain database. It then generates a response to the user and passes it to the language generator and other output modules, if present.

The Language Generator. This module accepts information from the dialogue manager in an abstract form and generates well-formed sentences to pass to the speech synthesis module. The complexity of the language generator ranges from simple fill-in of values (e.g. for time of day) into predefined carrier sentences to generating fully synthesised sentences based on an output vocabulary and grammar. Apart from the generation of the sentence, the language generator may in some cases also annotate the string with prosodic cues e.g. contrastive stress, to help the user extract the meaning. However, the task of generating detailed prosodic belongs to the speech synthesizer module.

The Speech Synthesizer. The speech synthesizer generates the acoustic output from the SDS. It can either be based on playback of pre-recorded human speech, well-known from voice-response systems. In this case, the only task is to select appropriate pieces from a database of audio clips and play them back to the user. The voice quality is often highly intelligible, but often the prosody suffers and becomes unnatural (where two clips are joined together), and in some cases even speech samples from more than one person has been used. Purely synthetic or text-to-speech (TTS) systems are also used. Since the spoken output is fully synthetic, a much greater flexibility can be achieved. However, many TTS systems still sounds too unnatural to the human ear, even though the intelligibility might be good. Unfortunately, the bandwidth limitation imposed by the telephone network (fixed and wireless) apparently affects synthetic speech more than human speech.

Other Modalities. There are many indications that speech-only SDS are sub optimal. Therefore, many research activities within recent years have added additional modalities, for example a graphical screen or the possibility to use a stylus or telephone touch buttons (DTMF) as a supplement or alternative to speech. This often introduces some synchronisation problems, but also offers great potential and flexibility. However, the present work is mainly concerned with speech-only systems.

Since the primary concern is SDS usability evaluation seen from a user perspective, this brief introduction is considered sufficient to understand the issues discussed in this report. The following part of this chapter will introduce the most basic concepts of spoken dialogues.

A.2 Dialogue Control

The issue of who is in control of the interaction is a deciding factor of how the user experiences the dialogue. By control is meant who (the user or the SDS) decides what the next step is. In principle there are three different strategies: System directed, user directed and mixed initiative. A brief characteristic of each is given below:

System directed approach. The system retains the initiative throughout the interaction and at no point allow the user to take control. This will in practise mean that the user is not allowed to ask questions, and must always keep within the scope defined by the systems' questions. If the task is one in which the system requires a series of specific pieces of information from the user, the task may safely be designed as one in which the system preserves the initiative throughout by asking focused questions of the user. However, for experienced users this will often be perceived as too tightly constrained and can be very frustrating. Inexperienced users, on the other hand, might be prefer this mode, since it provides a very clear guidance, and the user is never in doubt of his/her options, as they will be clearly stated by the system.

User directed approach. A user directed dialogue is the mirror of the one mentioned above. In this mode, the system will be completely passive and all initiative lies with the user. If the task is extremely unstructured this mode might be usable, but there is a high risk that the interaction might become "stuck", since the system will not try to take the initiative and e.g. provide guidance to the user. Another situation where a user directed mode is appropriate is in very short question-answer dialogues, where the dialogue reduces to single questions or commands, unrelated to the previous interaction.

Mixed initiative approach. The most popular dialogue control mode is mixed initiative. In this mode, the initiative can shift between the user and the system arbitrarily or at predefined points in the dialogue. This has many advantages. For example, an experienced user will most likely prefer to take the initiative and thus control the interaction exactly in the direction s/he prefers. On the other hand, an inexperienced user might prefer to let the system keep the control and thereby be led through the dialogue.

The dialogue mode is of course dependent on the level of complexity the dialogue manager and the speech input processing modules are capable of handling. Obviously, the requirements for complexity in system directed dialogues is much lower than for mixed initiative or user controlled systems. In the latter cases, the system must be prepared to handle abrupt switches in context, whereas for the system directed case, the user's responses can easily be predicted (since s/he will typically have a limited set of response options, defined by the system). This can be utilised to e.g. constrain the speech recognisers' search space accordingly and thereby reduce the error rate. For many applications, where robustness is of high priority, this can be a substantial advantage. Examples are noisy environments like in-car systems or low-bandwidth cellular networks.

A.3 Spoken dialogue concepts

The atomic entities in a dialogue are the **turn** and the **utterance.**

A dialogue is composed by a sequence of turns, where first 'A' says something, then it is 'B's turn to says something, then 'A' again, etc. Humans use a variety of (often quite subtle) cues to determine when it is their turn. These are often based on intonation, pauses, but also on e.g. eye-contact and other non-verbal signs. Overlap between two speaker will occur to a limited degree, either by mistake (in case the turn taking cue was misinterpreted) or by one speaker deliberately interrupting another.

An utterance might in some ways be compared to the written sentence, but there are some important distinctions. Firstly, an utterance can span more than one turn and more sentences can be included in one utterance. Secondly, since the utterance is composed by natural, unconstrained speech, it is rarely well-formed (grammatically correct), and might contain phenomena like false starts, hesitations, short pauses, non-verbal (e.g. breath or lip) sounds etc.

A discourse is a series of coherent or related sentences, or for spoken dialogues, utterances. An example dialogue from the OVID corpus (translated from Danish[1]) is given below:

```
OVID 1  Welcome to the OVID banks automatic telephone service.
        Please state your card number.
user 1  9643267
OVID 2  Thank you.
        Please state your access code.
user 2  8648
OVID 3  Do you wish information about the balance or latest
        movements on your account?
user 3  the balance on the budget account
```

[1] See "Transcription of Dialogue B:" on page 185.

```
OVID 4   the balance on your budget account is <amount>[1]
         Do you wish further information about the balance on
         your primary account, your budget account or your cash
         credit account?
user 4   No, I'd rather have some movements on my budget account
OVID 5   The latest movements on your budget account are
         <amount1, date1>, <amount2, date2>, <amount3, date3>.
         Do you wish further information about the movements on
         your primary account, your budget account or your cash
         credit account?
user 5   No - I'm fine, thank you very much. Good bye
OVID 6   Do you wish to continue?
user 6   No - Good bye
OVID 7   Thank you for the call - good bye
```

The example clearly constitutes a discourse, since the utterances are related. It consists of at total of thirteen turns, seven system and six user turns. There are no utterances spanning more than one turn, but many (turns number 1,3,4,5) of the systems (OVID) turns consist of two utterances. Only one of the user turns (number 5) contains two utterances. The dialogue mode is mixed initiative. The system has the initiative in the first three turns. In turn number four and five the user takes the initiative, which passes back to the system again in the two final turns.

[1] For reasons of clarity, the actual amount have been removed.

Appendix B Factoring and Multiple Linear Regression

This appendix provides some additional results and definitions on some of the methods used Chapter 6 and Chapter 7.

B.1 Cronbachs α

Cronbachs α is a measure of the internal consistency among the statements in a questionnaire. A high value indicates a strong consistency, which is desirable.

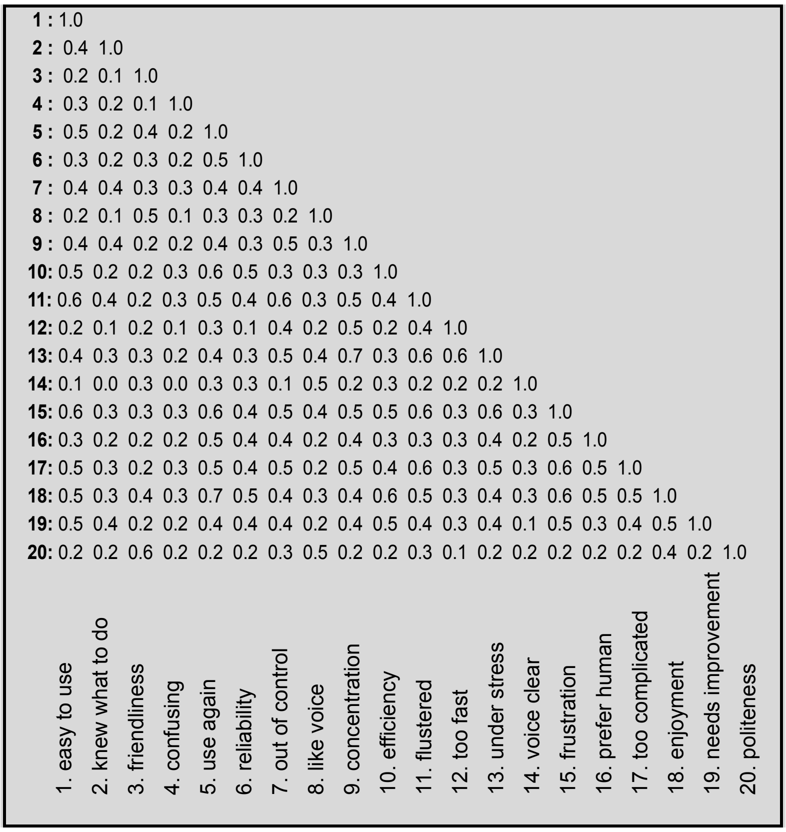
$$Cronbachs\ \alpha = \left(\frac{k}{k-1}\right) \cdot \left[1 - \sum_{i=1}^{items} s_i^2 / s^2{}_{sum}\right]$$

(EQ 1) i denotes the items (statements), *k* denotes the number of samples (the number of test users), $s^2{}_i$ is the item variance and $s^2{}_{sum}$ the total observed variance (estimate of true variability).

Given a typical (around 20-30) number of statements (*i*) in a test, high reliabilities (0.85) for Likert scales can often be achieved (Oppenheim 1966). One potential weakness of the Cronbach reliability measure is that there is an inherent relationship between the number of items and α, as α will increase with larger number of items. Therefore, one must take this into account when comparing Cronbachs α for different questionnaires, or when comparing a subset of the items to the overall questionnaire. Likewise, the number of different answer-categories (anchors) can be expected to influence the measure, as this obviously influences the variance. However, according to (Aiken 1996) this seems not to be the case, except for very low numbers of categories (two to three).

B.2 Factor Analysis

This section includes further details of the Factor Analysis of the OVID questionnaire, as well a brief definitions of some important concepts and variables associated with FA. It is largely based on the very comprehensive textbook: “Using Multivariate Statistics” by Barbara Tabachnick and Linda Fidell (Tabachnick and Fidell 2001). All calculations and generation of figures are carried out using the Statistics Toolbox in MATLAB (Mathworks 1999).



1 :	1.0																			
2 :	0.4	1.0																		
3 :	0.2	0.1	1.0																	
4 :	0.3	0.2	0.1	1.0																
5 :	0.5	0.2	0.4	0.2	1.0															
6 :	0.3	0.2	0.3	0.2	0.5	1.0														
7 :	0.4	0.4	0.3	0.3	0.4	0.4	1.0													
8 :	0.2	0.1	0.5	0.1	0.3	0.3	0.2	1.0												
9 :	0.4	0.4	0.2	0.2	0.4	0.3	0.5	0.3	1.0											
10:	0.5	0.2	0.2	0.3	0.6	0.5	0.3	0.3	0.3	1.0										
11:	0.6	0.4	0.2	0.3	0.5	0.4	0.6	0.3	0.5	0.4	1.0									
12:	0.2	0.1	0.2	0.1	0.3	0.1	0.4	0.2	0.5	0.2	0.4	1.0								
13:	0.4	0.3	0.3	0.2	0.4	0.3	0.5	0.4	0.7	0.3	0.6	0.6	1.0							
14:	0.1	0.0	0.3	0.0	0.3	0.3	0.1	0.5	0.2	0.3	0.2	0.2	0.2	1.0						
15:	0.6	0.3	0.3	0.3	0.6	0.4	0.5	0.4	0.5	0.5	0.6	0.3	0.6	0.3	1.0					
16:	0.3	0.2	0.2	0.2	0.5	0.4	0.4	0.2	0.4	0.3	0.3	0.3	0.4	0.2	0.5	1.0				
17:	0.5	0.3	0.2	0.3	0.5	0.4	0.5	0.2	0.5	0.4	0.6	0.3	0.5	0.3	0.6	0.5	1.0			
18:	0.5	0.3	0.4	0.3	0.7	0.5	0.4	0.3	0.4	0.6	0.5	0.3	0.4	0.3	0.6	0.5	0.5	1.0		
19:	0.5	0.4	0.2	0.2	0.4	0.4	0.4	0.2	0.4	0.5	0.4	0.3	0.4	0.1	0.5	0.3	0.4	0.5	1.0	
20:	0.2	0.2	0.6	0.2	0.2	0.2	0.3	0.5	0.2	0.2	0.3	0.1	0.2	0.2	0.2	0.2	0.2	0.4	0.2	1.0
	1. easy to use	2. knew what to do	3. friendliness	4. confusing	5. use again	6. reliability	7. out of control	8. like voice	9. concentration	10. efficiency	11. flustered	12. too fast	13. under stress	14. voice clear	15. frustration	16. prefer human	17. too complicated	18. enjoyment	19. needs improvement	20. politeness

Figure 1 Correlation matrix for the 20 core statements. 60% of the correlations are above 0.3. Only two ((14,2) and (14,4)) coefficients are not significantly different from zero at the 95% confidence level.

Correlation Matrix. The basic assumption for FA is that there exists a relationship between the observed variables, caused by an underlying set of factors. In the present case, the variables are the statements in the questionnaire, and the observations are the scores each user has assigned to the particular statement. Therefore, the first step is to inspect the correlation matrix **R** to verify the degree of correlations among the statements. If few or none of the correlation coefficients are above 0.3 the data is not suitable for FA. However, as can be seen from Figure 1 this not the case here.

Estimation of Factor Loadings. Matlab computes a Maximum Likelihood Estimate (MLE) of the Factor loadings matrix Λ and the unique (or specific) variance Ψ, from the covariance matrix of the statement scores *ss*:

(EQ 1) $$cov(ss) = \Lambda'\Lambda + \Psi$$

The factor loading matrix contains regression-like weights used to estimate the contribution of each factor to the variance in a variable (statement), i.e. the correlations between variables and factors[1]. The specific variance is of interest, since it represents the remaining variance for each variable, which is not included in the FA. Variables with a high specific variance can not be expected to be well modelled by the FA.

Factor Loadings	F_1	F_2	F_3	F_4	F_5	Ψ
1. easy to use	0.47	0.10	0.54	0.05	0.07	0.47
2. knew what to do	0.12	0.16	0.52	0.11	-0.01	0.68
3. friendliness	0.24	0.15	0.05	0.60	0.28	0.48
5. use again	0.71	0.19	0.20	0.12	0.15	0.38
6. reliability	0.52	0.11	0.18	0.13	0.15	0.64
7. out of control	0.27	0.44	0.50	0.18	-0.07	0.44
8. like voice	0.13	0.11	0.11	0.39	0.86	0.07
9. concentration	0.26	0.61	0.33	0.05	0.13	0.43
10. efficiency	0.64	0.05	0.24	0.07	0.22	0.47
11. flustered	0.36	0.32	0.63	0.08	0.15	0.34
12. too fast	0.10	0.69	0.12	0.08	0.06	0.49
13. under stress	0.22	0.76	0.32	0.08	0.23	0.21
14. voice clear	0.24	0.14	-0.00	0.13	0.51	0.65
15. frustration	0.59	0.29	0.46	-0.04	0.25	0.30
16. prefer human	0.50	0.36	0.08	0.07	0.03	0.60
17. too complicated	0.43	0.36	0.45	0.06	0.05	0.47
18. enjoyment	0.80	0.19	0.20	0.27	0.06	0.22
19. needs improve.	0.44	0.16	0.36	0.07	0.12	0.63
20. politeness	0.10	0.05	0.17	0.84	0.19	0.22
Captured Variance	18.5	11.9	11.7	7.5	7.3	

Table 1 Resulting Factor Loading Matrix of a FA with 5 factors[a]. The highest loading factors are shown in grey for all statements.

[a] Note that statement 4 has been removed, because of very low loadings and high specific variance.

[1] Only in the case of orthogonality between factors, see the discussion on rotation below

Table 4 shows the factor loading matrix (Λ) for the 20 core statements, except one. The results are summarised and compared to the CCIR FA in Table 8 on page 55. Note that there are a number of statements that are cross-loading, i.e. has loadings of approximately equal size on two factors. Ideally, each statements should only load (correlate) with one factor, but this is seldom the case in real-world situations.

Rotation. The matrix shown in Table 4 is rotated in order to reduce cross-loading, and hence make the interpretation easier. Rotation can be either orthogonal or oblique, but in most cases orthogonal rotation is chosen. Oblique rotation has the side effect that factors become correlated, and the resulting matrices are harder to interpret.

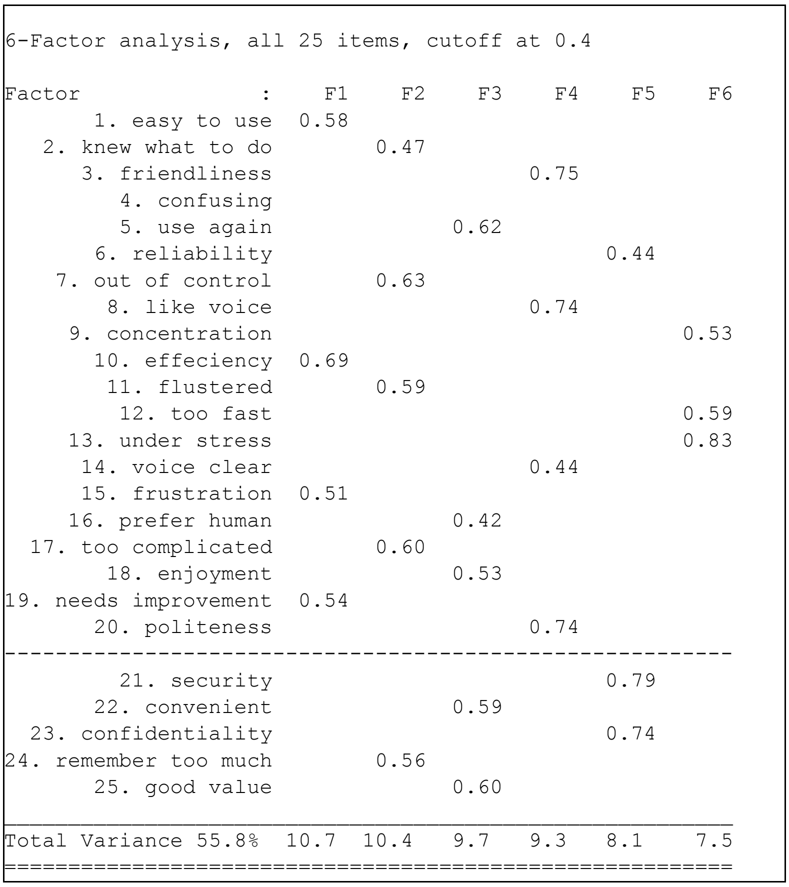
Several methods exist to aid the determination of the optimal number of factors. The most well-known are the Kaiser criterion and Cattell's Scree plot.

The Kaiser Criterion. This criterion assumes that a PCA is carried out on a correlation matrix C of the original items. The Eigenvalues of C will be equal to the variance of principal components. The Kaiser criterion simply requires that only components with variances above one are retained. This makes sense insofar as it would seem pointless to retain components with less variance than the original items.

The Scree[1] Plot. The Scree test is based on a plot of the eigenvalues. A graphical inspection can identify where the magnitudes of the eigenvalues levels off to the right of the plot and becomes nearly horizontal (the "scree"). The eigenvalues to the left of this point are retained. An example of a Scree plot is shown in Figure 12 on page 54

Factor Analysis of the full questionnaire. Table 4 showed a FA with only the core set of statements, in order to compare with a similar questionnaire from (Love et al 1994) and thereby validate the questionnaire. However, it is interesting to investigate the structure of the full questionnaire, including the domain-specific statements. The result is shown in Figure 2 below.

[1] The name "Scree" refers to the area at the foot of a mountain side with rubble and stones. Thus the reference here is to the area where the factors "level of" and the curve becomes flat.



```
6-Factor analysis, all 25 items, cutoff at 0.4

Factor              :    F1    F2    F3    F4    F5    F6
       1. easy to use  0.58
  2. knew what to do        0.47
     3. friendliness                    0.75
        4. confusing
         5. use again              0.62
       6. reliability                          0.44
   7. out of control        0.63
        8. like voice                   0.74
     9. concentration                                0.53
       10. effeciency  0.69
        11. flustered       0.59
         12. too fast                                0.59
     13. under stress                                0.83
      14. voice clear                   0.44
      15. frustration  0.51
     16. prefer human              0.42
  17. too complicated       0.60
        18. enjoyment              0.53
19. needs improvement  0.54
       20. politeness                   0.74
---------------------------------------------------------
         21. security                          0.79
       22. convenient              0.59
  23. confidentiality                          0.74
24. remember too much       0.56
       25. good value              0.60

_________________________________________________________
Total Variance 55.8%  10.7  10.4   9.7   9.3   8.1    7.5
=========================================================
```

Figure 2 FA with 6 factors and 25 statements. The captured variances are shown in the bottom line. Only loadings above 0.4 are shown and cross-loadings have been removed (4). The total capture variance is 55.3%

When all statements are included in the FA, a six factor structure is found to yield the best factor structure. Figure 2 shows the resulting factor loading matrix, with loadings below 0.4 and cross loadings removed. Note that statements 4. “Confusing” does not load on any factor above the threshold - the load is distributed across several factors. The labels and statements assigned to each factor is shown in Table 10 on page 68.

B.3 Multiple Linear Regression

Similar to the previous section on FA, some background and additional results are provided here on MLR used to estimate the coefficients in the PARADISE performance function. The definitions are based on (Tabachnick and Fidell 2001), and a set of excellent lecture notes by Dave Meko (Meko 2003). Microsoft Excel and Matlab are used for the calculations and generation of figures.

Initial selection of regression parameters. A number of regression experiments with various configurations of the factor clusters indicated that the F_1 cluster[1] was (slightly) better estimated by the MLR and was chosen for further analysis. These experiments all show approximately the same trends and resulting coefficients, and will not be discussed further here.

	Beta	St.Error	P-value
Kappa	0,53	0,17	0,0047
task success	-0,31	0,19	0,1137
Recognition	0,61	0,24	0,0174
Total Turns	-0,19	0,47	0,6844
Total Time	0,17	0,55	0,7592

Regression Statistics	
R^2	0,50
Adjusted R^2	0,41
Standard Error	0,76

Table 2 Result of initial Regression for 35 selected users and 5 independent variables. The Betas are the coefficients estimated by the regression.

Table 2 shows the statistics for an initial regression with five (normalised) parameters (c.f. Chapter 4 for a description of the measures). The Betas are the regression coefficients, shown with standard error and P-values. The table shows that Kappa and Recognition are significant predictors ($p < 0.05$), while the others are not. This is also indicated by the numerically smaller Beta coefficients (which are directly comparable due to the Z-normalisation of the variables), and the larger error coefficients for task success, Turns and Time. Of the regression statistics, the adjusted R^2 is the most interesting, since it represents the amount of variance in the independent variable explained by the regression. In this case it is only 0.41, or 41%[2] Thus, Kappa and Recognition

[1] F_1 is the factor cluster concerning statements about "Quality of Interface/Performance". See Table 8 on page 55

[2] The R^2 is adjusted to compensate for the number of independent variables that exceeds one. This is done to enable comparison across different numbers of parameters, since a higher number of variables can always be expected to capture a larger portion of the variance, without the regression actually being better.

are chosen as the independent variables for the MLR and F_1 as the dependent variable.

In order to establish the degree of normality of the variables a graphic inspection is often useful.

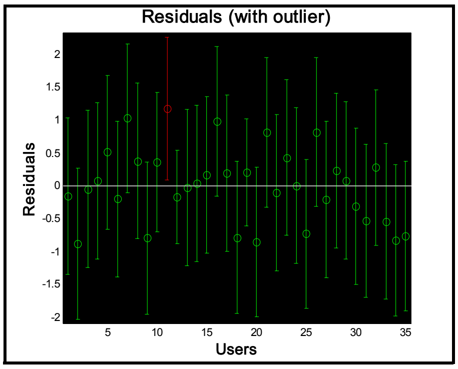


Figure 3 This plot shows the residuals with 95% confidence intervals. One outlier is indicated (in red) and is removed before the final regression is carried out.

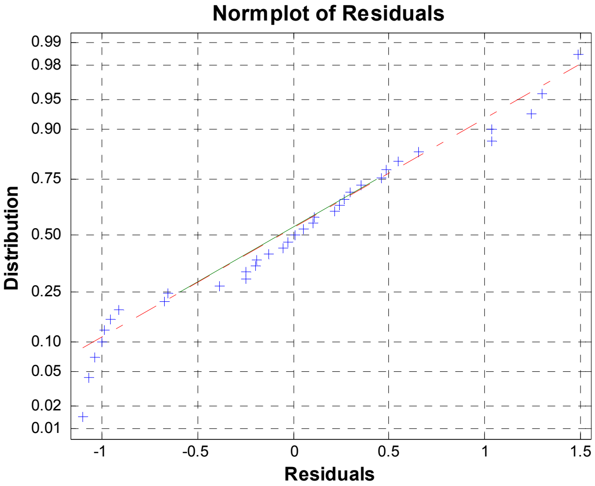
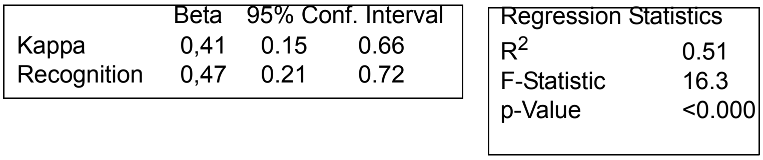


Figure 4 The residuals are shown in a normal plot, where a normally distributed variable would lie on a straight line. This is only partly the case here.

Figure 3 and Figure 4 shows the residuals (difference between the predicted and the observed) values of F_1. The outlier shown in Figure 3 is removed before the final regression is carried out. Figure 4 indicates that some non-linearity is present, since the values are not on a straight line. A similar plot is shown for the dependent variable, F_1 in Figure 16 on page 71.

Common remedies for non-linearity is to apply various transformations, e.g. squared or logarithmic. However, this would make a comparison with similar performance equations impossible. Since the main purpose of the MLR here is to compare the resulting performance equation with the corresponding ones, obtained by Walker and colleagues at AT&T, it is decided not to attempt linearisation through a transformation of variables.

The final result is shown in Table 3 below.



	Beta	95% Conf. Interval	
Kappa	0,41	0.15	0.66
Recognition	0,47	0.21	0.72

Regression Statistics	
R^2	0.51
F-Statistic	16.3
p-Value	<0.000

Table 3 Result of the final regression for 34 users and 2 independent variables. The Betas are the coefficients estimated by the regression.

The table shows that 51% of the variance is explained by the model. The coefficients are significant predictors of the dependent variable, but as can be seen the confidence intervals for the identified beta coefficients are fairly large. Figure 17 on page 73 shows the observed and predicted values for F_1.

B.4 PARADISE usability questionnaire

Statement	Usability Aspect
Was ELVIS easy to understand in this conversation?	TTS Performance
In this conversation, did ELVIS understand what you said?	ASR Performance
In this conversation, was it easy to find the message you wanted?	Task Ease
Was the pace of interaction with ELVIS appropriate in this conversation?	Interaction Pace
In this conversation, did you know what you could say at each point of the dialogue?	User Expertise
How often was ELVIS sluggish and slow to reply to you in this conversation?	System Response
Did ELVIS work the way you expected him to in this conversation?	Expected Behaviour
In this conversation, how did ELVIS'S voice interface compare to the touch-tone interface to voice mail?	Comparable Interface
From your current experience with using ELVIS to get your email, do you think you'd use ELVIS regularly to access your mail when you are away from your desk?	Future Use

Table 4 Usability questionnaire used in (Walker et al 1998, p.18)

Appendix C Articles and Reports

This appendix contains six articles and one technical report published in the period 1996-1999 and 2002-2003. A brief summary of each is presented below to aid the reader in selecting the appropriate article. The articles appear exactly as originally published, except that they have been reformulated to fit the format of the present report. Thus, some minor inconsistencies might be found in e.g. notation, number of users and recorded dialogues and other details, as they represent work in progress covering a span over several years.

A list of more recent articles (not part of the present thesis) are included in order to complete the list. The articles are concerned with on multi modal user interaction and are not part of the thesis however.

Included Articles

"Voice Controlled Home Banking - Objectives and Experiences of the Esprit Ovid Project". *In proceedings of the IVTTA-96 workshop, Murray Hills, NJ. 1996*. (on page 111 ff).

This paper describes the methods applied in the requirements gathering phase, and the resulting specification of the functionality and dialogue style for the prototype: The findings of the requirements capture process are reported and discussed. Most notably, a number of requirement interviews uncovered that almost identical requirements exist for all the OVID banks. This was the case even though the banks have very different profiles and market strategies. This observation suggests that the conclusions may be extended to cover voice controlled home banking services in general

"A Strategy for Mixed-initiative Dialogue Control ". *In the Proceedings of Eurospeech '97, Rhodes, Greece 1997.*(on page 119 ff.)

This paper presents the dialogue model applied within the OVID project and in particular discusses a strategy for mixed-initiative dialogue management. The strategy utilises the guidance of system-directed dialogues, while accommodating user initiated focus shifts by the inclusion of *short-cuts* in the dialogue. The paper reports on the two OVID experiments, one with a simulated speech recogniser (WOZ), and the second with a fully automated system (results are only reported from a a subset of the dialogues). Both experiments shows that users use the possibility for *short-cuts*, even when not instructed of their existence. A tendency towards user habituation is also demonstrated.

"Investigating a Mixed-Initiative Dialogue Management Strategy". In the *Proceedings of the IEEE Workshop on Speech Recognition and Understanding, ASRU 1997*, Santa Barbara, Cal. (on, page 129 ff.)

Like the previous paper, this one focuses on the issue of designing mixed-initiative spoken dialogues with only a partial recognition of the user utterances (recognition of concepts or phrase spotting). A mixed-initiative dialogue management model has been developed and implemented. The paper analyses the turn-taking in details and use the results to conclude that users are able to grab the initiative at natural points in the dialogue. Automatic speech recognition performance is compared to the results from the subjective evaluation, but does not show a clear correspondence between user satisfaction and ASR performance. This paper reports on the full dialogue corpus.

"The OVID Project Objectives and Results". Technical Report 98-0201 CPK Aalborg University, 1998 (on page 137 ff.)

The report covers the results presented in the three papers described above, but goes into more detail concerning the applied methods. Furthermore, the report contains a number of appendices with the Danish and British versions of the user attitude questionnaire, a full list of all OVID deliverables, log files and summaries from example dialogues and additional background information. While to some degree redundant to the three papers described above it has been included because it: a) provides more in-depth details and b) presents the results in a more fluent way than three separate papers.

"Combining Objective and Subjective Data in Evaluation of Spoken Dialogues", In the *Proceedings of the ESCA ETRW on Interactive Dialogue Systems,* Kloster Irsee, Germany 1999 (on page 187 ff.)

This paper combines the evaluation based on objective data obtained from log files with subjective data, where the test subjects express their attitudes to a number issues directly related to the usability of the service. It is shown how the joint analysis can be used to support, but also question findings from either source.

The two final papers included in Appendix C are set apart from the previous ones by the fact that they present much more recent work (2002-03) on the corpus. The subjects are covered in the main part of the report in more details and the articles are included in order to give the reader a more condensed version of the recent findings.

"Assessment of Spoken Dialogue System Usability - What are We really Measuring?". *To appear in the proc. of Eurospeech'03* (on page 197 ff)

This paper attempts to clarify some of the reasons why the penetration of speech-based applications has not reached it's expected success 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. Thus, the paper revolves around the themes discussed in Chapter 2 and Chapter 5.

"Applying The PARADISE Evaluation Scheme to an Existing Dialogue Corpus". Submitted to *ASRU'03* (on page 205 ff)

This paper presents results and conclusions about the current evaluation methodologies for Spoken DIalogue Systems (SDS). The PARADISE paradigm, used for evaluation in the DARPA Communicator project is briefly introduced and discussed through the application to the OVID home banking dialogue system. It is shown to provide results consistent with those obtained by the DARPA community, but a number of problems and limitations are pointed out. The issue of user attitude measures through questionnaires is discussed. This is an area that have not received much attention from the speech technology community, but is important in order to obtain valid results and conclusions about usability. The paper gives a condensed version of the results and discussions presented in Chapter 5 to Chapter 7

Recent publications, not included in the present report

Conference Papers

"Multi Modal User Interaction in an Automatic Pool Trainer". Lars Bo Larsen, M. D. Jensen, W. K. Vodzi, "Multi Modal User Interaction in an Automatic Pool Trainer", Fourth IEEE International Conference on Multi modal Interfaces (ICMI 2002), October, 2002

"A Multi Modal Pool Trainer". Lars Bo Larsen, Tom Brøndsted, Proc. of the International Workshop on Information Presentation and Natural Multimedia Dialogue - IPNMD-2001, Verona Italy, December, 2001, pp. 107-111

"Issues on Globalisation of Engineering Educations". Lars Bo Larsen, F.K. Fink, Proceedings of SEFI annual conference 2000, Paris September 2000

"Building Affective Robots with LEGO Mindstorms". Lars Bo Larsen, J. Bang, T. Madsen Proceedings of Affect in Interactions - towards a new generations of interfaces, October, 1999

"CHAMELEON: a general platform for performing intellimedia"[1]. Tom Brøndsted, Paul Dalsgaard, Lars Bo Larsen, Michael Manthey, Paul Mc Kevitt, Thomas B. Moeslund, Kristian G. Olesen, The Eight International Workshop on the Cognitive Science of Natural Language Processing, pp. 110-122, August, 1999, Galway, Ireland

"Setting Up a Masters Programme in Intelligent Multi Media - Approach and Applications". Thomas. Moeslund, Lars Bo Larsen, Proceedings of: The 11th Scandinavian Conference on Image Analysis - SCIA '99, June 1999, Søndre Strømfjord, Greenland, June, 1999

"CHAMELEON: a general platform for performing intellimedia"[1]. Tom Brøndsted, Paul Dalsgaard, Lars Bo Larsen, Michael Manthey, Paul Mc Kevitt, Thomas Moeslund, Kristian G. Olesen, Proceedings of the Ninth Irish Conference on Artificial Intelligence (AICS-98), August, 1998, pp. 73-90

"Combining Speech and Vision Processing in a Platform for Intelligent Multi-Media"[1]. Tom Brøndsted, Lars Bo Larsen, Michael Manthey, Paul Mc Kevitt, Thomas Moeslund, Kristian G. Olesen, The 7th Danish Conference on Pattern Recognition and Image Analysis, August, 1998, pp. 75-79

"Enhancing a WIMP based interface with Speech, Gaze tracking and Agents", L. Bakman, M. Blidegn, M. Wittrup, L.B. Larsen, T. B. Moeslund, Proceedings of International Conference of Spoken Language Processing, ICSLP 1998, Sydney, Australia 1998, December, 1998

"The Intellimedia WorkBench - a Generic Environment for Multi Modal Systems"[1], Tom Brøndsted, Lars Bo Larsen, Michael Manthey, Paul Mc Kevitt, Thomas Moeslund, Kristian G. Olesen, The 5th International Conference on Spoken Language Processing (ICSLP), October, 1998, pp. 273-276

"The Intellimedia WorkBench - an environment for building multi modal systems"[1], Tom Brøndsted, Lars Bo Larsen, Michael Manthey, Paul Mc Kevitt, Thomas Moeslund, Kristian G. Olesen. Proceedings of the Second International Conference on Cooperative Multi modal Communication, Theory and Applications, January, 1998, Tilburg, pp. 166-170

Journal Papers and Book Chapters:

"CHAMELEON: a general platform for performing intellimedia"[1], Tom Brøndsted, Paul Dalsgaard, Lars Bo Larsen, Michael Manthey, Paul Mc Kevitt, Thomas B. Moeslund, Kristian G. Olesen, Language, Vision & Music, John Benjamins, October, 2002, ISBN 90 272 5155 X pp. 79-96

[1] Authors are listed in alphabetical order

“Developing Intelligent Multimedia Applications”[1], Tom Brøndsted, Lars Bo Larsen, Michael Manthey, Paul Mc Kevitt, Thomas B. Moeslund, Kristian G. Olesen, Multi modality in Language and Speech Systems, Kleuwer, July, 2002, ISBN 1-4020-0635-, pp. 149-171

“The Internationalisation of Postgraduate Programmes”, F. K. Fink, O. K. Andersen, T. Bak, L. B. Larsen, Global Journal of Engineering Education, Vol. 6, No 2, UICEE, September, 2002.

“The Automated Pool Trainer - A multi modal system for learning the game of Pool”, Lars Bo Larsen, P. M. Jensen, K. Kammersgaard, L. Kromann. In: Intelligent Multi Media, Computing and Communications: Technologies and Applications of the Future. John Wiley and Sons ISBN 0-471-20435-8, June, 2001, pp.90-96

“The IntelliMedia WorkBench. An Environment for building multi modal systems”[1], Tom Brøndsted, Lars Bo Larsen, Michael Manthey, Paul Mc Kevitt, Thomas B. Moeslund, Kristian G. Olesen, Cooperative Multi modal Communication, Springer Verlag, 2001, pp. 217-233

Reports

“A platform for developing Intelligent Multi Media Applications”[1], Tom Brøndsted, Paul Dalsgaard, Lars Bo Larsen, Michael Manthey, Paul Mc Kevitt, Thomas Moeslund, Kristian G. Olesen, Technical Report R-98-1004, May, 1998, CPK, Aalborg University, 157 pages.

“Evaluation Methodologies for Spoken and Multi Modal Dialogue Systems”. Lars Bo Larsen, Research Report, COST278 (in progress)

[1] Authors are listed in alphabetical order

"Voice Controlled Home Banking - Objectives and Experiences of the Esprit Ovid Project"
In the proceedings of IVTTA-96 workshop, September, 1996.

Voice Controlled home banking - objectives and experiences of the ESPRIT OVID project

Lars Bo Larsen
Center for PersonKommunikation, CPK
Aalborg University
DK-9220 Aalborg, Denmark
Email: lbl@cpk.auc.dk

Abstract

This paper reports on the ESPRIT OVID[1] project. It describes the objectives behind the project, and the results up till now.

The OVID project deals with the task of phone based home banking services. The findings of the requirements capture process are reported in detail. Most notably a number of requirement interviews uncovered that almost identical requirements exist for all the OVID banks. This was the case even though the banks have very different profiles and market strategies. This observation suggests that the conclusions may be extended to cover voice controlled home banking services in general.

1. Introduction

The first part of the present paper is devoted to a general description of the OVID project. In particular the background and motivation of the project are described. The following sections cover the application in more detail and present the requirements capture methodology and results. The planned user trials are described briefly and a short introduction to the implementation of the resulting dialogue description are presented.

1.1. Background

Two factors play an important role for the formulation of the OVID project. A business objective and a technical objective.

The technical objective is to evaluate the current state-of-the-art within the rapidly growing field of commercial voice technology applications. That is, to asses the usability of the technology available now or in the very near future in a domain which is expected to offer a potentially large number of applications for voice technology. The business objective for the banks is to offer new and more efficient services to their customers. Touch tone telephone banking systems have been in use for more than a

[1] The OVID Esprit 20717 Project consortium comprises The Royal Bank of Scotland and Barclays Bank in the U.K., Lån & Spar Bank in Denmark, CCIR Edinburgh University, U.K, Center for PersonKommunikation, Aalborg University, Denmark, Brite Voice Technology U.K., and AGORA Consult, France as coordinating partner.

decade, but have changed little over that period in functionality and user interface. In the same period, call centres have emerged.

Therefore the commercial incentive for the banks to switch to voice technology is very high. With a cost reduction of 90% for transactions via call centres compared to ordinary branches, and a further reduction of 90% for fully automated interactive touch tone voice response (IVR) transactions, there is a total cost reduction of 90-99% for each transaction that the bank can relocate from branches to an automated service. Furthermore, indications are that more than 80% of all transactions are suitable for automation [1] Thus the banks have a very high motivation in making automated services as attractive as possible. One way to do that is to replace current IVR and call centre (CCS) operated services with voice controlled technologies.

Although the business strategies differ substantially for the three banks, the requirements analysis uncovered almost identical demands for the voice controlled system. The results are discussed in more detail in section 3

1.2. The Project Consortium

The project consortium comprises British and Danish banks, research institutions and an industrial voice technology provider. The banks differ in a number of aspects. The British banks are well established with large number of branches, whereas the Danish bank is small measured in branches and employees, and focuses directly towards a business strategy centred on telephone based services. Either as IVR, PC-based or via a branch only accessible via the telephone.

The British banks have established call centres, which handle a growing proportion of customer transactions, and recently also IVR services. Call centres does not exist in Denmark at all. Because of this, the outlook of the bank partners differs. For the Danish bank, introducing voice technology would mean increasing the functionality and attractiveness of an already automated service, whereas the British banks are looking for ways to automate or supplement current CCS services.

The role of the research institutions is to bring expertise within speech recognition and spoken dialogue systems into the project. As no new technology is being developed within the project, the tasks are centred upon establishing, testing and evaluating the specified trial dialogues. The technology provider will bring current commercial leading edge telephone voice system technology into the project.

2. Application Domain

The application domain of the OVID project is within voice controlled telephone based home banking. Basically, a customer call to a telephone banking system - either IVR or CCS involves four phases.

- *Customer identification.* The customer typically identifies him- or herself to the system by supplying a 7-12 digits number.
- *Customer Verification.* A password consisting of 2-5 digit from the customers' PIN is required for caller identity verification.
- *Balance and account status.* Most often, the customer enquires about a balance and recent activity on the account.
- *Transactions.* A more complicated dialogue where the customer requests among a possible set of transactions, e.g transfers money to other accounts, or sets up standing orders, etc.

Not all calls involve the last type of transaction. Often the customer only wants information on a balance or whether e.g. a certain payment has taken place. This is discussed in more detail in section 3 below.

3. Results of the User Requirement Capture

This section describes the findings of the user requirements capture in detail. In the present context, the "user" is the bank, not to be confused with the end-users of the service, who are referred to as customers.

3.1. Methodology

Two sources of information have formed the basis for the formulation of the user requirements. The first concerns data collected by the banks on existing IVR systems and CCS on statistics of transactions types, duration of calls, customer profile (age, gender, calls per month, etc.). This source gave accurate statistical information on the usage of the current services that the voice controlled service is intended to replace.

The other source of information was an extensive series of semi-structured interviews with representative personnel from the three banks. The interviews allowed the interviewees to answer freely within a broad framework of the service domain. By using this approach, the interviewees were not constrained to only answer very specific predefined questions, but were allowed to introduce new issues not anticipated by the interviewer, such that novel ideas and concepts were likely to be uncovered.

3.2. Structuring of the interviews

The questions included in the interviews were concerned with the current situation with telephone banking for each bank and what the requirements for the OVID trials should be. To systemise the interviews, the "Six

Witches” method, developed at CCIR [1], was adopted. The technique is built upon the *Who? What? Where? When?* and *Why?* questions of jounalism extended with a *How?*, and provides the framework for the semi-structured interviews.

The following part of this section presents the resulting user requirement specifications obtained by the approach. In total, 24 key user requirements were identified through 27 interviews and grouped according to the Six Witches.

Who?

The questions concerns the gender, accent, habituation and age profiles of the potential customers. All banks report an almost equal distribution between male and female callers. Main Accent types Table 1 shows the main accent types anticipated by the banks. Especially for the British banks it is observed that accents are an important factor, and must be taken into consideration when designing the speech recogniser. For the Danish bank, regional accents must be taken into account.

Accent	**Lån & Spar Bank**	**Royal Bank of Scotland**	**Barclays Bank**
Native Danish	98%	-	-
Native English	-	60%	70%
Scottish English	-	35%	5%
Other	2%	5%	25%

Table 1 Main Accent types

The interviewees assessed that customers might experience difficulties at first but that they can be assumed to be ‘experienced’ users after having used the service two to three times. Very few customers use an informal style of address, and consequently the service should maintain a formal mode of addressing. The distribution according to age of the customers is similar for all the banks with a strong dominance for younger customers (age groups 20 - 40 years).

Why?

The question of Why? explains why customers use or stop using the existing IVR and CCS services. The interviewees were asked to rank the reasons why (in their opinion) customers use the services. The results are shown in Rank order for service featuresTable 2 below.

Rank	**Lån & Spar Bank**	**Royal Bank of Scotland**	**Barclays Bank**
Highest	Convenience	Convenience	Convenience
	24-hour Service[a]	24-hour Service	24-hour Service
	Speed	Speed	Speed
	Security	Operator helpful	Security
	Informative	Security	Confidentiality
	Confidentiality	Confidentiality	Informative
Lowest	Operator helpful	Informative	Operator helpful

Table 2 Rank order for service features

[a] The need for a 24 hour service cannot be documented as the present IVR service closes between 00 and 04 hours (c.f. Figure 1).

The table shows a high degree of agreement across the banks, and the three highest ranking features all relate to convenience and availability of the services. The most common reasons given by customers to stop using the service are loss of confidence and lack of functionality.

When?

The question *When?* discusses the times and durations of calls to the services. The average duration for CCS calls is 2.5 minutes. More complex calls may take up to five minutes. Calls made in the evening tends to be longer than those made during day time. The average duration for IVR calls are as low as one minute. Customers typically call once or twice per month, with more calls towards the end of the month. A few customers use the service very often. Average call densities during the day. shows the distribution of the average call density during the day, and it is observed that a large proportion of the calls are made outside normal banking hours. The service must be able to handle a peak of no less than 10% of the expected daily calls at any time.

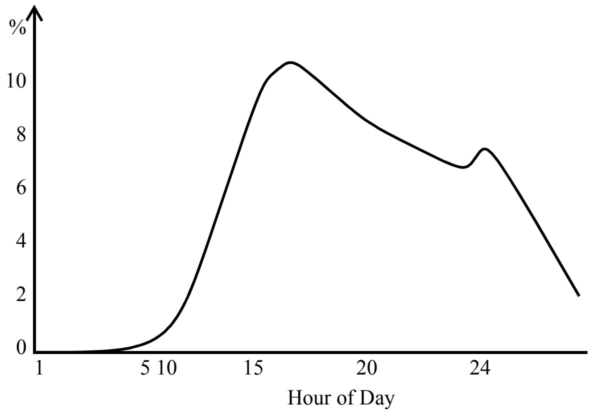


Figure 1 Average call densities during the day.

What?

The question of *what?* discusses the types of transactions handled by the telephone banking services.

All the three banks require the customer to supply identification and security information. Typically the identification requires 7 to 12 digits and a security PIN. This is typical a mixture of 2 to 5 digits and alphanumericals. Transaction Densities for OVID banks Table 3 shows the average densities for the transaction types. For example, 93% of all calls to

Rank	Lån & Spar Bank	Royal Bank of Scotland	Barclays Bank
Balance Enquiry	93%	54%	38%
Account Enquiry	42%	43%	22%
Bill Payments	n/a	28%	21%
Transfer Own acc	28%	9%	8%
Transfer 3rd party	8%	n/a	n/a
Transfer Giro	10%	n/a	n/a
Order Statement	2%	1%	2%
Direct Debit	n/a	1%	2%
Exchange Rates	3%	unknown	unknown
Change Password	2%	unknown	unknown
Standing Orders	n/a	n/a	3%
New Checkbook	n/a	1%	n/a

Table 3 Transaction Densities for OVID banks

Lån & Spar bank involves a balance enquiry and 8% involves a third party transfer. It can be observed that enquiries for balances and account statements clearly dominate. When more than one transaction occurs in a call, the first request is typically for a balance. The interviewees expressed a need to include further transaction types in the future, such as third party payments for the U.K. banks and establishing standing orders for the Danish banks.

How?

The How? question relates to the profile of the call, how customers are greeted, how they address the system, etc. Greeting and ending phrases. The banks' name must be included in the message at the beginning and the end of a call. The dialogue design must anticipate that many customers do not catch the first words after the call is established, and consequently no crucial information must be given here. The voice is considered important. It must be clear, and carry a perceived 'bank' personality. The mode of addressing should be friendly, yet formal.

In case of the customer not reacting, the system should reprompt after a suitable interval. In case of some communication problem, the system should retry two times. If the problem is not solved, the system on request should pass the customer to a human operator. This option presupposes that a human operator (typically at a CCS) can be reached, which might no always the case.

The system must be tolerant of customer interruptions, i.e. allow barge-in. It is unavoidable that speech recognition errors will occur at some level. In extreme cases, e.g. in very noisy environments, or a very strong customer accent, this may seriously damage the interaction. Therefore, the system should offer the possibility of touch tone input in parallel to spoken input. Also, in some situations the customer might be forced to speak his Id- and Access codes when in public. This situation can be avoided by allowing the telephone keypad to be used instead when supplying this information.

4. Trial Specifications

The requirements have led to the specification of a series of usability trials. It was decided to set up three trials with increasing complexity to measure the usability that can be expected for a voice controlled service.

- Trial One. Focuses on the customer identification and verification phases and a simple balance and account statement (c.f. section . Application Domain)
- Trial Two. Same focus as trial one, but with more flexible user input.
- Trial Three. Focuses on robustness and more compre-

hensive service functionality.

5. Dialogue Implementation

The trial dialogues are implemented and the experiments carried by the academic partners. At CPK a Generic Dialogue System platform is used for the implementation. It has successfully been used for similar applications [2],[3], and provides a development as well as a runtime environment.

6. Conclusions

A set of salient user requirements have been obtained using a methodology of semi-structured interviews, organised according to the Six Witches questions. The methodology has proven an effective way of obtaining the information needed to design a voice controlled dialogue system. On the basis of these results a series of three trials have been set up to investigate the usability of the proposed services, and to further investigate the usability of such systems. A goal is also to obtain detailed knowledge of user behaviour, when actually confronted with a voice controlled service. It is the banks' belief that voice recognition will especially prove attractive for older age groups (over 50), whereas the present IVR services tend to used by a majority of younger customers. Lack of confidence in the services was given as a major reason by customers stopping to use the system. This implies that the speech recognition must have a very high level of accuracy in order to ensure that customers will use the system.

Acknowledgements

The work performed at CPK is partly funded by the EU through the OVID project, and partly by the Danish Technical Research Council, through its funding to CPK

The author wishes to thank all the participants in the OVID project. Especially Professor Mervyn Jack, Director of CCIR, Edinburgh University, who conducted most of the interviews.

References

[1] ESPRIT 20171 Project OVID - Trial application of Voice Processing in Automated Telephone Banking Services: “Technical Annex”, July 1995.

[1] ESPRIT 20171 Project OVID - Trial application of Voice Processing in Automated Telephone Banking Services: “User Requirements (Deliverable D1)” CCIR Edinburgh March 1996.

[2] “L.B. Larsen A. Baekgaard, “Rapid Prototyping of a Dialogue System using a Generic Dialogue Development Platform” in Proc. ICSLP-94, Yokohama 1994.

[3] L.B. Larsen, “Development and evaluation of a spoken dialogue for a telephone based transaction system”, in proc. EUROSPEECH-95, Madrid 1995.

A Strategy for Mixed-Initiative Dialogue Control

Lars Bo Larsen
Center for PersonKommunikation
Institute of Electronic Systems
Aalborg University
DK-9220 Aalborg Denmark
Email: ***lbl@cpk.auc.dk***

Abstract

This paper presents and discusses a strategy for mixed-initiative dialogue management within a home banking application. The strategy tries to utilise the guidance of system-directed dialogues, while accommodating user initiated focus shifts by the inclusion of *short-cuts* in the dialogue. The paper reports on two experiments, one with a simulated speech recogniser (WOZ), and the second with a fully automated system. Both experiments shows that users use the possibility for *short-cuts*, even when not instructed of their existence. A tendency towards user habituation is also demonstrated.

1. Introduction

This paper describes a strategy for mixed-initiative spoken dialogue management. The strategy is outlined, and two experiments are carried out to investigate the methodology. The experiments are carried out within the Esprit OVID[1]project. The OVID project is concerned with the development of trial applications within automated banking services [1]. Among other things, the OVID user specifications [1] states that the customer must be in control of the interaction. However, this may not always lead to the most natural or efficient mode of communication, as humans often expect others to hold or take the initiative in conversations. Therefore, a mixed-initiative strategy is proposed.

The overall goal of the OVID project is to measure user acceptance of voice controlled home banking systems. Apart from this, the purposes of the dialogue experiments reported here are twofold:

- To test the implemented dialogue management strategy.
- To identify and delimit the application vocabulary

The customers use unconstrained natural speech, and the speech recognition technology chosen for the task is a combination of digit string rec-

[1] The OVID Esprit 20717 Project consortium comprises The Royal Bank of Scotland and Barclays Bank in the U.K., Lån & Spar Bank in Denmark, CCIR Edinburgh University, U.K, CPK, Aalborg University, Denmark, Brite Voice Technology U.K., and AGORA Consult, France as coordinating partner. The work presented here is partly funded by the ESPRIT programme, and partly by CPK

ognition and spotting of keywords and -phrases. The experiments must therefore include identification of the application vocabulary. Consequently, the experiment is carried out in two phases. First with a simulated speech recogniser (Wizard of Oz.), denoted Trial 1 or “WOZ-trial” and the second with a fully automated system, denoted Trial 2. This paper focuses on the dialogue management issues, and reports on the experiments carried out in Trial 1 and preliminary results from Trial 2.

2. Dialogue Specifications

The overall functionality of the automated home banking application is:

- The service must first enquire the customer for his/her identification (Id) number, and subsequently a PIN code. The formats are identical to those used by Danish banks.
- The service provides the customer with a balance and an overview of the most recent transactions on his/her accounts (denoted a Mini Statement). Each customer has three accounts.
- DTMF interpretation of at least Id- and PIN codes must be available to ensure privacy.

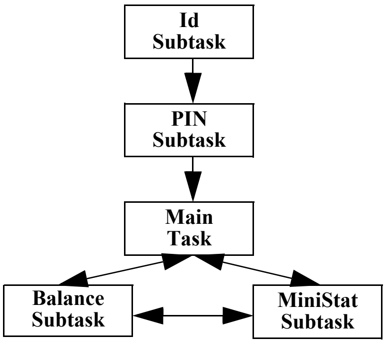


Figure 1Task Structure

On the basis of the overall specification a simple dialogue structure with five tasks is implemented. These are the Main task, Id- and PIN sub tasks and Balance and Mini-Stat subtasks.

The task structure is shown in Figure 1. Instead of building specific DTMF subtasks, all tasks accept spoken and DTMF key pad input in parallel. This is achieved by including two sets of prompts in the dialogue, and switching between them depending on which modality the user chooses.

3. Dialogue Initiative

The question of system directed vs. user-driven (or mixed-initiative) dialogue control strategies has been the focus of discussion for a number of years. In general, user controlled dialogues is considered preferable, as this allows the user to gain the control over the interaction, and hence achieve his goals more directly. In contrast, system-directed dialogues tend to be more rigid and menu-like.

However, this might not always be the case. A problem that might arise in user-driven dialogues, is that the user is left without a clear understanding of his options at a given point in the dialogue. This can cause frustrations or even breakdown of the communication. In [2] it is demonstrated that for a train information task, users actually preferred the system directed mode. On these grounds, and in the case of inexperienced users, the system directed mode might be preferable, while experienced users will choose to gain the initiative. Consequently, a combined system directed and user-driven dialogue (mixed-initiative) management strategy is employed in the present case. By default, the system has the initiative, and the user responds to system prompts. This works well for inexperienced users, who will be guided throughout the dialogue. However, for experienced (or impatient) users this strategy is too rigid. There clearly exists a need for the user to be able to take the initiative and directly request the desired information from the service. This is achieved by including a number of *short-cuts* in the rigid system directed dialogue structure.

By performing a short-cut, the user overrules the dialogue task structure, and forces the system to switch from one subtask to another. The short-cuts are shown as dashed (red) arcs in the simplified diagram of the overall dialogue flow structure shown in Figure 2.

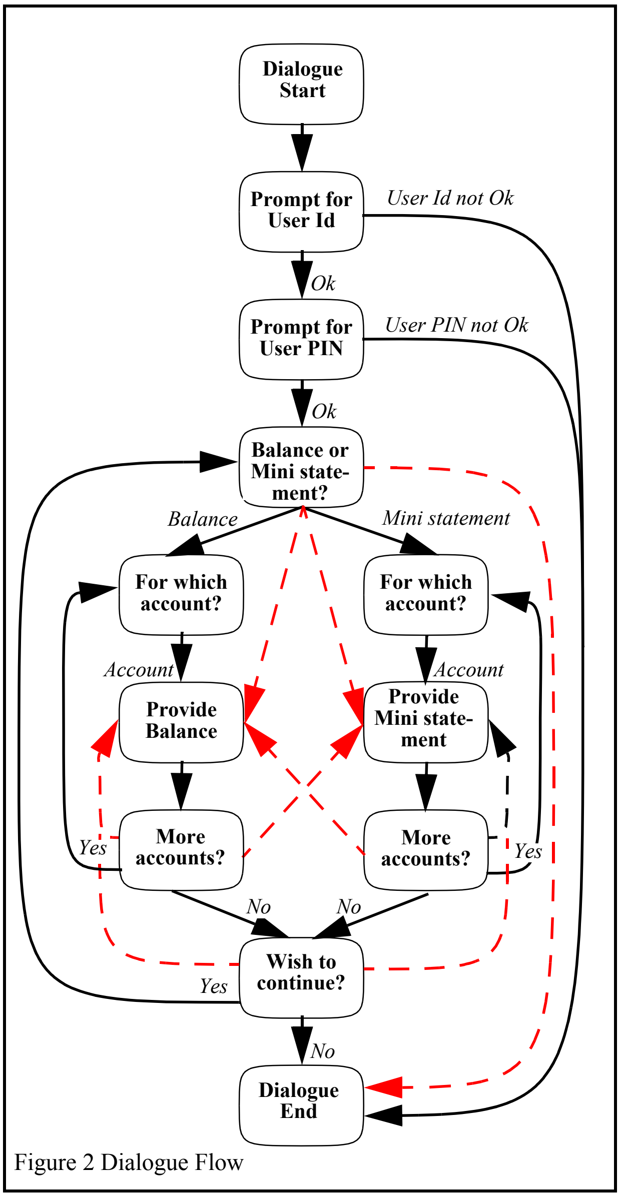


Figure 2 Dialogue Flow

The text in the boxes denotes system utterances, and the semantics of the user reasons are shown on the connecting arcs.

Two types of arcs are shown. The fully drawn lines show the system initiated transitions, and the dashed arcs depicts the user initiated transitions (*short-cuts*). This means that if the user answers all system prompts faithfully, the possible dialogue state transitions will reduce to the fully drawn lines. Incidentally, this corresponds to the dialogue that would be valid if only DTMF input was available.

Note that the user can generate a transition to “Dialogue End” from any point in the dialogue simply by hanging up.

4. Dialogue Model

Further to the task structure and flow control, the dialogue model comprises a number of elements. Among the most important are:

- User Profile.
- Dialogue History.

The system builds and maintains a profile of the user’s behaviour. This includes the information already given to the user, and whether the user has been given specific instructions about the use of the system. The user profile and dialogue history is used to determine the way the system will respond to specific user input. E.g. in the case of a rejected user utterance, the system response will be dependent on previous instructions given to the user.

5. Experiments

In the experiments, all subjects received a letter describing the application, and defining two scenarios. Furthermore, they received a usability questionnaire to be filled out and returned.

Deliberately, the subjects were not informed of the short-cuts in the dialogue. By this, it is possible to investigate to what extent users will naturally take the initiative, and also how quickly users can be termed “experienced”.

In both trials each user completed two scenarios; A and B.

- Scenario A: Obtain the balance for all three accounts
- Scenario B: Obtain the balance and a mini statement for the budget account.

6. Results

6.1. Results of Trial 1 (WOZ-Trial)

Trial 1 was carried out with a limited number (20) of participants. All the participants had either some connection with the university or with Tele Denmark. They had no prior knowledge of the application, although a number of them had experience with speech technology. They were not told that they participated in a simulated trial. The results of the experiments in Trial 1 is shown below in Table 1 in terms of number of turns and dialogue completion times. The figures are based on 40 dialogues. The nominal number of turns is the number of turns a user would have to go through if he/she answered all system prompts without gaining the initiative at any point.

If the user gains the initiative e.g by supplying additional information, or by answering a yes/no question with a new request, he can short-cut the rigid system controlled dialogue structure. Full utilisation of this yields the minimal number of turns.

The scenarios are designed in such a way that the nominal number of turns are almost equal for both, but scenario B can be completed with less that half the nominal number. Inspecting the average number of turns does not directly give an indication of how users perform, but a closer investigation of the transcribed dialogues shows that the subjects now are separated in two groups. This can also be observed in the confidence intervals, which have doubled for scenario B as compared to scenario A.

Scenario:	**A**	**B**
Nominal number of turns	7	9
Minimal[a] number of turns	5	4
Average number of turns	8.1[b]	7.8
95% confidence interval (turns)	0.5	1.0
Average duration of dialogues[c] (seconds)	105	112
95% confidence interval (duration)	6.1	13.9

Table 1 Turns and dialogue completion times

[a] This includes the user hanging up immediately after the desired information has been obtained.

[b] Some users requested the information twice, or asked for repetition. Therefore the average number of turns is larger than the nominal.

[c] It should be taken into account that the Mini Statement includes full description of three postings, and hence the average completion time for scenario B is influenced by this.

The other group have started utilising the short-cuts. This tendency is even more pronounced when taking the user identity and verification procedure into account. This “costs” two turns in all cases. This tendency is illustrated in Figure 3

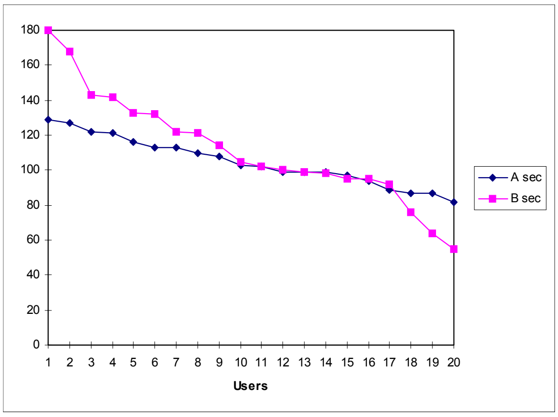


Figure 3Duration of scenario A and B in Trial 1

Another objective of Trial 1 was to identify the vocabulary for the word spotting speech recogniser to be used in Trial 2. A total of 20 words were found to be sufficient for the task.

6.2. Results of Trial 2

Trial 2 was carried out with 350 customers from the Danish Lån & Spar Bank. As Lån & Spar is a "Direct Bank" depending heavily on automatic services all users were experienced users of DTMF systems, but had not used a speech controlled system before. The users were selected evenly from geographic regions and age groups.

The results reported for Trial 2 are preliminary and are based on data from 80 users and a total of 176 dialogues. The users were given the same scenarios as in Trial 1, but 50% of the users were instructed to perform scenario B first. As in Trial 1, the CPK Generic Dialogue System (GDS) platform [3],[4],[5] was used to implement the dialogue. Trial 2 was carried out using the CPK SUNCAR real-time speech recogniser [6]. The Danish SpeechDat M 1000 speaker corpus [7] was used for training of the acoustic models. The dialogue model was identical to that of Trial 1.

It was not possible to identify a similar tendency for Trial 2 as shown in Table 1 and Figure 3 concerning the overall duration of the dialogues. However, the users did take the initiative throughout the dialogues, as indicated in Figure 4

The numbers shows whether it is the users' first or second dialogue.

The figure shows clearly that more experienced users tend to take the initiative more often (e.g. compare A1 to A2).

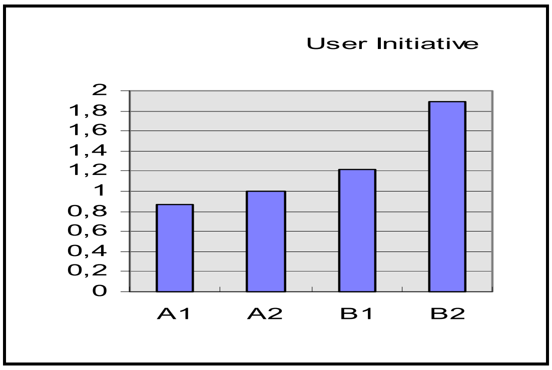


Figure 4User initiatives per dialogue

Turns	Total		Id.Num		PIN		Main		Balance		Mini St.	
Scenario	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**
Nominal	7	9	1	1	1	1	2	3	3	2	0	2
Actual	8.0	7.2	1.3	1.3	1.3	1.2	2.0	2.0	3.2	1.7	0.2	0.9
Minimal	5	4	1	1	1	1	1	1	3	1	0	1

Table 2 Nominal, Actual and Minimal number of turns per task

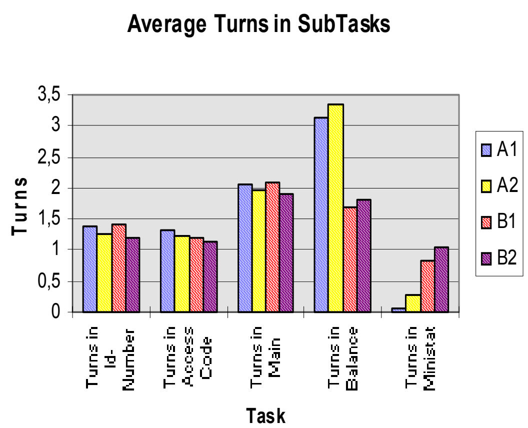


Figure 5Turns per Task

In Table 2 and Figure 5 it is shown how many turns the users spend on average in each subtask. Note that the average number of turns in the Id- and PIN code task are very close to one. The nominal number of turns in the Balance task is three for scenario A and two for scenario B. Again, the actual figures come very close, and is even a little below the nominal number. The nominal number of turns in Ministat is zero for scenario A and one for B. The average number of turns in scenario B actually drops below one, which indicates that not all users succeed in getting the mini statement required in the scenario.

7. Conclusions

The emphasis has been put on naturalness and flexibility of the spoken input/output and the dialogue structure. The user tests indicate that this goal has been accomplished, as the users were able to start gaining the initiative and short-cut the system controlled dialogue structure without prior instructions or informations about this opportunity.

Trial 1 indicates that after only one exposure to the dialogue, some of the users have become acquainted with the dialogue structure. This indication could also be found in Trial 2. It furthermore showed that users immediately started to go beyond the limits of the system directed dialogue structure and utilising the built-in short-cuts. An even more pronounced effect can be expected when users are exposed to a larger number of dialogues.

Speech recognition error rates for digit strings and word spotting were found to be approximately 10%. This seemed sufficient to ensure user acceptance. The preceding sections have shown that the strategy of maintaining a system directed dialogue on the surface and then provide *short-cuts* for more experienced users has proved successful. Trial 2 was, in fact, a usability trial, with the aim to investigate to what extent customers are prepared to accept voice controlled access to their bank accounts. The preliminary results strongly indicates that this is the case.The tested dialogue was very small, containing only a few sub tasks, so the next step will evidently be to expand the dialogue to cover a larger number of tasks, and a more complex task structure.

8. Acknowledgements

The author wishes to thank Børge Lindberg, Bo Bai and Jesper Olesen from CPK for their help in getting this experiment done.

9. References

[1] Lars Bo Larsen, "Voice controlled home banking - objectives and experiences of the Esprit OVID project" in Proc IVTTA-96, New Jersey September 1996 (ieee 96TH8178).

[1] ESPRIT 20171 Project OVID - Trial application of Voice Processing in Automated Telephone Banking Services: "User Requirements (Deliverable D1)" CCIR Edinburgh March 1996.

[2] Louis Boves et al. "Localization and Field Test of a Dutch Train Time Table Information System" oral presentation and in Proc. IVTTA-96, New Jersey September 1996 (ieee 96TH8178).

[3] Anders Bækgaard, “The GDS PLatform” Report 10 from the Danish Dialogue Project, CPK, Aalborg University 1996.“

[4] L.B. Larsen A. Baekgaard, “Rapid Prototyping of a Dialogue System using a Generic Dialogue Development Platform” in Proc. ICSLP-94, Yokohama 1994.

[5] L.B. Larsen, “Development and evaluation of a spoken dialogue for a telephone based transaction system”, in proc. EUROSPEECH-95, Madrid 1995.

[6] Børge Lindberg, Jan Kristiansen, “Realtime Speech Recognition within Spoken Dialogue Systems”, Report 8 from the Danish Dialogue Project, CPK, Aalborg University 1996.

[7] H. Christensen, B. Lindberg and P. Steingrimson, “Documentation of the Danish SpeechDat (M) Database”, CPK, Aalborg University, Aalborg 1996.

Investigating a Mixed-Initiative Dialogue Management Strategy

Lars Bo Larsen

Center for PersonKommunikation - CPK

Aalborg University, Denmark

Email: lbl@cpk.auc.dk

http://www.kom.auc.dk/~lbl/

ABSTRACT. This paper discusses how to design mixed-initiative spoken dialogues with only a partial recognition of the user utterances (recognition of concepts or phrase spotting). The objective is to investigate the potential of such a technique and in particular to develop a corresponding dialogue model. The work has been carried out within the ESPRIT OVID project[1], which addressed a voice controlled home banking task. A number of experiments have been carried out within the project and the present paper discusses the results of these.

A mixed-initiative dialogue management model has been developed and implemented, and the experiments have shown that users to a very high degree are able to grab the initiative at natural points in the dialogue.

1. Introduction

Within the most recent years very high performing spoken dialogue systems have emerged in a number of laboratories, e.g at MIT [1]. In contrast to the previous generations of spoken dialogue systems, these systems exhibit a more natural interaction style and high performance speech recognition. Common to them is that they rely on sophisticated speech recognition techniques as well as a combination of a powerful natural language grammar and a statistical language model. Previously, systems had a tendency to rely on a a more unnatural, machine directed dialogue mode in order to cope with the complexities at hand.

These systems show great promise, but they all represent a large invest-

[1] The OVID Esprit 20717 Project consortium comprises The Royal Bank of Scotland and Barclays Bank in the U.K., Lån & Spar Bank in Denmark, CCIR Edinburgh University, U.K, CPK, Aalborg University, Denmark, Brite Voice Technology U.K., and AGORA Consult, France as coordinating partner. A number of reports and papers from the project is located at the OVID homepage:
http://www.kom.auc.dk/CPK/Speech/OVID/
The work presented here is partly funded by the ESPRIT programme, and partly by CPK. The OVID project ended August 1 1997.

ment in terms of manpower and development time, which can only in part be transferred to new application domains.

1.1. Background

The experiments reported here are carried out within the Danish part of the ESPRIT OVID project and addresses the domain of phone based home banking. This task is well defined and as such represents a very broad class of well-structured tasks, which are suitable for voice controlled automation. Examples of these are: credit card information services, telephone ordering services (of e.g. travel catalogues), book clubs, number to name phone services and transaction systems in general. Common to these services are that a certain degree of structure can be imposed upon the discourse model without compromising the naturalness of the dialogue. As a consequence, the linguistic phenomena exhibited by the user tend not become too complicated.

Many of such services have been automated within the last 5 to 10 years using Interactive Voice Response (IVR) technology. The drawback of these systems are that they tend to grow into large menu based systems, which are tedious to navigate, and prompts become very long listings of the users' options. Obviously, IVR services also requires the customer to use a DTMF phone.

The class of services described above is very suitable for voice control, especially as they can be structured in a way which is perceived as natural to the users, while not requiring an extensive amount very domain specific language engineering. Consequently they have a very high commercial potential.

1.2. The Present Task

The overall goal of the OVID project is to measure user acceptance of voice controlled home banking systems [2]. This is achieved by setting up trial applications and measure the users attitudes by means of a usability questionnaire.

Among other things, the OVID user specifications [3] states that the customer must be in control of the interaction. However, this may not always lead to the most natural or efficient mode of communication, as humans often expect others to hold or take the initiative in conversations. Therefore, a mixed-initiative strategy is proposed. Consequently, the purposes of the dialogue experiments reported here are twofold:

- To develop and test a dialogue management strategy in accordance with the specifications.
- To measure user attitudes towards the service.

The focus of this paper is on the design of the dialogue management strategy, but user responses are included in the evaluation of the findings.

2. Application

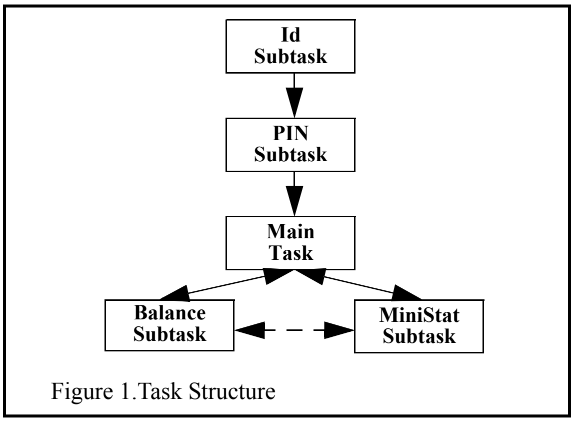


Figure 1.Task Structure

The overall functionality of the application is as follows: The system prompts for the user's Id and PIN numbers. After this, the main dialogue task is entered, and the user can request account information as either a balance or the most recent movements on his/her accounts. In the experiment all users have three accounts. The application is bi-modal as the user can switch to DTMF input at any point in the dialogue. Either on his own initiative (e.g. when submitting Id and PIN numbers), or on advise from the system (e.g. after repeated misrecognitions). The overall task structure is shown in Figure 1

The arrows connecting the tasks indicates the possible transitions. The dashed arrow between the Balance and MiniStat sub tasks indicates that this transition can only be made when the user grabs the initiative in the dialogue. The tasks are identical for both DTMF and spoken input, except that the system's prompts change according to the input mode, and that user initiative is only possible with spoken commands.

2.1. Dialogue Management strategy

The question of system directed vs. user-driven (or mixed-initiative) dialogue control strategies has been the focus of discussion for a number of years. In general, user controlled dialogues are considered preferable, as this allows the user to gain the control over the interaction, and hence achieve his goals more directly. In contrast, system-directed dialogues tend to be more rigid and menu-like.

However, this might not always be the case. A problem arising in user-driven dialogues, is that the user is left without a clear understanding of his options at a given point in the dialogue. This can cause frustrations or even breakdown of the communication. In [4] it is demonstrated that for a train information task users actually preferred the system directed mode. On these grounds, and in the case of inexperienced users, the system directed mode might be preferable, while experienced users will choose to gain the initiative. Consequently, a combined system directed and user-driven dialogue (mixed-initiative) management strategy is employed in the present case. By default, the system has the initiative, and the user responds to system prompts. This works well for inexperienced users, who will be guided throughout the dialogue. However, for experienced (or impatient) users this strategy is too rigid. There clearly exists a need for the user to be able to take the initiative and directly request the desired information from the service. This is achieved by including a number of *short-cuts* in the rigid system directed dialogue structure.

By performing a short-cut, the user overrules the dialogue task structure, and forces the system to switch from one sub task to another.

2.2. Speech Recognition

The users must be able to use unconstrained natural speech. As discussed in Section 1, this often calls for an elaborate language model, both in terms of the acoustic recognition task (typically a bi- or trigram) and the following natural language parsing. However, when addressing well structured tasks as the present one, this might be avoided, and a much simpler model can be employed.

Hence, a shorter development time can be expected, and the extent of linguistic expertise needed may be reduced accordingly. In the present case, the speech recogniser [5] is used to do a combination of word- and phrase spotting. The acoustic decoding is based on a mixture of phoneme and whole word (digits and a limited set of function words) models.

Robustness is achieved through the use of adaptive garbage modelling and dynamic estimation of the background noise levels.The acoustic models were trained on the Danish part of the SpeechDat(M) database [6].

2.3. Experiments

Two experiments were carried out. Trial 1 was a Wizard of Oz experiment, with a limited number of users. The purpose was to verify the dialogue model and establish the application vocabulary. Trial 2 was carried out with 350 customers from the Danish Lån & Spar bank and the fully automated system was used. Users could call from their homes at any time during the period of the trial. Calls from cellular phones were not possible.

Deliberately, the subjects were not informed of the short-cuts in the dialogue, i.e. they were not aware that they could take over the initiative in the dialogue. This makes it possible to investigate where in the dialogue and to what extent users will naturally take the initiative, and also to what extent user habituation occurs.

In both trials each user completed two scenarios; A and B.

- Scenario A: Obtain the balance for three named accounts
- Scenario B: Obtain the balance and a mini statement for the "budget" account.

3. Results

Trial 1 verified that the dialogue model was acceptable to users, and served as a preparation for Trial 2. The keyword vocabulary was established and was used in Trial 2.The first part of this section presents issues concerning of the dialogue model in terms of turns and user initiative. The second part combines the user responses with the speech recognition accuracy they encountered.

3.1. Verification of the Dialogue Model

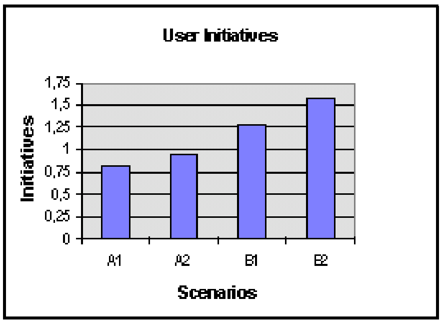


Figure 2User initiatives per dialogue

The results reported here for Trial 2 are based on the results from 350 users and 800 transcribed dialogues. Half of the users were instructed to perform scenario B first. As in Trial 1, the CPK Generic Dialogue System (GDS) platform ([7], [8], [9]) was used to implement the dialogue.

Turns	**Total**		**Id.Num**		**PIN**		**Main**		**Balance**		**Mini St.**	
Scenario	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**
Nominal	8	9	1	1	1	1	2	3	4	2	0	2
Actual	8.0	7.2	1.3	1.3	1.3	1.2	2.0	2.2	3.2	1.7	0.2	0.9
Minimal	5	4	1	1	1	1	1	1	3	1	0	1

Table 1 Nominal, Actual and Minimal number of turns per task

The enumerations on Figure 2 denotes the scenario and whether it is the users' first or second dialogue.

The figure shows clearly that more experienced users tend to take the initiative more often (e.g. compare A1 to A2).

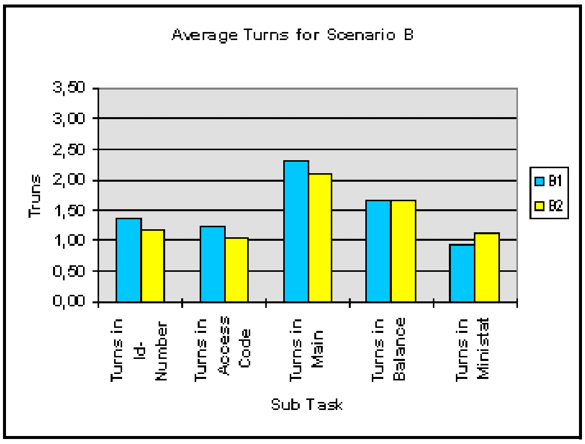


Figure 3Turns per Task

In Table 1 and Figure 3 it is shown how many turns the users spend on average in each sub task (only scenario "B" is shown in figure 3). Note that the average number of turns in the Id- and PIN code task are very close to one. The nominal number of turns in the Balance task is three for scenario A and two for scenario B. Again, the actual figures come very close, and is even a little below the nominal number. The nominal number of turns in Ministat is zero for scenario A and one for B. The average number of turns in scenario B actually drops below one, which indicates that not all users succeed in getting the mini statement required in the scenario.

3.2. User Attitudes

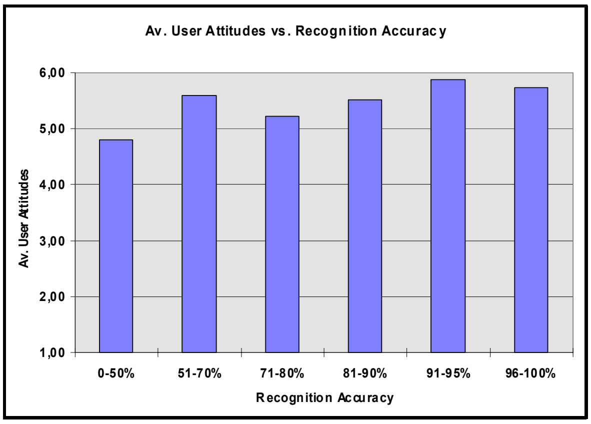


Figure 4. User Attitudes

All users were required express their attitudes towards the system by filing out a usability questionnaire developed at CCIR at Edinburgh university [10]. It consists of a set of Likert statements. The user expresses his/her attitude by ticking boxes ranging from "strongly agree" to "strongly disagree" to each statement. The statements are categorized into usability aspects such as: quality of interface/performance, cognitive effort/stress, conversational model, fluency and transparence. The responses are transformed into a scale ranging from one to seven, with four as a neutral attitude. Figure 4 shows the average of the users attitude as a function of the experienced speech recognition accuracy. In the figure, users have been grouped according to the degree of recognition accuracy they experienced when using the system. As can be observed from the figure, the accuracy has very little influence on the attitude towards the system.

It may seem surprising, but the recognition accuracy is one of many factors influencing users attitudes towards an application. The categories "0-50%" and "50-70%" only comprises approximately 10% of the users (see figure 5). The overall speech recognition accuracy is 84%.

This also includes errors made by the speech detector. On average each user say nine keywords[1] per dialogue, which in turn means that they may expect about one misrecognition per dialogue.

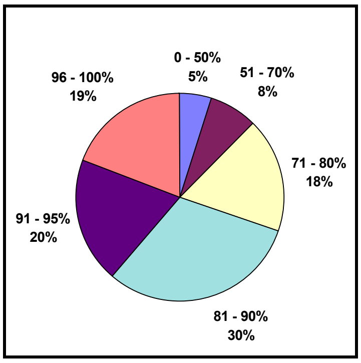


Figure 5. Users Sub grouped

4. Discussion

The emphasis has been put on naturalness and flexibility of the spoken input/output and the dialogue structure. The user tests indicate that this goal has been accomplished, as the users were able to start gaining the initiative and short-cut the system controlled dialogue structure without prior instructions or informations about this opportunity. Figure 2 shows that users immediately start to go beyond the limits of the system directed dialogue structure and utilizing the built-in short-cuts. An even more pronounced effect of user habituation can be expected when users are exposed to a larger number of dialogues.Speech recognition error rates for digit strings and word spotting were found to be 16% on average. This seemed sufficient to ensure user acceptance, but further stresses the importance of the dialogue design to ensure high user acceptance. The preceding sections have shown that the strategy of maintaining a system directed dialogue on the surface and then provide *short-cuts* for more experienced users has proved successful.

The aim of Trial 2 was to investigate to what extent customers are prepared to accept voice controlled access to their bank accounts. The prelim-

[1] The Id- and PIN codes are here counted as one "keyword", although they actually consist of 7 and 4 digits. It turns out that digit string and keyword accuracy are roughly similar.

inary results strongly indicates that this is the case. The tested dialogue was very small, containing only a few sub tasks, so the next step will evidently be to expand the dialogue to cover a larger number of tasks, and a more complex task structure.

Another very interesting experiment will be to subject users to a number of different dialogue models, in order to verify to what extent this influences their attitudes towards the system.

5. Acknowledgements

The author wishes to thank the OVID team and in particular Børge Lindberg, Aalborg University and Lars Rud, Lån & Spar Bank for their help and encouragement

6. References

[1] Victor Zue, “Conversational Interfaces: Advances and Challenges”, in proc. EUROSPEECH-97, Rhodes 1997.

[2] Lars Bo Larsen, “Voice controlled home banking - objectives and experiences of the Esprit OVID project” in Proc. IVTTA-96, New Jersey September 1996 (ieee 96TH8178).

[3] ESPRIT 20171 Project OVID - Trial application of Voice Processing in Automated Telephone Banking Services: “User Requirements (Deliverable D1)” CCIR Edinburgh March 1996.

[4] Louis Boves et al. “Localization and Field Test of a Dutch Train Time Table Information System” oral presentation and in Proc. IVTTA-96, New Jersey September 1996 (ieee 96TH8178).

[5] Børge Lindberg, Jan Kristiansen, “Real-time Speech Recognition within Spoken Dialogue Systems”, Report 8 from the Danish Dialogue Project, CPK, Aalborg University 1996.

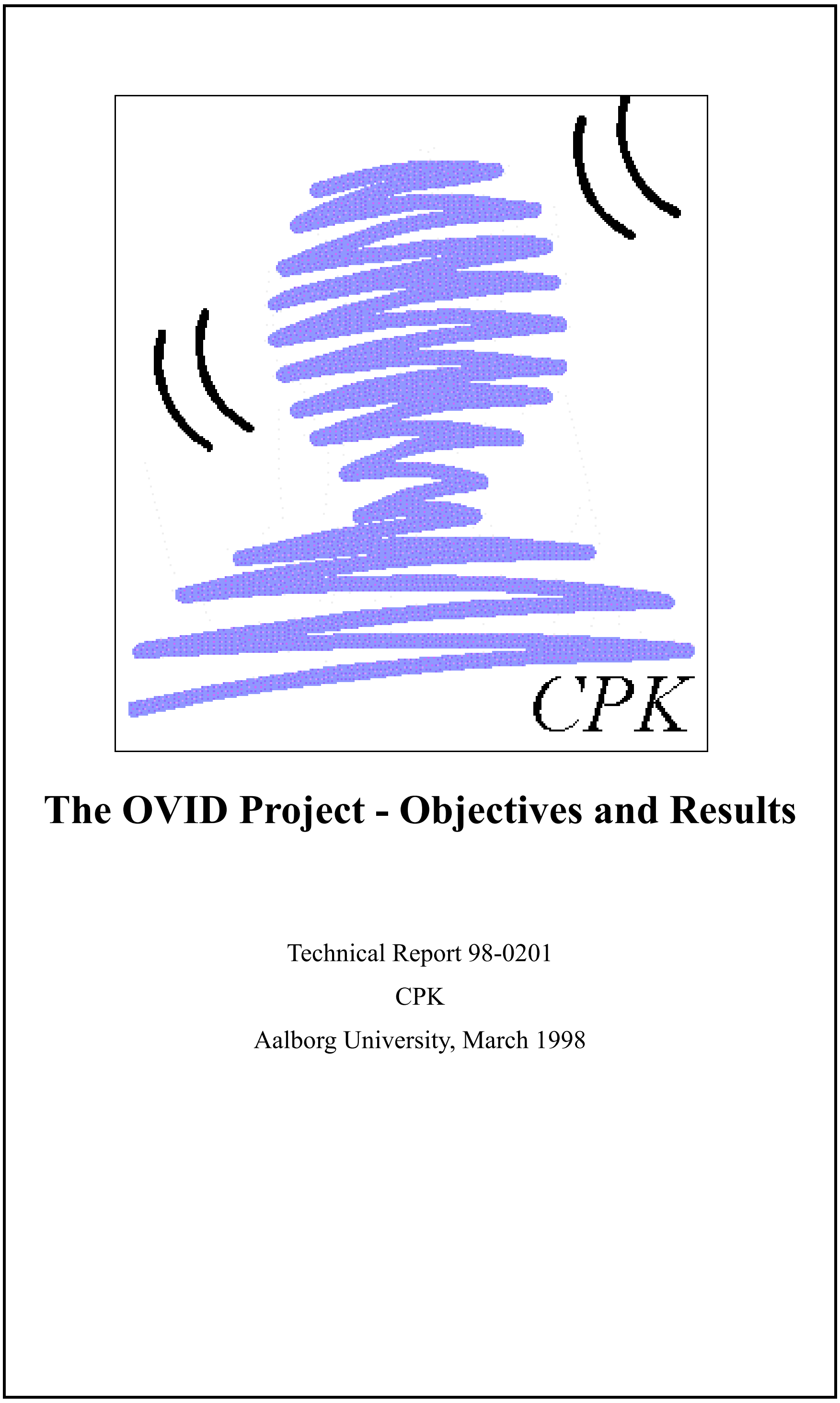
[6] H. Christensen, B. Lindberg and P. Steingrimson, “Documentation of the Danish SpeechDat(M) Database”, CPK, Aalborg University, Aalborg 1996.

[7] Anders Bækgaard, “The GDS PLatform” Report 10 from the Danish Dialogue Project, CPK, Aalborg University 1996.“

[8] L.B. Larsen, “Development and evaluation of a spoken dialogue for a telephone based transaction system”, in proc. EUROSPEECH-95, Madrid 1995.

[9] L.B. Larsen, “A Strategy for Mixed-Initiative Dialogue Control”, in proc. EUROSPEECH-97, Rhodes 1997 (to appear).

[10]D. Poulsen “Towards simple indices of the perceived quality of software interfaces” In IEE Colloquium - *Evaluation Techniques for Interactive System Design,* IEE, Savoy Place, London



The OVID Project - Objectives and Results

Technical Report 98-0201

CPK

Aalborg University, March 1998

Keywords

Spoken Dialogue Systems, Speech Recognition, Dialogue Management, Usability Engineering, Tele Banking, Home Banking.

Summary

The OVID project deals with the task of voice controlled, phone based home banking services.

This report documents the results mainly achieved at CPK within the ESPRIT OVID project. It describes the objectives behind the project, the methodologies applied and the results of the performed trials.

Project Consortium

The OVID Esprit 20717 Project consortium comprises:

AGORA Consult, France (coordinating partner)
The Royal Bank of Scotland and Barclays Bank in the U.K.
Lån & Spar Bank in Denmark
Centre for Communication Interface Research, Edinburgh University, U.K.
Center for PersonKommunikation, Aalborg University, Denmark
Brite Voice Technology, U.K.,
Voice Corporation Systems, U.K.

Distribution

This report may be freely distributed. Any part may be quoted provided references is given to this report. The report is available from CPK in printed form (see below), or can be downloaded directly from:

http://www.cpk.auc.dk/~lbl/OVID/public/CPK_tec_rep.ps

Contact Points (Danish Partners)

CPK:
Lars Bo Larsen
Center for Personkommunikation
Lars Bo Larsen
Aalborg University
Fr. Bajers Vej 7A, DK-9220 Aalborg Denmark.
Email: lbl@cpk.auc.dk phone: +45 9635 8635 Fax: +45 9815 1583

Lån & Spar Bank:
Lars Rud,
IT Department
Lån & Spar Bank,
Højbro Plads 9-11, P.O. Box 2117 DK-1014 København K. Denmark
Email: Lars.Rud@laanspar.dkphone +45 6536 1110 / +45 3378 2000

The OVID homepage:
http://www.kom.auc.dk/~lbl/OVID/

Document Information:
Status: Version 1.0
Total Pages: 186
File: G:\phd\Reports\phd report\Included Articles\ovid 1_tech_report Reformatted.fm
Created: 15 Feb. 1998
Last modified: 4 July 2003 6:42 AM
Printed Copies:15

1 Foreword

This report summarises the results of the Danish part of the ESPRIT 20717 OVID Project. The project ran from ultimo 1995 until media 1997. The work presented here was carried out at CPK in collaboration with Lån and Spar Bank and the British partners, most notably CCIR at Edinburgh university.

The work presented here has been supported by the Commission of the European Union, DG. XIII and the Danish Technical Research Council (STVF) through CPK.

The report is based on a number of deliverables and papers produced by the OVID project consortium.

OVID Deliverables

Deliverable D1 “User Requirements” March 1996

Deliverable D2.1 “Trial 1: Usability Experiment Design” May 1997

Deliverable D2.2: “Trial 2: Usability Experiment Design” May 1997

Deliverable 3: “Trial Application Software” July 1977

Deliverable 4: “Spoken Dialogue Software” July 1997

Deliverable D5.1: “Trial 1: Results” May 1997

Deliverable D5.2: “Trial 2: Results” July 1997

Deliverable D6: “Project Dissemination and User Group Achievements” July 1997

“Final Report” Sep. 1997

Scientific Papers

L.B. Larsen “Voice Controlled Home Banking - Objectives and Experiences of the ESPRIT OVID project”, in proc. 3hrd IEEE Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA), New Jersey, USA Sep. 1996

L.B. Larsen: “A Strategy for Mixed-initiative Dialogue Control”, in proc. Eurospeech '97, Rhodes Greece Sep. 1997.

L.B. Larsen: “The Danish OVID Trials”, in proc. of “Workshop on Language Engineering and Telebanking”, the LINGLINK project, Anite Systems, Brussels Belgium, May 1997. (invited talk)

L.B. Larsen: “Investigating a Mixed-Initiative Dialogue Management Strategy”, in proc. of the ieee workshop on Automatic Speech Recognition and Understanding (ASRU), Santa Barbara USA, Dec. 1997.

Note! Apart from the scientific publications only D1, D6 and the Final Report are publicly available outside the project consortium.

2 Contents

3 Introduction

This report contains four major parts. First a study is conducted in order to capture the user's (banks) expectations and requirements to the OVID project. The findings of this study then form the basis for the design and implementation of the OVID prototype service. The proceeding section describes the trial dialogue service in some detail, and a simulation experiment (Trial 1) is conducted in order to verify the dialogue design and capture the application vocabulary for a fully automated system. The experiment with the fully automated system (Trial 2) is carried out as a field test with approximately 350 banking customers. This is documented in the following section. Finally the conclusions are drawn up.

3.1. The background and motivation of the OVID project

Two factors played an important role for the formulation of the OVID project. A business objective and a technical objective.

The technical objective was to evaluate the current state-of-the-art within the rapidly growing field of commercial voice technology applications. That is, to asses the usability of the technology available at the present or in the very near future in a domain which is expected to offer a potentially large number of applications for voice technology.

The business objective for the banks is to offer new and more efficient services to their customers.

Touch tone telephone banking systems have been in use for more than a decade, but have changed little over that period in functionality and user interface. In the same period, call centres have emerged.

Therefore the commercial incentive for the banks to switch to voice technology is very high. With a cost reduction of 90% for transactions via call centres compared to ordinary branches, and a further reduction of 90% for fully automated interactive touch tone voice response (IVR) transactions, there is a total cost reduction of 90-99% for each transaction that the bank can relocate from branches to an automated service. Furthermore, indications are that more than 80% of all transactions are suitable for automation ESPRIT 20171 Project OVID - Trial application of Voice Processing in Automated Telephone Banking Services: "Technical Annex", July 1995..

Thus the banks have a very high motivation in making automated services as attractive as possible. One way to achieve this is to replace current IVR and call center (CCS) operated services with voice controlled technologies.

4 The Requirements Capture

The findings of the requirements capture process are reported in detail in this section. Most notably a number of requirement interviews uncovered that almost identical requirements exist for all the OVID banks. This was the case even though the banks have very different profiles and market strategies. This observation suggests that the conclusions may be extended to cover voice controlled home banking services in general.

4.1. The Participating Banks

The project consortium banks comprises two British (BARCLAYS and the Royal Bank of Scotland) and one Danish bank (Lån og Spar Bank).

The banks differ in a number of aspects. The British banks are well established with large number of branches, whereas the Danish bank is small measured in branches and employees, and focuses directly towards a business strategy centred on telephone based services. Either as IVR, PC-based or via a branch only accessible via the telephone.

The British banks have established Call Centres (CC), which handle a growing proportion of customer transactions, and recently also Interactive Voice Response (IVR) services. Call centres do not exist in Denmark at all.

Because of this, the outlook of the bank partners differs. For the Danish bank, introducing voice technology would mean a potential increase of the functionality and attractiveness of an already automated service, whereas the British banks are looking for ways to automate or supplement current CC services.

4.2. Application Domain

The application domain of the OVID project is within voice controlled telephone based home banking. Basically, a customer call to a telephone banking system - either IVR or CCS involves four phases.

- *Customer identification.* The customer typically identifies him- or herself to the system by supplying a 7-12 digits number.
- Customer Verification. A password consisting of 2-5 digit from the customers’ PIN is required for caller identity verification.
- Balance and account status. Most often, the customer enquires about a balance and recent activity on the account.
- Transactions. A more complicated dialogue where the customer requests among a possible set of transactions, e.g transfers money to other accounts, or sets up standing orders, etc.

Not all calls involve the last type of transaction. Often the customer only wants information on a balance or whether e.g. a certain payment has taken place.

4.3. Results of the User Requirement Capture

This section describes the findings of the user requirements capture in detail. In the present context, the "user" is the bank, not to be confused with the end-users (the bank customers) of the service.

4.3.1Methodology

Two sources of information have formed the basis for the formulation of the user requirements.

The first concerns data collected by the banks on existing IVR systems and CC on statistics of transactions types, duration of calls, customer profile (age, gender, calls per month, etc.). This source gave accurate statistical information on the usage of the current services that the voice controlled service is intended to emulate.

The other source of information was a series of semi-structured interviews with representative personnel from the three banks. The interviews allowed the interviewees to answer freely within a broad framework of the service domain.

By using this approach, the interviewees were not constrained to only answer very specific predefined questions, but were allowed to introduce new issues not anticipated by the interviewer, such that novel ideas and concepts were likely to be uncovered.

4.3.2Structuring of the Interviews

The questions included in the interviews were concerned with the current situation with telephone banking for each bank and what the requirements for the OVID trials should be.

To systemise the interviews, the "Six Witches" method, developed at CCIR ESPRIT 20171 Project OVID - Trial application of Voice Processing in Automated Telephone Banking Services: "User Requirements (Deliverable D1)" CCIR Edinburgh March 1996., was adopted. The technique is built upon the *Who? What? Where? When?* and *Why?* questions of journalism extended with a *How?*, and provides the framework for the semi-structured interviews.

The following part of this section presents the resulting user requirement specifications obtained by the approach. In total, 24 key user requirements were identified through 27 interviews and grouped according to the Six Witches.

Who? This question concerns the gender, accent, habituation and age profiles of the potential customers.

All banks report an almost equal distribution between male and female callers. Table 1 [2] shows the main accent types anticipated by the banks.

Accent	Lån & Spar Bank	RBOS	Barclays Bank
Native Danish	98%	N/A	N/A
Native English	N/A	60%	70%
Scottish English	N/A	35%	5%
Other	2%	5%	25%

Table 1 Main Accent types

Especially for the British banks it is observed that accents are an important factor, and must be taken into consideration when designing the speech recogniser. For the Danish bank, regional accents must be taken into account.

The interviewees assessed that customers might experience difficulties at first but that they can be assumed to be 'experienced' users after having used the service two to three times.

Very few customers use an informal style of address, and consequently the service should maintain a formal mode of addressing.

The distribution according to age of the customers is similar for all the banks with a strong dominance for younger customers (age groups 20 - 40 years).

Why? The Why? question explains why customers use or stop using the existing IVR and CCS services.

The interviewees were asked to rank the reasons why (in their opinion) customers use the services. Table 2 below shows the results.

Rank	Lån & Spar Bank	RBOS	Barclays Bank
Highest	Convenience	Convenience	Convenience
	24-hour Service[a]	24-hour Service	24-hour Service
	Speed	Speed	Speed
	Security	Operator helpful	Security
	Informative	Security	Confidentiality
	Confidentiality	Confidentiality	Informative
Lowest	Operator helpful	Informative	Operator helpful

Table 2 Rank order for service features

[a] The need for a 24 hour service cannot be documented as the present IVR service closes between 00 and 04 hours (c.f. Figure 2.).

A high degree of agreement across the banks is evident, and the three highest ranking features are all related to convenience and availability of the services.

The most common reasons given by customers to stop using the service are loss of confidence and lack of functionality.

When? The question *When?* discusses the times and durations of calls to the services.

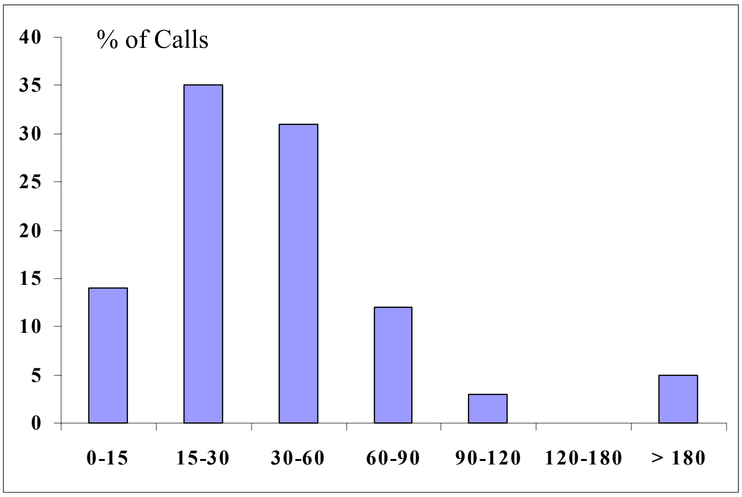


Figure 1 Duration of Calls (seconds)

Figure 1 shows the duration of the calls for the IVR service[1]. The average duration for IVR calls are as low as one minute. Customers typically call once or twice per month, with more calls towards the end of the month. A few customers use the service very often. Figure 2 shows the distribution of the average call density during the day[1]., and it is observed that a large proportion of the calls are made outside normal banking hours

[1] For the Lån & Spar IVR service

.

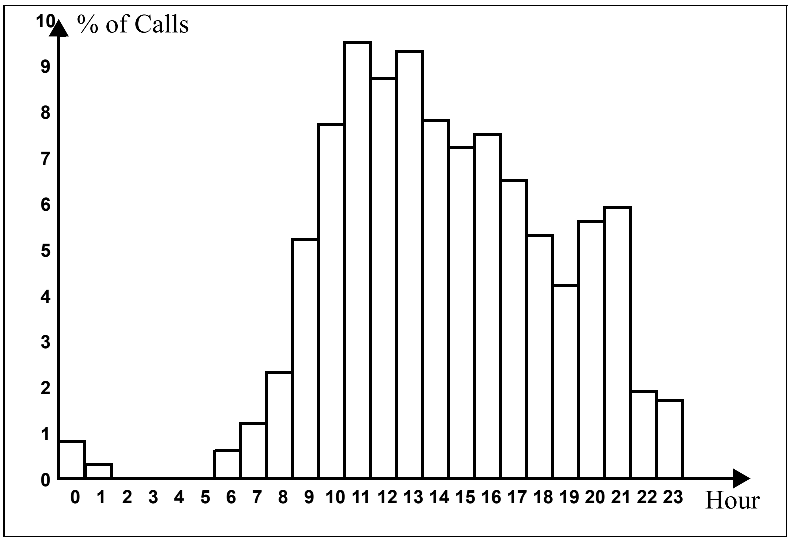


Figure 2. Average call densities during the day

The service must be able to handle a peak of no less than 10% of the expected daily calls at any time.

What? The question of *what?* discusses the types of transactions handled by the telephone banking services.

All the three banks require the customer to supply identification and security information. Typically the identification requires 7 to 12 digits and a security PIN. This is typical a mixture of 2 to 5 digits and alphanumericals.

[Table 3] [2] shows the average densities for the transaction types. For

Rank	Lån & Spar Bank	RBOS	Barclays Bank
Balance Enquiry	93%	54%	38%
Account Enquiry	42%	43%	22%
Bill Payments	N/A	28%	21%
Transfer Own acc.	28%	9%	8%
Transfer 3rd party	8%	N/A	N/A
Transfer Giro	10%	N/A	N/A
Order Statement	2%	1%	2%
Direct Debit	N/A	1%	2%
Exchange Rates	3%	N/A	N/A
Change Password	2%	N/A	N/A

Table 3 . Transaction Densities for OVID banks

Rank	Lån & Spar Bank	RBOS	Barclays Bank
Standing Orders	N/A	N/A	N/A
New Checkbook	N/A	1%	N/A

Table 3 . Transaction Densities for OVID banks

example, 93% of all calls to Lån & Spar bank involves a balance enquiry and 8% involves a third party transfer. It can be observed that enquiries for balances and account statements clearly dominate.

When more than one transaction occurs in a call, the first request is typically for a balance.

The interviewees expressed a need to include further transaction types in the future, such as third party payments for the U.K. banks and establishing standing orders for the Danish bank.

How? The *How?* question relates to the profile of the call, how customers are greeted, how they address the system, etc.

Greeting and ending phrases. The banks' name must be included in the message at the beginning and the end of a call. The dialogue design must anticipate that many customers do not catch the first few words after the call is established, and consequently no crucial information must be given here.

The voice is considered important. It must be clear, and carry a perceived 'bank' personality. The mode of addressing should be friendly, yet formal.

In case of the customer not reacting, the system should reprompt after a suitable interval.

In case of some communication problem, the system should retry two times. If the problem is still not solved, the system as a last resort should offer to pass the customer to a human operator. However, this option presupposes that a human operator (typically at a CC or a branch) can be reached, which might no always the case for the Danish bank.

The system must be tolerant of customer interruptions, i.e. allow barge-in.

It is unavoidable that speech recognition errors will occur at some level. In extreme cases, e.g. in very noisy environments, or a very strong customer accent, this may seriously damage the interaction. Therefore, the system should offer the possibility of touch tone input in parallel to spoken input. Also, in some situations the customer might be forced to speak his Id- and Access codes when in public. This situation can be avoided by allowing the telephone keypad to be used instead when supplying this information. However, it must be anticipated that not all of the service functionality will be available for touch-tone input.

4.4. Description of the OVID home banking service

Based on the previous sections, an overall specification of the Home Banking prototype can be formulated. The service main characteristics are:

- The user must able to speak naturally to the system, i.e. no artificial speaking styles must be required, and no explicit vocabulary must be imposed on the user.
- The user must be in control of the communication. This means that the user should be free to request any information or give any command (within the capability of the system) that he may wish at any point in the dialogue. The system must provide guidance in the case of mistakes or to point out the options to the user.

The actual functionality of the service has been chosen to be:

- The customer identification and verification procedure must be compatible to the one used today by Danish banks.

The service will provide information of the balance and latest movements of three named user accounts.

4.5. Definition of experiments

Two experiments (Trial 1 and 2) are set up in order to measure user responses to the proposed service.

Trial 1. Focuses on a first evaluation of the dialogue design. The trial is carried out with a simulated recogniser. This experiment will establish whether the chosen speech recognition paradigm (word- or phrase spotting, see section [5.1.]) is applicable to the present application. The service will accept telephone keypad input supplementary to the voice input.

The keyword spotting methodology allows the user to speak in a fully fluent, natural mode, but on the expense that only a limited set of key words will be recognised.

The specific purposes of the trial is to investigate:

- Whether it is possible to select a limited set of key words in such a way that the semantics of the user's utterances can be extracted correctly.
- If the accuracy of the speech recognition device is sufficient to be accepted by the users.
- Whether the resulting dialogue is acceptable to the users.

Trial 2. Except for corrections of errors uncovered in Trial 1, the dialogue as such remains unchanged. Trial 2 is accomplished with the fully automated system and is a field test with bank customers calling from their homes. The focus is on system performance and usability measures.

4.6. Conclusions

A set of salient user requirements have been obtained using a methodology of semi-structured interviews, organised according to the Six Witches questions. The methodology has proven an effective way of obtaining the information needed to design voice controlled dialogue systems.

On the basis of these results the functionality of the OVID prototype service has been defined. Two trial experiments have been defined in order to establish the service prototype and examine user responses to the service.

5 Dialogue Functionality

This chapter discusses the functionality of the OVID home bank prototypes. The functionality remains unchanged, except for minor corrections, for both trials.

5.1. Outline of the trial application service

This section outlines the actual functionality of the proposed dialogue.

One of the main conclusions of the requirement analysis was that the customer must be in control of the dialogue situation. This requirement cannot be achieved in all aspects, as it will be impossible for the system react to any user command occurring at any point in the dialogue without a full recognition and interpretation of the user’s spoken input. Therefore, the service will operate by prompting the user in a way that will elicit responses within the scope of the predefined key word vocabulary. The design of the dialogue messages will try to guide the user to answer within this scope without directly demanding that specific words or phrases are used. Thus, although the dialogue will be system directed, the user will still experience the dialogue as free and natural.

For example, the system will at no point automatically transfer the call to a human operator, or hang up on the user.

5.1.1Greeting, user identification and verification

The service will welcome the user to the Ovid telephone bank service. A request for the user Id and Access numbers is then presented (separately). The system asks the user to speak the numbers as connected digits without pauses in between - not natural numbers. The user is given three attempts to supply the information. The prompts will become increasingly specific after each attempt, and the last attempt will directly ask the user to use the telephone keypad instead of voice input. If this also fails, the system will advice the user to call his/hers local branch within normal opening hours (there is no possibility to transfer the call to a call center in Denmark).

5.1.2Main dialogue loop

The dialogue now enters a main loop where the user is asked whether he wishes information about account balances or a mini statement giving the latest activity on the specified account. As in the user verification sub dialogue the prompts will be increasingly specific, if the user fails to answer or is misrecognised. Three default account names have been defined (a primary- (løn-), a budget- and a cash credit account). All account references are to the names, thus eliminating the user to memorising the actual account numbers.

5.1.3Balance and Mini Statement Sub Tasks

Depending on the information sought for by the user the dialogue enters two different sub-tasks, Balance or Mini Statement. The Balance sub-task provides information of the balance of the user accounts and the Mini Statement sub-task relates the last three movements on the specified account.

The resulting overall dialogue task structure is shown in Figure 3 (simplified). As mentioned above the overall functionality is identical in Trial 1 and 2.

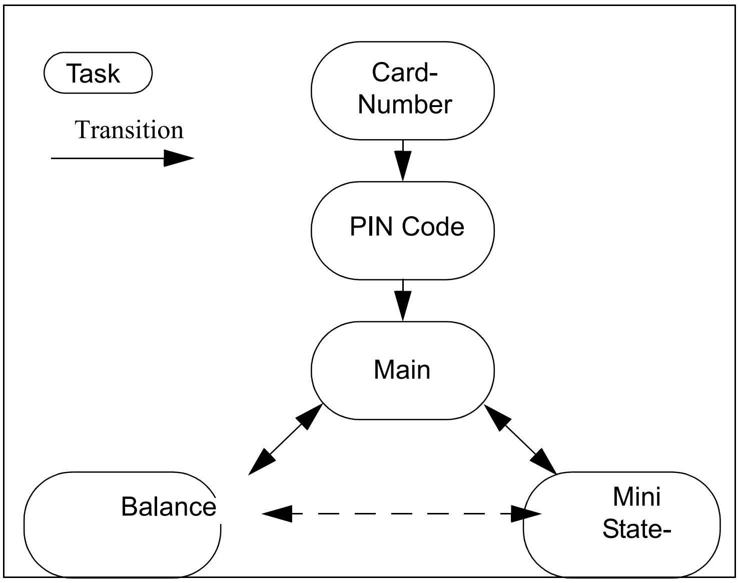


Figure 3. Overall Dialogue Task Structure

The figure shows the five major subtasks constituting the dialogue implementation. In Figure 3 a dashed line shows a transition which can only be performed by the user taking over the dialogue initiative.

5.1.4Error handling - transfer to human operator

If communication errors occur repeatedly, the service will attempt to remedy this, but as a last resort advice the user to contact the bank directly. This will be in the form of asking the user to call his/hers local branch within normal opening hours. In the trial situation this will still be the case, but the subjects will be instructed not to make the call.

5.1.5Dialogue repair and help messages

The dialogue will try to identify and repair communication errors. Implicit confirmation of user utterances will be given in the following system message (where applicable). This ensures that the user always knows the systems assumptions, allowing the user to immediately identify and correct misrecognitions or -understandings. This kind of implicit confirmation saves dialogue turns and lead to a more compact and natural dialogue. In case of repeated errors, the system will use increasingly specific prompt messages to guide the user. Examples of implicit confirmation and repeated prompting are given below.

Example 1. Implicit confirmation (translated from Danish):
USER: “*I want the balance of my cash credit account*”
SYSTEM:“*The balance of your* ***cash credit account*** *is: 3.500 kr.*”

Example 2. Repeated prompting:
SYSTEM:*“Please state your PIN code”.*
USER: “*twenty-three oh six*”
SYSTEM:“*Please repeat your PIN code. Remember to say the number as digits.*”
USER:*“twenty-three - oh sorry, I forgot”*
SYSTEM:*“Please repeat your PIN code. If the system has difficulties understanding you, please use the telephone keypad instead.”*

In general the user is given three attempts to supply e.g Id numbers. Context dependent help messages will be given upon request, or automatically in case of repeating errors.

5.1.6Vocabulary

As mentioned in 4.4. on page 149 speech recognition technique is based on a spotting of predefined keywords and -phrases. This technique allows the user to use natural, unconstrained speech, while at the same time keeping the complexity of the speech recognition task at a reasonable level. Especially the language.

The selection of the keywords is crucial for the success of the trial. The keywords will be selected in a Wizard of Oz (WOZ) simulation experiment prior to the actual trial. In the simulation, the speech recogniser will be replaced by a human "the Wizard", typing the user utterances. The dialogue specification and system messages will be identical to the trial to simulate the trial situation as closely as possible. A set of 30 subjects will be selected to perform the WOZ simulation. The WOZ dialogues will take place over the public telephone network, and be recorded and transcribed and used as a test database to evaluate the recognition accuracy prior to the trial.

As mentioned above a set of keywords is defined during Trial 1. A total of approximately 25 - 50 keywords (including numbers and variations (e.g. case)) are expected to be identified and included in the dialogue. For the speech recognition device, word models are generated from phoneme models previously trained on a 1000 speaker corpus, recorded over the public telephone network.

Account names - commonly used by Danish banking customers (e.g. "cash credit") will be used instead of account numbers.

5.1.7Dialogue initiative.

In 4.4. on page 149 it is stated that the customer must be in control of the interaction. At the most general level, this means that the customer is free to make any statement or command at any time in the dialogue, and the system will react accordingly.

However, this may not always lead to the most natural or efficient mode of communication, as humans in certain situations expect the other part in a conversation to hold the initiative. This is particularly true in situations when one part requests something from another, as in the present case.

Also, the customer may not be fully aware of his or hers options at a given point in the dialogue, and expect some guidance from the system. Therefore a strategy of mixed-initiative dialogue management is chosen to allow the initiative to pass from the user to the system and vice versa in a natural way. This will accommodate novice users, who will be in need of the systems guidance, as well as experienced users, who will want to control the dialogue in order to quickly obtain the desired information.

5.2. The GDS - Generic Dialogue Platform

The dialogues for both Trial 1 and 2 are implemented using the CPK GDS dialogue platform [5]. The platform has been developed over a number of years at CPK and is specially designed for implementation of spoken dialogue systems. It takes care of a long range of tasks that are common to spoken dialogues. This includes:

- Handling of communication between devices

- Allocation and de-allocation of devices
- Error handling
- Execution of the specified dialogue description
- A formalism for implementing dialogues
- Dialogue compiler and runtime debugger

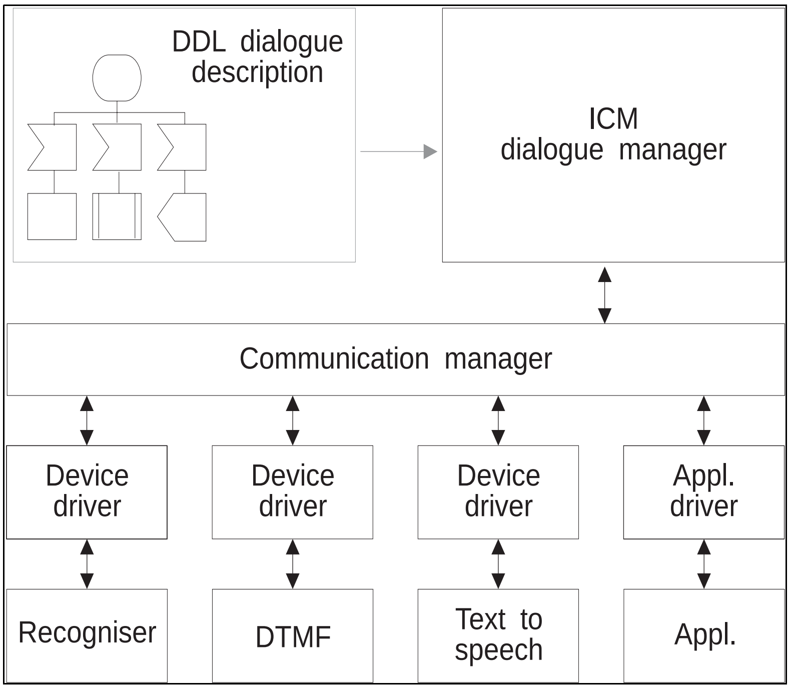


Figure 4. Architecture of the Generic Dialogue System Platform

- All generic tasks are thus handled by the platform, and only the tasks specific to a particular application must be implemented each time a new dialogue is created.

For the OVID application this includes:

- A dialogue graph specification
- Definition of spoken input and output in terms of vocabulary and grammar
- A database of customers and accounts for simulation of the bank

The platform consists of a number of modules constituting the core of a dialogue system as shown on Figure 4.

These are a generic dialogue manager (Interpretation and Control Manager - ICM), a communication system, a well-defined protocol for interfacing to input/output devices and applications, and a graphical tool for describing dialogues. The tool supports the dedicated language DDL - Dialogue Description Language [5].

DDL is a compound language consisting of three levels: a graphical level for describing the dialogue structure and the control structure, a frame level for describing data structures such as lexica, and textual level for implementing computations etc. as in traditional textual based programming languages. The three levels are used in two ways: 1) the frame and textual levels are used for declaring data structures, functions etc. with a global scope, and 2) each graphical symbol occurring in a diagram (see Figure 5) has frame and textual levels attachments that in detail defines the meaning of the symbol.

The DDL Tool supports the developments of dialogue descriptions in DDL. It has the necessary drawing facilities and provides a number of features such as instant syntax check, interactive debugging of the prototype during development etc.

The ICM dialogue manager controls the dialogue on the basis of the description provided by the DDL Tool. It interacts with the speech recogniser in a way which ensures that strong constraints are applied during the speech recognition process. The constraints are determined dynamically during the course of the dialogue. This results in improved reliability for the speech recogniser, and increased performance in terms of speed and overall size of grammar/vocabularies for an application.

5.3. The dialogue specification

The actual discourse of the dialogue is implemented as a state transition diagram, or flow chart like formalism. The dialogue is composed of a set on interconnected diagrams, which again are composed by a number of symbols and interconnecting arcs. Important symbols are the State symbol and Input symbol. Table 4 shows the key figures for the OVID dialogue:

Feature	
Sub Tasks	5 (Id number, PIN code, Main, Balance and Mini Statement)
Dialogue States	24
Prerecorded System Messages	138
Diagrams	The dialogue comprises 43 sub diagrams
Dialogue Symbols	762
Vocabulary	29 (15 digits, yes/no, keywords (c.f. section 7.0.2[7.1.]))

Table 4 Key figures for the dialogue description

Figure 5 shows the overall dialogue flow graph for Trial 2.:

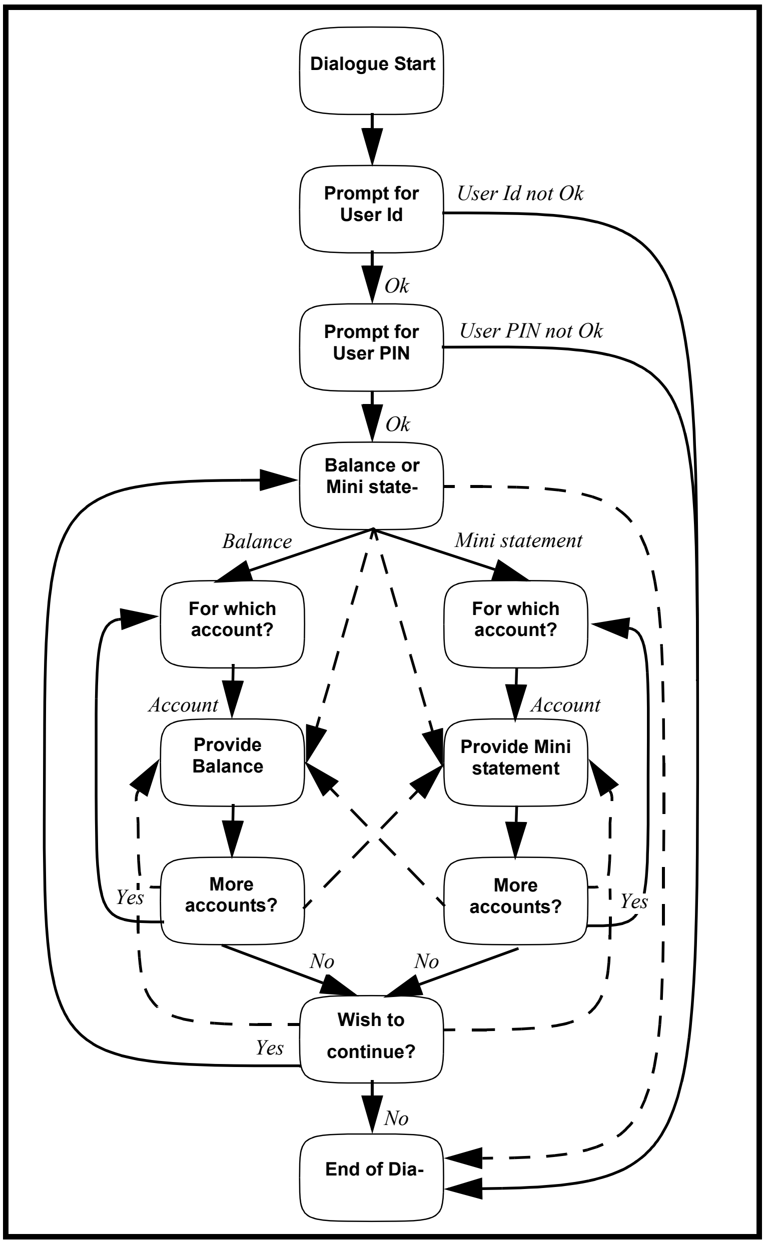


Figure 5. Dialogue flow graph (Trial 2)

The dashed arcs represent user initiated transitions. The text in the boxes denotes system prompts and the text on the arcs denotes the semantics of the user's utterances.

6 Set up of the Trials

As described in the introduction the scope of Trial 1 and Trial 2 is quite different. This is also reflected in the selection of test persons for the trials.

Trial 1:

The objective is to verify the dialogue model and capture the keyword vocabulary for Trial 2. Therefore, no usability statistics are collected (although the questionnaire is sent in order to receive feedback on the translation of the questions, etc.). Consequently, the demographics become of less importance and only a sufficient number of test persons is needed in order to ensure that a reliable estimate of the vocabulary can be made. 20 users were recruited at the university, which seemed acceptable.

Trial 2:

The objective is to capture user responses at a statistical significant level in a field test with real-life banking customers. Therefore a large number of Lån & Spar customers are contacted accordingly to predefined demographic criteria. These are: gender, age, and geographic location (accent). In total 1.200 customers were contacted by Lån & Spar bank, resulting in 330 completing the trial and returning the questionnaires. The aim was no less than 100 participants, which would assure a sufficient number for reliable statistical analyses.

Set up of the trials

An 800 (free phone) number is set up at CPK to allow subjects participating in the trials without costs. The dialogue host (a Linux based PC equipped with a telephone interface board) is coupled directly to the telephone line, allowing the system to run unsupervised throughout the test period.

For Trial 1, test persons were recruited by email, which proved to be an effective means of communication. For Trial 2, a letter was sent from Lån & Spar bank explaining the experiment and asking whether the customer would participate. A phone card (Value 50 DKK) was offered as a reward for participating.

After accepting to participate all subjects receive more detailed information in a letter from CPK containing (see Appendix A):

- A letter introducing the experiment and the Ovid project in general terms.
- Individual Id- and PIN codes.
- A short description of how to use the service.
- Two scenarios for each subject.

- A Questionnaire (a Danish version of the CCIR usability questionnaire)
- An envelope for returning the questionnaire

The actual tests are carried out by the participants calling from their home (or work) at a convenient time. For Trial 1, however, an initial call was required to verify that "the system was ready", but really to ensure that the Wizard was ready to type the user utterances into the system.

6.1. Evaluation

Objective performance measures

All dialogues are recorded and transcribed and information concerning call success rate, individual and overall task duration, and speech recognition accuracy, etc. will be extracted and evaluated. If special problems occur at certain points in the dialogue, they will be investigated and reported.

Attitude measures - the Questionnaire

The CCIR attitude questionnaire for subjective measures will be translated into Danish and used for retrieving the subjects attitudes to different aspects of the trial service.

This will enable the project to compare the British and Danish trial dialogues directly also on the level of subjective measures.

The objective data will be cross-correlated with the findings from the subjective measures, to investigate e.g. relationships between speech recognition accuracy and user confidence/acceptance. Below the British Likert statements are shown.

Sample Questionnaire - Likert statements
The automated banking service was easy to use
When I was using the automated banking service I didn't know what I was expected to do
The automated banking service was friendly
The automated banking service was confusing to use
I would be happy to use the automated banking service again
I felt that the automated banking service was reliable
I felt out of control while using the automated banking service
I liked the voice
I had to concentrate hard to use the automated banking service
I thought the automated banking service was efficient
I got flustered when using the automated banking service
The automated banking service was too fast for me
I felt under stress whilst using the automated banking service
I thought the voice was very clear
Using the automated banking service was frustrating
I would prefer to be given account information by a human being
I thought the automated banking service was too complicated
I enjoyed using the automated banking service
I feel that the service needs a lot of improvement
I thought the automated banking service was polite
I would be confident in the security of the automated banking service
The automated banking service is a convenient way of accessing my account information
I would worry about the confidentiality of information with this service
The details given to me by the automated banking service were accurate
There were too many different things to remember
I think the automated banking service would be good value

Table 5 British Likert statements

7 Trial 1 - Results

The notion of several iterations of test/refinement is commonly used in the design of spoken dialogue systems. In this case three versions (iterations) of the test had originally been planned, but it turned out that a robust dialogue was achieved after two iterations of the test had been completed.

7.0.1Results of the first iteration

The first iteration was carried out with very few users (7), all associated with CPK. The goal of this iteration was to verify that the system would not malfunction during test, that the dialogue did not contain obvious errors, and that the user instructions was adequate and unambiguous.

The users were not required to fill out the usability questionnaire, but gave their comments directly to the experimenter. Most users realised that it was a simulation test.

As a result of the first iteration, the user instructions were corrected at several points. More important, a design error in the dialogue specifications was uncovered. As can be seen in Figure 5 the dialogue splits into two branches one concerning information on balances, the other information on transaction statements. In the first iteration, the dashed arcs allowing transition between the two branches did not exist. A majority of the users tried to make this transition in scenario B (which is a very natural thing to do), and most reported that they felt this should be possible. Therefore, the transitions were included in the second iteration. Evidently, the users focused on the account instead of the type of information.

7.0.2Results of the second iteration

This section presents and discusses the findings of the main experiment. 20 users participated in the experiment. Most were university staff, but a number of users from Tele Denmark also participated.

All users were given the letter with instructions (Appendix A) and the usability questionnaire (the Danish version of the CCIR usability questionnaire (Appendix B)). They were instructed to fill out the questionnaire immediately after completing the two scenarios. They were encouraged to include any comments and suggestions that occurred to them.

As stated previously, the trial is not a usability trial. The number of users is too low to make any statistically significant analysis and conclusions from the questionnaire. In addition, the users were not selected from demographic criteria. Instead, the main purpose of using the questionnaire was to gain experience with it and for it to serve as a channel of feedback for refinements and corrections for Trial 2.

This succeeded. All users returned the questionnaires and only very few of the users did not supply any comments.

Another and more important source of information is the quantitative data collected by logging the dialogues. All dialogues were recorded on tape and transcribed. This serves two purposes: To extract the vocabulary for the word-spotting speech recogniser, and to serve as a test database to verify the performance of the recogniser.

A log file was created for each dialogue containing information on timing, user and system utterances, turn taking etc. Examples of a dialogue transcription and log file are shown in

7.1. Analysis of the quantitative data

A number of parameters were recorded for each dialogue. The number of dialogue turns and the overall completion time are the most important and can be used to evaluate the success of the adopted strategy for the dialogue.

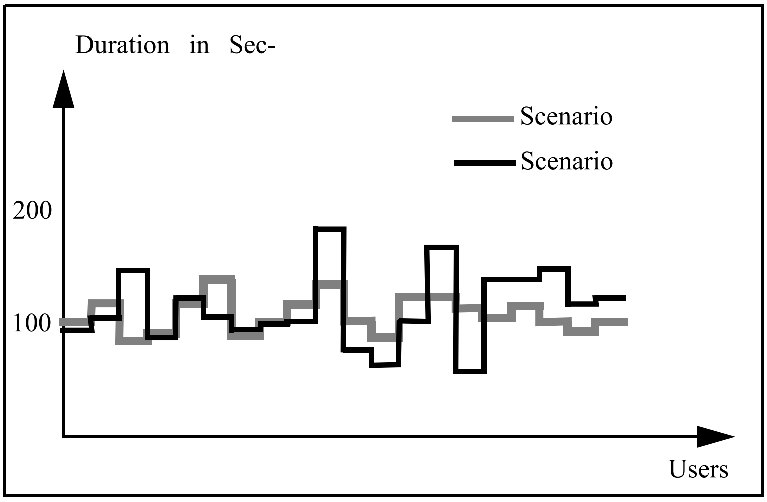


Figure 6. Completion times for scenario A and B

[Figure 6] shows the completion times for the two scenarios. In [Table 6] the key figures for the durations and number of turns of the dialogues are shown. All users completed both scenarios. The nominal number of turns is the number of turns a user would have to go through if he/she answered all system prompts without at any point gaining the initiative. In some cases (as shown in [Figure 5]) the user can optionally supply addi-

Scenario:	A	B
Nominal number of turns	7	9
Minimal number of turns	5	4
Average number of turns	8.1	7.8
Standard Deviation (turns)	1.1	2.3
95% confidence interval (turns)	0.5	1.0
Average duration of dialogues	105	112
Standard deviation (duration)	13.8	31.7
95% confidence interval (duration)	6.1	13.9

Table 6 . Turns and dialogue completion times tional information in many cases, and thereby short-cut the rigid system controlled dialogue structure. Full utilisation of this yields the minimal number of turns[1].

Some indications that this in fact happens can be observed from [Figure 6] and [Table 6]. Although the nominal number of turns is higher for scenario B, the average is the same for scenario A and B. This indicates that the users to some extent have started to learn how to make short-cuts in the dialogue. From [Figure 6] (and the last two rows in [Table 6]) it can be seen that there is much greater variation in the completion times for scenario B than for A. This indicates that some users have "learnt the tricks", whereas others still follow the more rigid dialogue structures. As no information of the dialogue short-cuts were given to the users prior to the experiment this separation into two groups was to be expected.

This effect is even more clear when the two turns used to gain access to the service is excluded from the calculations as the access procedure cannot be circumvented or reduced[2].

7.1.1Identification of user vocabulary

One of the main goals of the present trial is to obtain a keyword vocabulary to be used for Trial 2. In total 39 keywords were identified:

Digits + yes/no:15 (including some variants)
Accounts, etc.14
Other forms:10 (almost all keywords appear with two endings, due to Danish definite/indefinite forms)

In total more that 100 different words were recorded. The vocabulary seems to be closed, and all users assumed the formulations in the system messages.

All the used keywords were in fact anticipated in the dialogue, with the possible exception of "goodbye", which many users said when they had obtained the informations they desired.

7.2. Additional comments from the users

Barge-In A large proportion of the users proposed that the service should allow barge-in (that the user can cut off the system). This was not the case in Trial 1. None of the users tried it, though (but it was stated in the instructions that barge-in wasn't possible).

[1] This includes the user hanging up immediately after the desired information has been obtained.

[2] Unless the users are allowed to say both their Id- and PIN numbers in one sentence. However this clashes with the requirement that the user Id verification procedure must adhere to the present Danish practice.

Speed Most users complained that the service was too slow to answer. This was due to the wizard setup.

Dialogue short-cuts Some users complained that transition from balances directly to mini statement was not possible. This was not evident from the system prompts, but as discussed earlier it was possible anyway, and a number of users utilised it.

Privacy Many users felt they might be overheard when saying their Id- and PIN codes, and suggested that DTMF should be included in the service. This is the case already, but the users were not informed (due to reasons explained earlier). Some users tried it (successfully) on their own initiative.

7.3. Conclusions

A spoken dialogue system has been designed, implemented and evaluated in accordance with the specifications described in [1] and [2].

The Trial was a WOZ simulation test, with the additional purpose to capture the expected user vocabulary in preparation of Trial 2. The implementation of the spoken dialogue has been designed so it can be used with minimal modifications in Trial 2.

The emphasis has been put on naturalness and flexibility of the spoken input/output and the dialogue structure. The user test indicates that this goal has been accomplished, as the users with minimal instructions were able to gain the initiative and short-cut the system controlled dialogue structure.

A limited set of keywords has been identified as adequate for the service to perform the commands issued by the users. The set consists of the digits and yes/no and 14 other keywords identifying accounts etc. Furthermore, the set of keywords can be split in two, as digits are only used in the access procedure, and the other keywords only when eliciting informations from the service. All keywords were anticipated beforehand, so the trial confirmed this.

Feedback from the users indicates that handling of barge-in should be included, as well as optional DTMF input.

8 Trial 2 - Results

This section presents the results of Trial 2. First, the demographic distribution of the test subjects are briefly discussed. This is followed by a presentation of the actual findings. These fall into two categories, results of the usability questionnaires and statistical information obtained from the logging of the dialogues.

8.1. Test Subjects

The users are all recruited among Lån & Spar Bank (L&S) customers. Initially, L&S provided a list with 5.000 customers with information on names, address, age and gender. This formed the basis for a selection of a balanced set of potential test subjects. The list was evenly distributed according to age, gender and geographic region. The users were divided into 5 age groups and 4 geographic regions. By dividing into geographic regions it is assumed that Danish regional accents[1] will be evenly represented, and that both users from urban and non-urban areas will be represented.

Criteria	Category	Percentage
Sex	Male	55%
	Female	45%
Age	18 to 29	16%
	30 to 39	23%
	40 to 49	22%
	50 to 59	23%
	Above 60	17%
Region	København	29%
	Sjælland, Bornholm	23%
	Fyn, Sønderjylland	21%
	Midt- and Nordjylland	27%

Table 7 Demographics of test subjects

L&S mailed a letter explaining the objectives of the OVID project to a total of 1128 customers. 369 (33%) of the contacted customers responded positively, and of these 329 (30%) later completed the dialogue scenarios and returned the questionnaire. [Table 7] shows the demographic distribution of the test subjects. Apart from a bias against the youngest and oldest age groups the resulting distribution seems satisfactory.

8.2. Results

This section presents the actual findings of the Trial 2 experiment. Two information sources are utilised in order to obtain the results presented here. One is the usability questionnaires filled out and returned by the users. The other is the statistics of the dialogue logfiles. This combination of subjective and objective information can provide valuable information about possible causes for e.g. negative user responses and hence lead to improvements of dialogue design of speech recognition modules.

[1] The regions correspond only very roughly to the major Danish accent regions, but still serves to ensure that users from some geographical area do not dominate the test.

Overall results from the usability questionnaires.

The questionnaire is an adapted and translated version of the CCIR Likert questionnaire, and the theory is not explained in detail here. The user is asked to express his/her degree of agreement to a total of 25 so-called Likert statements. The categories ranging from "strongly disagrees" over "neutral" to strongly agrees" are translated into a scale from 1 to 7 (with 4 as the neutral).

□	□	□	□	□	□	□
Strongly Agree	**Agree**	**Slightly Agree**	**Neutral**	**Slightly Disa-**	**Disagree**	**Strongly Disagree**
□	□	□	□	□	□	□
Meget Enig	**Enig**	**Lidt Enig**	**Neutral**	**Lidt Uenig**	**Uenig**	**Meget Uenig**
7 (1)	**6 (2)**	**5 (3)**	**4**	**3 (5)**	**2 (6)**	**1 (7)**

Figure 7. Check boxes from usability questionnaire.

Some of the statements are "inverted", i.e. the user must disagree in order to express a positive opinion of the system. This is done to force the users not to tick all boxes more or less uniformly. The corresponding responses are negated in order to make comparisons (shown in parentheses in Check boxes from usability questionnaire.). Thus, the higher the numbers, the more positive attitude. A number of diagrams showing the results are presented and discussed on the following pages.

The averages of the user attitude responses are shown in Figure 8 together with the 98% confidence intervals. In general, the responses are positive with an overall average of 5.5. Among the categories with the most positive responses are voice, ease of use, convenience and complexity. Categories such as concentration, speed, needs improvement and confidentiality are below the average. As can be seen from the figure, there seems to be a tendency towards higher confidence intervals for the least scoring categories. This implies that users tend to disagree more in these categories.

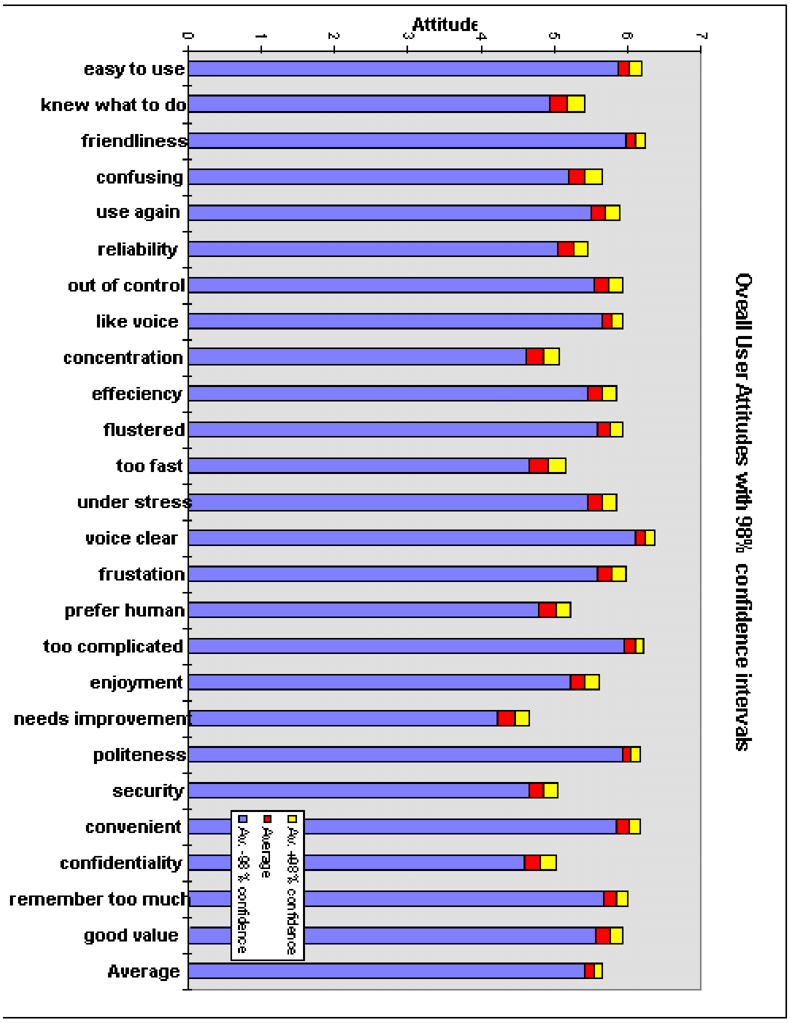


Figure 8. Overall user attitudes with confidence intervals

The overall responses are broken down into male/female in [Figure 9]. It is clear from the figure that no significant differences occurs (note that the scale has been changed to emphasize differences). The overall average is slightly higher for females than for males. Female responses tend to be slightly higher for questions concerning the appearance of the service, e.g. voice, friendliness and politeness, whereas male responses are slightly higher for categories related to the cognitive load, such as stress, distraction, concentration and complexity.

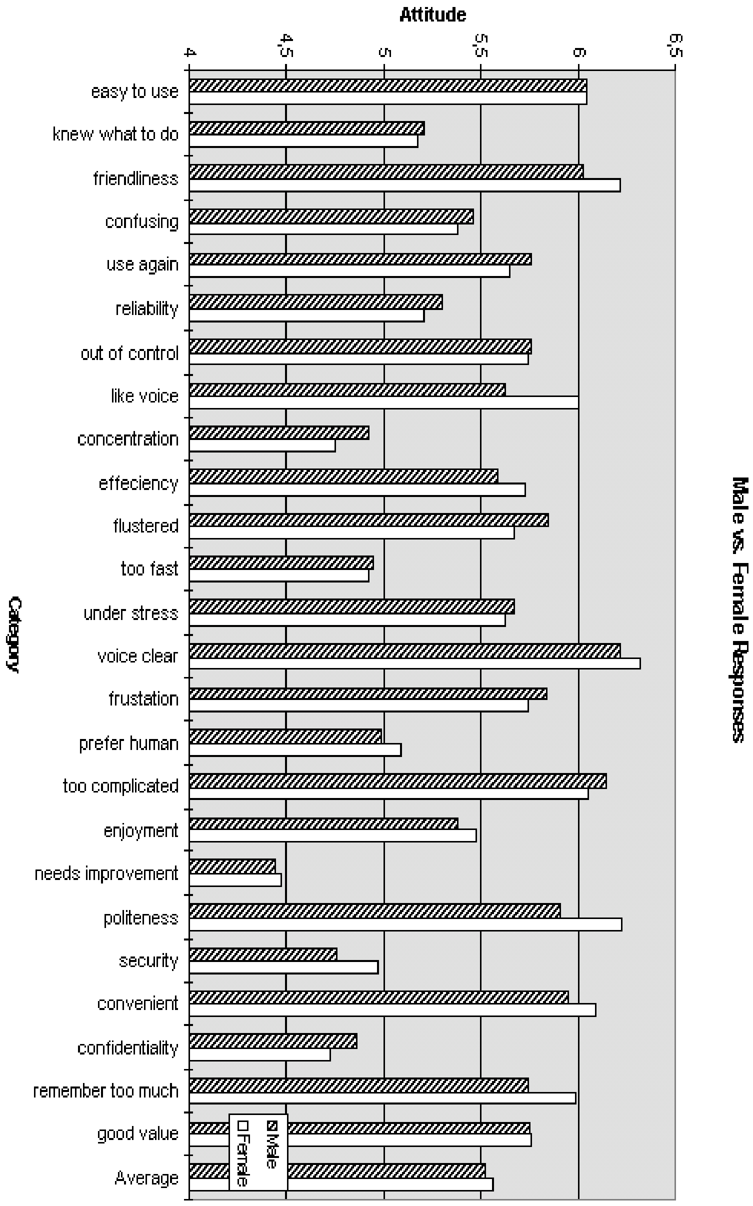


Figure 9. Male vs. female responses

The figures presented below shows the average attitudes according to age and region, as discussed in section [8.1.]. The overall attitude as a function of age is shown in [Figure 10]. Slightly surprising the attitude for younger users seems to be less positive than for the other age groups.

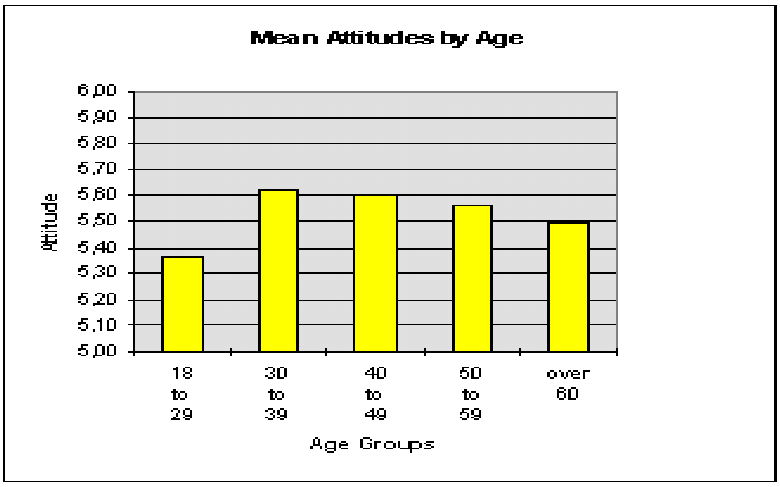


Figure 10. Attitudes according to age

Figure 11 shows the average user responses as a function of geographic location. Only very minor differences can be observed.

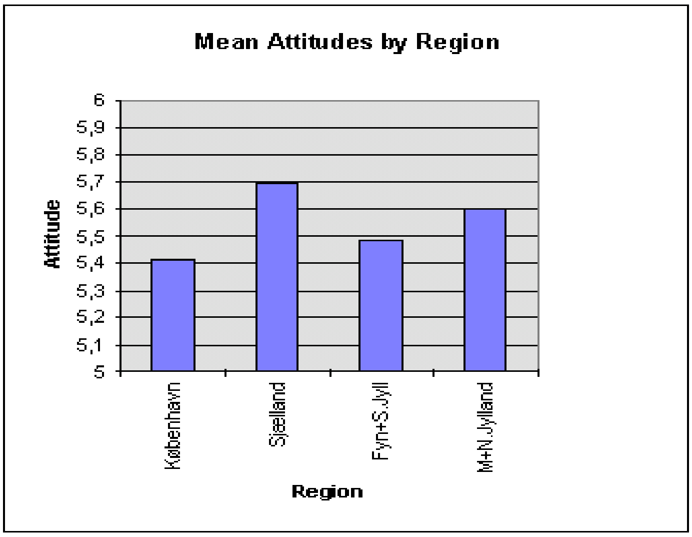


Figure 11. Attitudes according to region

8.3. Quantitative information derived from logging of dialogues

This section presents and discusses the statistics derived from the logging of the dialogues. Focus has been to verify whether the users behave as anticipated by the dialogue model discussed in Chapter [5].

The number of dialogue turns spent in each subtask is inspected as well as the extent the users actually take the dialogue initiative. All users are requested to do to dialogues: “A” and “B”. The task for scenario A is to obtain a balance for three different accounts. In scenario B the user has to obtain a balance and mini statement for one account. The scenario index refers to whether the scenario was carried out as the first or second call. Thus, one half of the users performed (A1,B2) and the other half (B1,A2).

Scenario:	**A1**	**A2**	**B1**	**B2**
Nominal number of turns	7		9	
Minimal[a] number of turns	5		4	
Average number of turns	8.4	7.8	7.5	7.1
Average duration of dialogues (seconds)	94	86	92	84
Nominal number of user initiatives[b]	1		2	
Average number of user initiatives	0.8	1.0	1.3	1.6

Table 8 Key figures for scenario A and B

[a] This includes the user taking the dialogue initiative whenever possible, and hanging up immediately after the desired information has been obtained, which none of the users did. This can be expected of more experienced users, though.

[b] User hang-up is not counted as a user initiative

Table 8, Figure 12 and Figure 13 show that users behave as expected and that some habituation effects are present in the second dialogue, as the average times and number of turns are less in all cases for the second dialogue. For scenario A the expected number of turns are 1 for Id- and PIN

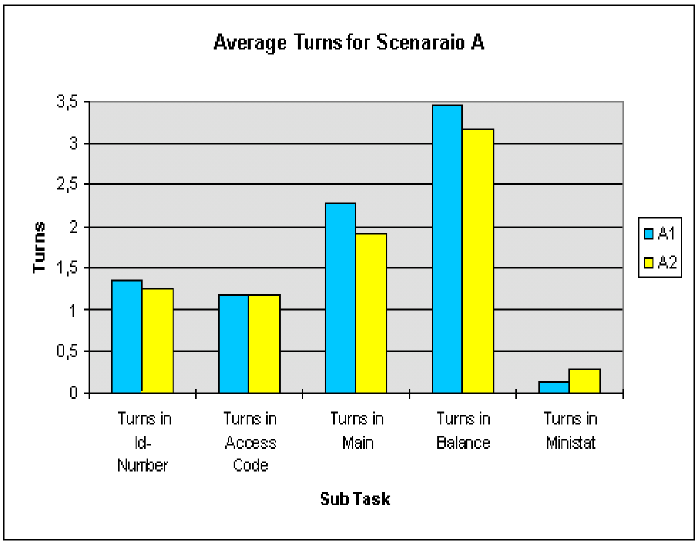


Figure 12. The number of turns spent in each subtask in scenario A

number verification, two in the main task and three to obtain the three account balances. For scenario B one turn is expected to obtain the balance and one to get the ministatement. [Figure 12] and [Figure 13] show that this is very close to the actual numbers, thus demonstrating that the dialogue model behaves as predicted and that few speech recognition errors occur.

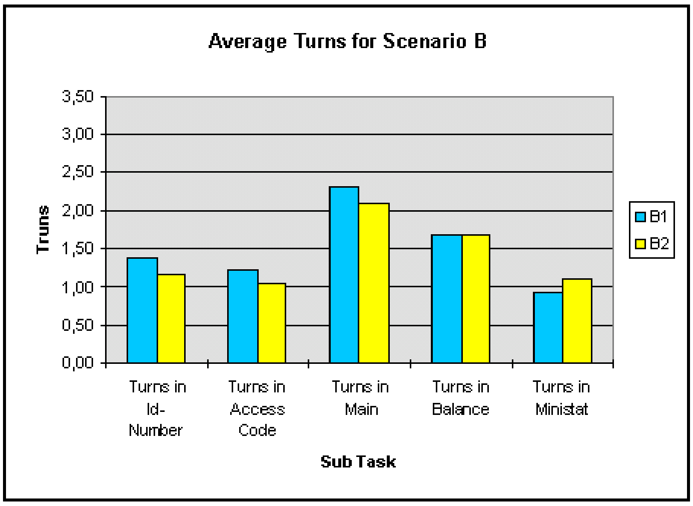


Figure 13. The number of turns spent in each subtask in scenario B

Table 2 and figure 4.7 shows the average number of user initiatives per dialogue. The expected number is 1 for scenario A and 2 for scenario B. The figure shows that this is the case for scenario A, whereas scenario B is less clear. Habituation effects are also evident in this case as the number of user initiatives clearly increases for the second dialogue.

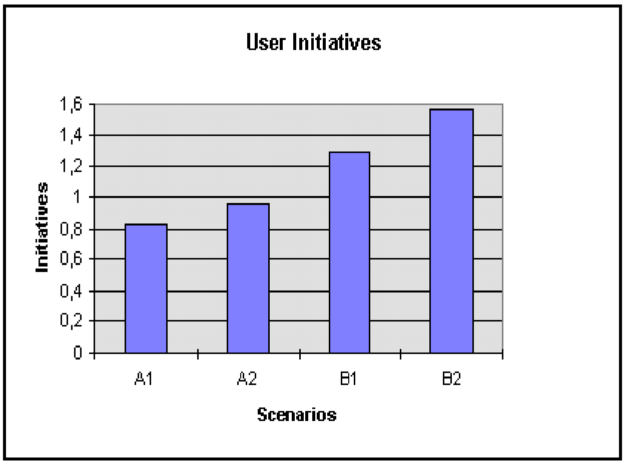


Figure 14. User initiatives per dialogue

An interesting point to examine is the duration of the user identification and verification procedures as compared to the total dialogue. This is shown in [Table 9] below.

Task	A1,B1	A1,B1	A2,B2	A2,B2
Id number	20.8 secs	22 secs	17.4%	20%
Access code	11.3 -	12 -	9.8 -	12 -
Id + Access code	32.1 -	33 -	27.2 -	32 -
Total dialogue duration	93 -	100 -	85 -	100 -

Table 9 . Duration of user authentication procedure

From the table it can be observed that on average one third of the call is spent on the user authentication procedure, corresponding to approximately 30 seconds. It is also evident that users spend almost twice the time entering the id number compared to the access code. The reason for this that the id number is a 7 digit code whereas the access code only contains 4 digits.

8.4. Further comments from the users

The users were invited to express their comments to the service, and almost all did so. The most common impression was that the mini statement was to fast to note down. This is also evident from [Figure 8], where statements about concentration and speed are below average.

A number of the users expressed very positive or negative opinions, depending on their experiences.

Most users described any specific problems they encountered, such as: "*It* (the service) *could not recognise my id number the first time, but the second time it was ok*". The comments prove a valuable source of information when examining which errors users perceive as more severe, and which doesn't seem to be of importance.

Users were not informed about the DTMF option. Only in case of repeated recognition errors did the system suggest that they use the telephone keypad. Consequently many users expressed concern that their Id and Access numbers might be overheard. This concern has also influenced the attitude towards confidentiality, as shown in [Figure 8].

8.5. Speech recognition performance

All users were required express their attitudes towards the system by filing out a usability questionnaire developed at CCIR at Edinburgh university. It consists of a set of Likert statements shown in section [6.1.] and the Danish translation (see [Appendix B]). The user expresses his/her attitude by ticking boxes ranging from "strongly agree" to "strongly disagree" to each statement. The statements are categorized into usability aspects such as: quality of interface/performance, cognitive effort/stress, conversational model, fluency and transparence. The responses are transformed into a scale ranging from one to seven, with four as a neutral attitude. Figure 4 shows the average of the users attitude as a function of the experienced speech recognition accuracy. In [Figure 15] users have been grouped according to the degree of recognition accuracy they experienced when using the system. As can be observed from the figure, the accuracy has very little influence on the attitude towards the system.

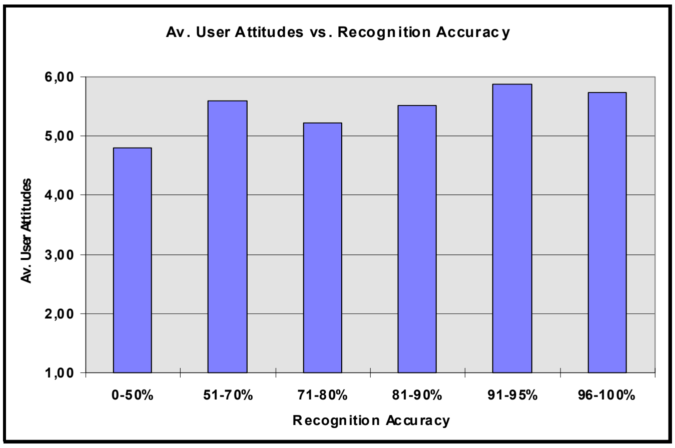


Figure 15. Average user attitudes vs recognition accuracy

It may seem surprising, but the recognition accuracy is one of many factors influencing users attitudes towards an application. The categories “0-50%” and “50-70%” only comprises approximately 10% of the users (see [Figure 16]). The overall speech recognition accuracy is 84%., including errors made by the speech detector. On average each user say nine keywords[1] per dialogue, which in turn means that they may expect about one misrecognition per dialogue.

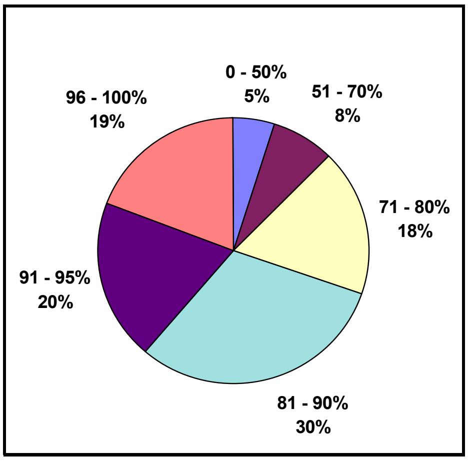


Figure 16. Users grouped according to experienced recognition accuracy

[1] The Id- and PIN codes are here counted as one “keyword”, although they actually consist of 7 and 4 digits. It turns out that digit string and keyword accuracy are roughly similar.

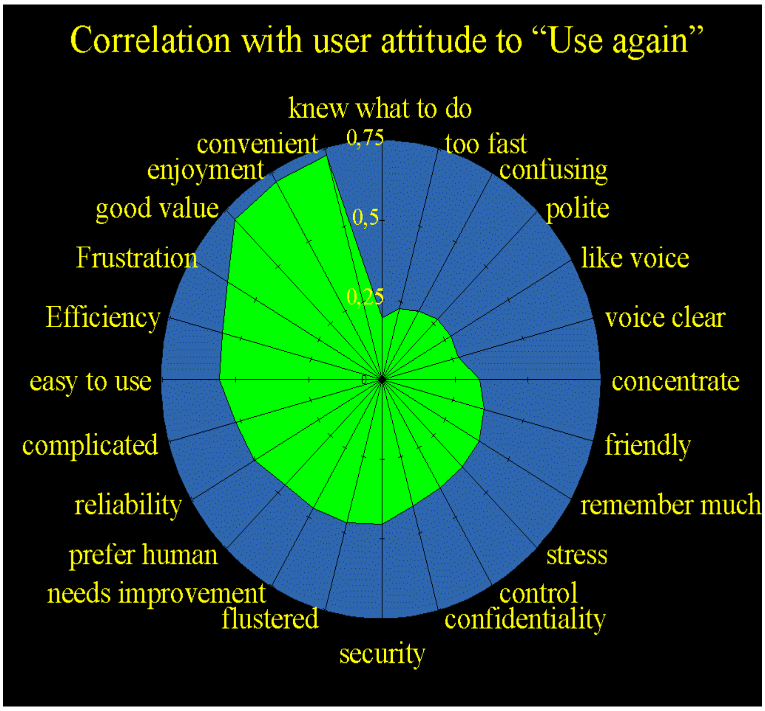


Figure 17. Likert statements correlation with the statement "I would like t use the OVID home banking service again"

Figure 17 shows the correlation between a specific statement and the remaining statements in the questionnaire. The question of whether the user "would like to use the service again" has been chosen to depict the overall attitude towards the service, and the figure shows how well the remaining statements corresponds to this. As the figure shows, issues like "convenience" and "efficiency" seems to be important to users, whereas e.g. the quality of the voice is less so. The speech recognition accuracy has a correlation of 0.3, i.e. none or very little impact on the overall attitude towards the service.

9 Conclusions

9.1. The User Attitude questionnaire

The results clearly show that:

- The service was accepted positively by the users as is evident from the results of the user attitude measurements. It must be noted that all test subjects are experienced VR users
- No major differences in attitudes were found across age, sex and region
- Users found that the account information given by the system was quoted too fast - a problem that can be easily solved
- Some users expressed concern about the confidentiality of the service. DTMF as an alternative input mode was not communicated to the users before the test. If this had been the case this problem would have been solved or reduced
- The dialogue model proved to perform as predicted and users were able to gain the initiative in the dialogue
- Although not perfect, the performance of the speech recogniser seemed to be acceptable to the users.
- The overall recognition accuracy was 84% on average, but differed substantially for individual users
- No major differences in user attitude depending on the experienced speech recognition accuracy could be documented.
- Issues like convenience and efficiency correlated well with the users overall attitude towards the service, whereas e.g. speed, confusion and voice were less important.

9.2. The dialogue model

The emphasis has been put on naturalness and flexibility of the spoken input/output and the dialogue structure. The user tests indicate that this goal has been accomplished, as the users were able to start gaining the initiative and short-cut the system controlled dialogue structure without prior instructions or informations about this opportunity. Figure 2 shows that users immediately start to go beyond the limits of the system directed dialogue structure and utilizing the built-in short-cuts. An even more pronounced effect of user habituation can be expected when users are exposed to a larger number of dialogues.Speech recognition error rates for digit strings and word spotting were found to be 16% on average. This seemed sufficient to ensure user acceptance, but further stresses the importance of the dialogue design to ensure high user acceptance. The preceding sections have shown that the strategy of maintaining a system directed dialogue on the surface and then provide *short-cuts* for more experienced users has proved successful.

The aim of the experiments was to investigate to what extent customers are prepared to accept voice controlled access to their bank accounts. The preliminary results strongly indicates that this is the case. The tested dialogue was very small, containing only a few sub tasks, so the next step will evidently be to expand the dialogue to cover a larger number of tasks, and a more complex task structure.

The key requirements for the service were formulated by the OVID banks and stated that:
"The user must feel in control" and
"The user must be able to speak naturally"

The conclusions suggest that these requirements have been fulfilled successfully.

9.3. Further experiments

Although the OVID experiments give some clear indications of the expected user attitudes towards this kind of service, some questions remains unanswered. The prototype service was very limited in functionality. A further investigation with a more complex, realistic application domain is needed before conclusive answers can be given, both with regard to the dialogue management paradigm, speech recognition accuracy and customer acceptance. Ideally, two (or more) prototypes with identical functionality but with different dialogue management strategies should be implemented and evaluated in order to obtain a conclusive answer.

Appendix A Information to Subjects

- Aalborg, den 1 Juli 1997

Til: fornavn efternavn

Tak for at du vil hjælpe med at teste den talestyrede OVID telefonbank service. Denne beskrivelse indeholder dels nogle praktiske oplysninger, samt et spørgeskema som jeg vil bede dig om at udfylde *umiddelbart efter* du har gennemført opkaldene. Desuden er der et telekort som tak for din hjælp.

Baggrund

Center for Personkommunikation på Aalborg Universitet og Lån & Spar Bank deltager for tiden i et EU-projekt, der går ud på at undersøge mulighederne for at anvende taleteknologi indenfor homebanking ("ring til din konto"). Projektets navn er OVID. Ideen er, at man i stedet for at benytte telefonens taster blot taler til systemet.

Hvad skal du gøre?

På bagsiden er der beskrevet to opgaver. Du ringer til det angivne telefonnummer og udfører opgaverne. Det tager ca. 3-4 minutter pr. opkald. Derefter udfylder du spørgeskemaet og returnerer det. Alt i alt tager det ca. et kvarter.

Hav tålmodighed, hvis nummeret skulle være optaget. Du kan ringe når som helst på døgnet alle dage indtil den 30. maj.

Hvad bliver det brugt til?

Resultaterne af testen vil blive brugt til at vurdere om det er realistisk at gå fra de kendte trykknap-styrede systemer til talsestyrede systemer.

Du deltager i et videnskabeligt eksperiment. Dit spørgeskema bliver behandlet fortroligt, og alle personlige oplysninger bliver slettet efter testen.

Yderligere oplysninger.

Hvis du har adgang til internettet, kan du se en nærmere beskrivelse af projektet og også resultaterne af denne test, når de foreligger.

Adressen er: http://www.cpk.auc.dk/~lbl/OVID/

Har du spørgsmål eller kommentarer er du velkommen til at kontakte mig.

Med venlig hilsen

Civilingeniør Lars Bo Larsen,

Center for Person Kommunikation, Aalborg universitet

Fredrik Bajers Vej 7A 9220, Aalborg Ø,

Telefon: 9635 8635, Fax: 9815 1583,Email:lbl@cpk.auc.dk

Opgaver

Beskrivelse af Opgaverne:

Du er kunde i den opdigtede OVID bank, hvor du har en lønkonto, en budgetkonto og en kassekredit. Du kan ringe til banken og få oplyst hvad der står på dine konti, samt hvad de seneste bevægelser har været. Du skal altid opgive dit *kortnummer* og din *adgangskode* for at få adgang til banken. Numrene skal altid udtales som *cifre* i sammenhæng (altså ni-to-tre-... og ikke ni hundrede tre og tyve...).

Bemærk at systemet *ikke kan afbrydes,* når det taler. Du kan svare så snart systemet har talt færdigt.

Når du har sagt noget lyder der et kort **Bip!**. Det betyder at systemet har hørt at du sagde noget og nu er i gang med at tolke det. Der kan forekomme en kort pause efter Biplyden. Vent blot til systemet svarer.

Hvis der skulle være problemer med at gennemføre opgaverne, så læg på og prøv igen lidt senere. Du bedes under alle omstændigheder udfylde og returnere spørgeskemaet i den vedlagte kuvert.

Nedenfor er de to opgaver beskrevet.

Opgave 1:

a. Ring op til OVID bank på tlf. 8081 5535 (modtageren betaler opkaldet). Du kan **kun** ringe fra en almindelig telefon, **ikke** fra en mobiltelefon,

b. Opgiv dit kortnummer: **9236702** og din adgangskode: **8234** når systemet beder om det.

c. Få oplyst: Indestående på din lønkonto,

indestående på din budgetkonto,

samt indestående på din kassekredit

d. Afslut opkaldet.

Opgave 2:

a. Ring op til OVID bank på tlf. 8081 5535

b. Opgiv dit kortnummer: **9236702**, og din adgangskode: **8234** når systemet beder om det.

c. Få oplyst:Indestående på din budgetkonto,

og de seneste bevægelser på budgetkontoen.

d. Afslut opkaldet.

Til Slut:

Husk at returnere spørgeskemaet. Det er en vigtig del af eksperimentet.

Appendix B Usability Questionnaire

Spørgeskema til OVID Telefonbank Eksperimentet.

Navn: ____ Dato: __

1 OVIDs Telefonbank var let at bruge

□	□	□	□	□	□	□
Meget Enig	Enig	Lidt Enig	Neutral	Lidt Uenig	Uenig	Meget Uenig

2 Da jeg anvendte OVIDs Telefonbank var jeg af og til i tvivl om hvad jeg skulle gøre

□	□	□	□	□	□	□
Meget Enig	Enig	Lidt Enig	Neutral	Lidt Uenig	Uenig	Meget Uenig

3 OVIDs Telefonbank var venlig

□	□	□	□	□	□	□
Meget Enig	Enig	Lidt Enig	Neutral	Lidt Uenig	Uenig	Meget Uenig

4 OVIDs Telefonbank var uoverskuelig at bruge

□	□	□	□	□	□	□
Meget Enig	Enig	Lidt Enig	Neutral	Lidt Uenig	Uenig	Meget Uenig

5 Jeg ville gerne benytte OVIDs Telefonbank igen

□	□	□	□	□	□	□
Meget Enig	Enig	Lidt Enig	Neutral	Lidt Uenig	Uenig	Meget Uenig

6 Jeg synes OVIDs Telefonbank var pålidelig

□	□	□	□	□	□	□
Meget Enig	Enig	Lidt Enig	Neutral	Lidt Uenig	Uenig	Meget Uenig

7 Jeg mistede overblikket når jeg brugte OVIDs Telefonbank

□	□	□	□	□	□	□
Meget Enig	Enig	Lidt Enig	Neutral	Lidt Uenig	Uenig	Meget Uenig

8 Jeg kunne lide stemmen

□	□	□	□	□	□	□
Meget Enig	Enig	Lidt Enig	Neutral	Lidt Uenig	Uenig	Meget Uenig

9 Jeg måtte koncentrere mig meget for at bruge OVIDs Telefonbank

□	□	□	□	□	□	□
Meget Enig	Enig	Lidt Enig	Neutral	Lidt Uenig	Uenig	Meget Uenig

10 Jeg synes OVIDs Telefonbank var effektiv

□	□	□	□	□	□	□
Meget Enig	Enig	Lidt Enig	Neutral	Lidt Uenig	Uenig	Meget Uenig

11 Jeg blev forvirret af at benytte OVIDs Telefonbank

□	□	□	□	□	□	□
Meget Enig	Enig	Lidt Enig	Neutral	Lidt Uenig	Uenig	Meget Uenig

12 OVIDs Telefonbank var for hurtig for mig

□	□	□	□	□	□	□
Meget Enig	Enig	Lidt Enig	Neutral	Lidt Uenig	Uenig	Meget Uenig

13 Jeg følte mig presset når jeg brugte OVIDs Telefonbank

□	□	□	□	□	□	□
Meget Enig	Enig	Lidt Enig	Neutral	Lidt Uenig	Uenig	Meget Uenig

14 Jeg synes stemmen var meget tydelig

□	□	□	□	□	□	□
Meget Enig	Enig	Lidt Enig	Neutral	Lidt Uenig	Uenig	Meget Uenig

15 Det var frustrerende at bruge OVIDs Telefonbank

□	□	□	□	□	□	□
Meget Enig	Enig	Lidt Enig	Neutral	Lidt Uenig	Uenig	Meget Uenig

16 Jeg ville foretrække at få oplysningerne fra en person

□	□	□	□	□	□	□
Meget Enig	Enig	Lidt Enig	Neutral	Lidt Uenig	Uenig	Meget Uenig

17 Jeg synes OVIDs Telefonbank var for kompliceret

□	□	□	□	□	□	□
Meget Enig	Enig	Lidt Enig	Neutral	Lidt Uenig	Uenig	Meget Uenig

18 Jeg kunne lide at bruge OVIDs Telefonbank

□	□	□	□	□	□	□
Meget Enig	Enig	Lidt Enig	Neutral	Lidt Uenig	Uenig	Meget Uenig

19 Jeg synes OVIDs Telefonbank kunne trænge til mange forbedringer

□	□	□	□	□	□	□
Meget Enig	Enig	Lidt Enig	Neutral	Lidt Uenig	Uenig	Meget Uenig

20 Jeg synes OVIDs Telefonbank var høflig

□	□	□	□	□	□	□
Meget Enig	Enig	Lidt Enig	Neutral	Lidt Uenig	Uenig	Meget Uenig

21 Jeg vil være tryg ved sikkerheden i OVIDs Telefonbank

□	□	□	□	□	□	□
Meget Enig	Enig	Lidt Enig	Neutral	Lidt Uenig	Uenig	Meget Uenig

22 OVIDs Telefonbank er en bekvem måde at få konto oplysninger på

□	□	□	□	□	□	□
Meget Enig	Enig	Lidt Enig	Neutral	Lidt Uenig	Uenig	Meget Uenig

23 Jeg ville være bekymret over oplysningernes fortrolighed i OVIDs Telefonbank

□	□	□	□	□	□	□
Meget Enig	Enig	Lidt Enig	Neutral	Lidt Uenig	Uenig	Meget Uenig

24 Der var for mange forskellige ting at huske på

□	□	□	□	□	□	□
Meget Enig	Enig	Lidt Enig	Neutral	Lidt Uenig	Uenig	Meget Uenig

25 Jeg tror OVIDs Telefonbank ville være en god ide

□	□	□	□	□	□	□
Meget Enig	Enig	Lidt Enig	Neutral	Lidt Uenig	Uenig	Meget Uenig

26 Var der noget du havde særlige problemer med?

27 Lykkedes det at gennemføre opgaverne?

Ja, begge to Kun opgave 1 Kun opgave 2 Ingen af dem

28 Har du nogen yderligere kommentarer, forslag eller indtryk?

27) Må vi kontakte dig igen senere til en eventuelt lignende undersøgelse?

Spørgeskemaet bedes returneret til Center for PersonKomminikation i den frankerede svarkuvert. Tak for hjælpen.

Appendix C Trial 1 Sample dialogue transcription and log file

This appendix contains transcriptions and log summaries for two dialogues: one of type A and one of type B:

Transcription of Dialogue A:

```
sys      velkommen til OVID banks automatiske
         telefon service.
         oplys venligst Deres kortnummer.
user     9286553
sys      tak
         oplys venligst Deres adgangskode.
user     8287
sys      tak
sys      ønsker De at få oplyst indestående
         eller de seneste bevægelser på deres
         konto?
user     (øhh) indestående
sys      saldoen på deres lønkonto er.
sys      ønsker De yderligere oplysninger om
         indestående på Deres lønkonto, budget-
         konto eller kassekredit
user     budgetkonto
sys      saldoen på deres budgetkonto er.
sys      ønsker De yderligere oplysninger om
         indestående på Deres lønkonto, budget-
         konto eller kassekredit
user     kassekredit (host)
sys      ønsker De yderligere oplysninger om
         indestående på Deres lønkonto, budget-
         konto
         eller kassekredit
user     nej
sys      ønsker De at fortsætte?
user     nej
sys      tak for opringningen, farvel
```

Summary of log file for Dialogue A:

```
OVID Trial WOZ experiment, version 1.1 November
         1996 CPK/lbl
-----------------------------------------------
         -------------
Summary of Dialogue: 961107.102527
Dialogue type             : A
User Id                   : 9286553
User PIN                  : 8287
Duration of Dialogue      : 87 seconds
# turns                   : 7
# recorded user utterances: 6
Transaction Report:
Total # of Transactions: 5
 Transaction Type        : occurrences
 Customer number      : 1
 Customer PIN           : 1
 Balance account 1      : 1
 Balance account 2      : 1
 Balance account 3      : 1
End of Summary
===============================================
         ==
```

Transcription of Dialogue B:

```
sys      velkommen til OVID banks automatiske
         telefon service.
         oplys venligst Deres kortnummer.
user     9643267
sys      tak
         oplys venligst Deres adgangskode.
user     8648
sys      ønsker De at få oplyst indestående
         eller de seneste bevægelser på deres
         konto?
user     indestående på budgetkontoen
sys      saldoen på Deres budgetkonto er...
sys      ønsker De yderligere oplysninger om
         indestående på Deres lønkonto, budget-
         konto
         eller kassekredit
user     nej jeg vil hellere have nogle bevægel-
         ser på min budgetkonto
sys      de seneste bevægelser på deres budget-
         konto er....
sys      ønsker De yderligere oplysninger om de
         seneste bevægelser på Deres lønkonto,
         budgetkonto eller kassekredit?
user     nej - det er fint tak skal du have -
         farvel
sys      ønsker du at fortsætte?
user     nej - farvel
sys      tak for opringningen - farvel
```

Summary of log file for Dialogue B:

```
OVID Trial WOZ experiment, version 1.1 November
         1996 CPK/lbl
-----------------------------------------------
         ------------------------------
Summary of Dialogue: 961107.135722
Dialogue type            : B
User Id                  : 9643267
User PIN : 8648
Duration of Dialogue     : 95 seconds
# turns                  : 6
# recorded user utterances: 5
Transaction Report:
Total # of Transactions   : 4
 Transaction Type       : occurrences
 Customer number         : 1
 Customer PIN            : 1
 Ministat account 2     : 1
 Balance account 2       : 1
End of Summary
===============================================
         ==
```

10 References

[1] ESPRIT 20171 Project OVID - Trial application of Voice Processing in Automated Telephone Banking Services: “Technical Annex”, July 1995.

[2] ESPRIT 20171 Project OVID - Trial application of Voice Processing in Automated Telephone Banking Services: “User Requirements (Deliverable D1)” CCIR Edinburgh March 1996.

[3] “L.B. Larsen A. Baekgaard, “Rapid Prototyping of a Dialogue System using a Generic Dialogue Development Platform” in Proc. ICSLP-94, Yokohama 1994.

[4] L.B. Larsen, “Development and evaluation of a spoken dialogue for a telephone based transaction system”, in proc. EUROSPEECH-95, Madrid 1995.

[5] A. Baekgaard, “The Generic Dialogue System Platform”, Report 10, The Danish Dialogue Project. CPK, Aalborg University, Aalborg 1996.

“Combining Objective and Subjective Data in Evaluation of Spoken Dialogues“, in *Proceedings of the ESCA ETRW on Interactive Dialogue Systems,* Cloister Irsee, Germany, 1999.

Combining Objective and Subjective Data in Evaluation of Spoken Dialogues

Lars Bo Larsen
Center for
PersonKommunikation,
Aalborg University
Fredrik Bajers Vej, 7-A6,
Aalborg, DK-9220, Denmark
Email: lbl@cpk.auc.dk, Web:
http://www.cpk.auc.dk/~lbl/

Abstract

Evaluation of human-computer spoken dialogues is often based on analyses of objective metrics such as task completion rate, turn-taking, time consumption, etc. While these data are easily obtained and processed from log files, they offer or very little information about the actual usability of the given service as perceived by the test subjects.

This study combines the evaluation based on objective data obtained from log files with subjective data, where the test subjects express their attitudes to a number issues directly related to the usability of the service. It is shown how the joint analysis can be used to support, but also question findings from either source.

1. Introduction

The study is based on a field trial of a home banking service. The trial was carried out within the Esprit OVID[1] project and involved a total of 320 users (see [1],[2],[3]). The analysis reported here is based on the test subjects’ response questionnaires and the corresponding transcribed dialogues. Each user was instructed to make two calls to the service from a time and place of their choice. Immediately after the calls they were required to express their attitudes towards different aspects of the interaction by responding to a number of statements. The responses were quantified and analysed together with logging information of task and sub task completion rates, time per task, turn-taking, speech recognition accuracy and user-initiatives.

[1] The OVID Esprit 20717 Project consortium comprises The Royal Bank of Scotland and Barclays Bank in the U.K., Lån & Spar Bank in Denmark, CCIR Edinburgh University, U.K, CPK, Aalborg University, Denmark, Brite Voice Technology U.K., and AGORA Consult, France as coordinating partner. The work presented here is partly funded by the ESPRIT programme, and partly by CPK

1.1. Background

The experiments reported here were carried out within the Danish part of the OVID project and addresses the domain of phone based home banking. This task is well defined and as such represents a broad class of well-structured tasks, which are suitable for voice controlled automation. Examples of these are: credit card information services, telephone ordering services (of e.g. travel catalogues), ticket ordering systems, book clubs, and transaction systems in general. Common to these services are that a certain degree of structure can be imposed upon the discourse model without compromising the naturalness of the dialogue. As a consequence, the task structure and the linguistic phenomena exhibited by the users tend not to become too complicated.

1.2. The Present Task

The overall goal of the OVID project was to measure user acceptance of voice controlled home banking systems [1]. This is achieved by setting up trial applications and carrying out usability tests in field tests with bank customers.

Among other things, the OVID banking partners required, that the customer must be in control of the interaction [2]. However, this may not always lead to the most natural or efficient mode of communication, as humans often expect the counterpart to hold or take the initiative in conversations.

Therefore, a mixed-initiative strategy is implemented. Consequently, the purposes of the dialogue experiments reported here can be formulated as:

- To develop and test a dialogue management strategy in accordance with the specifications.
- To measure the degree to which this has been achieved
- To measure the user attitudes in general towards the service.

2. Methodology

Two distinct sets of information were collected within the experiment. These were then merged for a joint analysis.

2.1. Objective Measurements

Objective measurements were collected from the logfiles of each dialogue. These include time stamped information of all events during the dialogue, combined with transcriptions of the user utterances. From this, the following parameters were derived:

- Number of turns within each subtask
- Time spent in each subtask

- Number of user initiatives
- Number of speech misrecognitions

2.2. Subjective Measurements

Subjective information was obtained by asking each test subject to fill out a questionnaire immediately after performing the two scenarios. The questionnaire had the form of a set of Likert statements [4], [5] to which the subject expressed her degree of agreement/disagreement. This methodology was developed at CCIR at Edinburgh University. The statements covered five general usability factors:

- Quality of interface/performance,
- Cognitive effort/stress,
- The conversational model,
- Fluency and transparency of service.
- Transparency

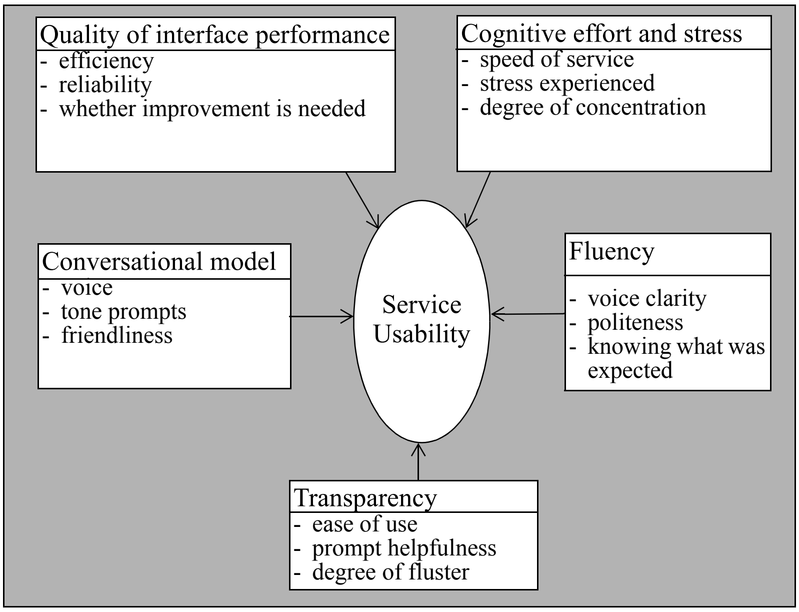


Figure 1. Service Usability Factors (from [5])

These comprised a *core set* of 22 statements, to which 4 application specific statements were added. The factors were identified and ranked in [5]. Figure 1 below illustrates this.

The test subjects express their attitude to each statement by marking one of seven boxes ranging from “strongly disagrees” over “neutral” to strongly agrees”. The marks are translated into a scale from one to seven (with four as the neutral point) [4]. By using this translation, the users attitudes can be quantified and subjected to statistical analysis.

2.3. Combined Analysis

The objective and subjective information can now be combined, as each test subject were issued unique Id and PIN codes for identification, which were used to link the logging information to the questionnaires. This makes it possible to subdivide the test subjects according to e.g. the number of speech recognition errors they experienced and analyse their attitude towards the service as a function of this parameter.

3. Results

This section presents the results of the field trial. As mentioned in section 1, one of the goals is to evaluate whether the dialogue model is acceptable to the users. In order to verify this, the turn-taking was analysed in detail. The other goal was to evaluate the users attitudes towards different usability aspects.

3.1. Subjective measurements

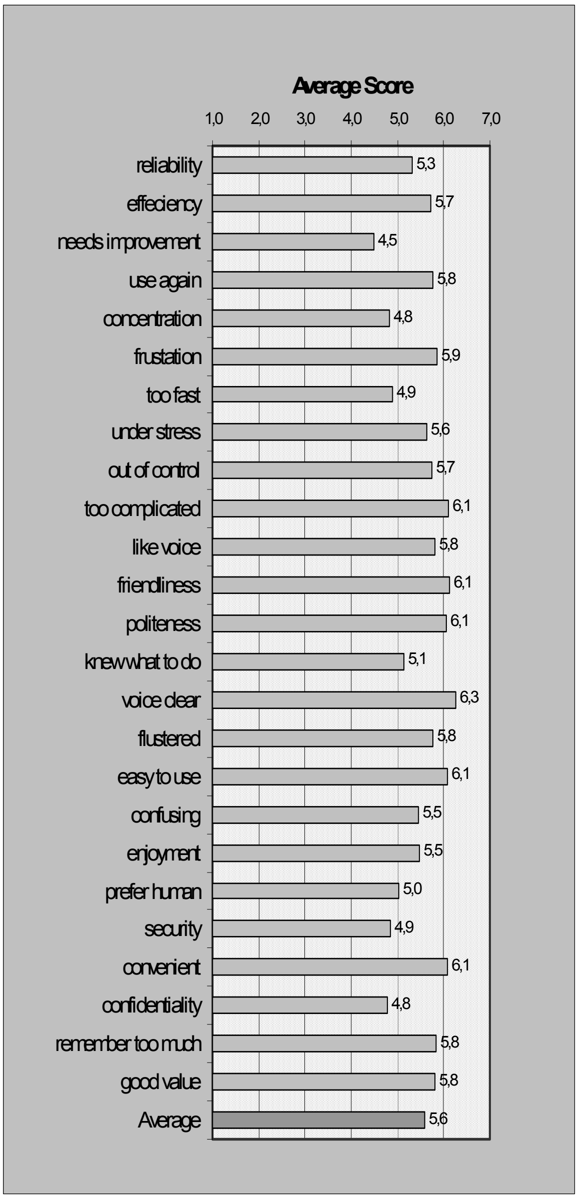


Figure 2 User Questionnaire responses

As described in the previous section, the questionnaires were quantified for statistical analysis. Figure 2 shows the overall averages for the 25 statements (and the combined average). The statements are ordered according to the categories discussed above. Taken as a whole, the results show that the users generally have a positive attitude towards the service. The overall average is 5.6. Note that some of the statements are negated, e.g. “confusing”. Thus, a high value for a negative statement indicates disagreement, and consequently a high value in the chart will always be interpreted to indicate a positive attitude towards the service, making the reading of the figure easier. Negative statements are necessary in order to balance the questionnaire.

The users were invited to express their comments to the service, and almost all did so. The most common impression was that an account mini statement was quoted to fast to note down. This is also evident from Figure 2, where attitudes towards the statements about concentration and speed are below average.

Users were not informed about an optional DTMF input modality. Only in case of repeated recognition errors did the system suggest that they use the telephone keypad. This decision was made in order not to bias users towards a specific input modality, and because DTMF was purely intended as a backup option. Consequently, many users expressed concern that their Id and Access numbers might be overheard. This concern has also influenced the attitude towards confidentiality, as shown in Figure 2.

The user population was broken down into even-sized subgroups with respect to:

- Age groups (18-29, 30-39, 40-49, 50-59, above 60)
- Region (three major regions in Denmark + Copenhagen).
- Male / Female

The users were not asked whether they had any prior experiences with ASR systems, as no such services existed in Denmark at the time. However, all subjects had long-standing experiences with automated home-banking.

T-tests were employed to uncover any significant differences, but at a confidence level of 5% none were found. This indicates that no difficulties or diverging attitudes towards the service could be ascribed a specific subgoup (e.g. age) of the users.

3.2. Analysis of the Dialogue Model

In order to determine whether the users experienced a degree of “being in control” in the dialogue, the turn-taking was analysed. In particular, the number of user initiatives, i.e. points in the dialogue, where control passed to the user, was investigated. The results are shown in Table 1 below. As mentioned above, each user was asked to complete two different scenarios, denoted A and B. In order to avoid bias, one half of the users were asked to carry out scenario A as the initial call and then proceeding with scenario B. The other half completed B first, then A. In Table 1 A1/B1 denotes scenario A/B as the users’ first task, and A2/B2 the second.

Scenario:	**A1**	**A2**	**B1**	**B2**
Nominal number of turns	7		9	
Minimal[a] number of turns	5		4	
Average number of turns	8.4	7.8	7.5	7.1
Avg. duration of dialogues (seconds)	94	86	92	84
Nominal number of user initiatives[b]	1		2	
Average number of user initiatives	0.8	1.0	1.3	1.6

[a] This includes the user taking the dialogue initiative whenever possible, and hanging up immediately after the desired information has been obtained, which very few of the users did. This can be expected of more experienced users, though.

[b] User hang-up is not counted as a user initiative

Table 1 Key figures for turn taking and user initiatives

Two conclusions can be drawn from the results. A tendency towards shorter (smoother) dialogues for experienced users can be deduced, as the average number of turns and task duration drop for both scenarios. This is taken as a sign of the user becoming familiar with the service.

The proportion of the user initiated turns are increasing for both scenarios. This is interpreted as the user, when fully familiarised with the service, can be expected to be able to gain the initiative whenever he feels it natural or convenient to do so.

This is reflected in Figure 2 where the statement referring to the degree of control scores averages at 5.7.

3.3. Analysis of the Speech Recogniser Performance

One of the crucial factors for the success of a spoken dialogue system is the performance of the speech recognition engine. Therefore much effort goes into ensuring a high level of performance.

However, it might be very costly or time-consuming to aim at a perfect level of performance without taking into account the actual influence on the overall usability of the service. Ignoring this might lead to efforts better spent on improving other parts of the system.

This is illustrated in Figure 3 below, where the average user attitude is shown as a function of the experienced speech recognition accuracy during the experiment. Note that very few users (less than 25) experienced a recognition performance below 70%. 90% accuracy roughly corresponds to the user experiencing a total of one error in the two dialogue scenarios. It can be observed that there seems to be no significant decrease in user attitude from perfect recognition down to a level of approximately 70%. However, other studies [6] report from an WOZ experiment that a decrease in user attitude was found at 90%.

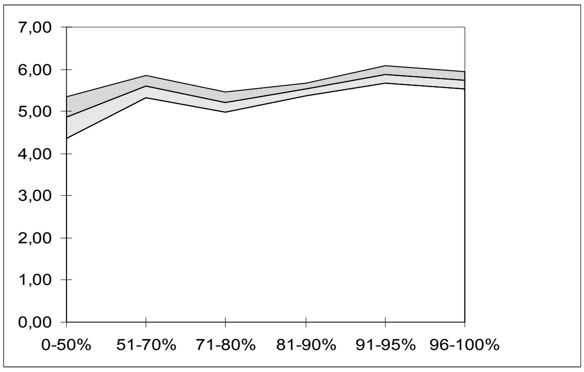


Figure 3 Average user attitude as a function of the experienced speech recognition accuracy. The band denotes the 5% significance level

This suggests that the impact of the speech recognition performance is dependent on the actual application. One plausible explanation is that other factors, such as the (re)formulating of prompts, error detection and -recovery strategies might in fact be as important for the overall usability as the "raw" speech recognition performance. The present dialogue was designed using implicit confirmation, which made errors immediately obvious to the users. Thereby, an error typically only had a "cost" of one additional dialogue turn. As mentioned, this seemed to be acceptable to most users, although the application domain (homebanking) could be expected to be particularly sensitive to errors.

The average speech recognition accuracy was 89%, but this figure covers a large diversity between individual users. 30% experienced 0 or 1 error (in two dialogues), 50% from 2 to 5 errors, 16% from 6 to 9, and 3% more than 10 errors. Even though the service in general performed at an acceptable level, improvements are certainly needed for the small percentage of users for which it didn't work at all. Inspections showed that many of these users had strong accents (e.g. Norwegian or Swedish), or spoke with a very soft voice. However, for other users no obvious reason could be detected.

4. Conclusions

As stated in the introduction, the overall aim of the experiments was to investigate to what extent customers are prepared to accept voice controlled access to their bank accounts. The results presented here indicate that this will be the case. The tested dialogue was very small, containing only a few sub tasks, so the next step will evidently be to expand the dialogue to cover a larger number of tasks, and a more complex task structure.

The key requirements for the service were formulated by the OVID banks and stated that:

"The user must feel in control" and

"The user must be able to speak naturally"

This was achieved by using multiple keyword spotting within the speech recogniser, thus allowing the user to speak naturally, while keeping the linguistic complexity low. The dialogue model supported this, and the mixed-initiative dialogue model proved to successfully anticipate the user behaviour, allowing the user to control the interaction, while retaining guidance for novice users. A tendency towards a greater proportion of user control for experienced users were found.

Regarding the combination of the information obtained from logging of the dialogues with the subjective user attitude questionnaire. A number of interesting conclusions emerged, in particular with respect to the influence of the speech recogniser performance. It was seen that there seems to be little influence on the overall attitude towards the service. However, users for whom the recogniser performed extremely poorly, it is hard to put credibility into their responses. Consequently, the quantitative data can also be used to validate the user responses.

5. References

[1] ESPRIT 20171 Project OVID - Trial application of Voice Processing in Automated Telephone Banking Services: "Technical Annex", July 1995, EU commission.

[2] Lars Bo Larsen, "Voice controlled home banking - objectives and experiences of the Esprit OVID project" in Proc. of IVTTA 1996 IEEE third workshop on Interactive Voice Technology for Telecommunications Applications, New Jersey, September 1996. ISBN 0-7803-3238-5.

[3] Lars Bo Larsen "Investigating a Mixed-Initiative Dialogue Management Strategy" pp. 65-71, in Proc. of the 1997 IEEE workshop on Automatic Speech Recognition and Understanding (ASRU), Santa Barbara, California, December 1997. ISBN 0-7803-3698-4

[4] Trochim, W. (1999). "The Research Methods Knowledge Base, 2nd Edition." Cornell Custom Publishing, Cornell University, Ithaca, New York. Also available on-line at: http://trochim.human.cornell.edu/kb/

[5] Love, S., Dutton, R.T., Jack, M. A., Stentiford, F. W. M.: Identifying Salient

Usability Attributes for Automated Telephone Services”. In proc of ICSLP’94, pp 1307-1311, Yokohama 1994.

[6] Foster, J.C. et al. “Intelligent Dialogues in Automated Telephone Services”, in Baber, C., And Noyes J.M (eds) “Interactive Speech Technology: Human factors Issues in the application of Speech Input/Output to Computers”, Taylor & Francis, U.K. 1993

"Assessment of Spoken Dialogue System Usability - What are We really Measuring?" To appear in the proceedings of Eurospeech'03, Geneva Switzerland, September 2003.

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 behaviour, 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.

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 consid-

[1] Spoken Language Dialogue System

ered 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.

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.

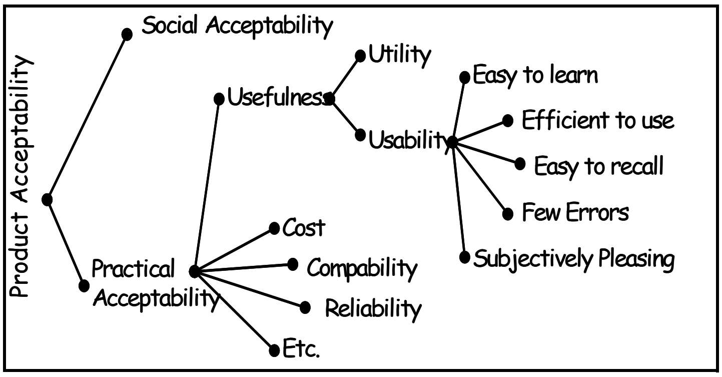


Figure 1. Jakob Nielsen's definition of usability (redrawn from [6], p.25)

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	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	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	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	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 identified subscales were labels are shown below in Table 2:

Sub Scales	**Alpha**	**Var**[a]
Quality of interface/efficiency/reliability	0.86	40%
Cognitive effort/stress	0.83	11%
Transparency/confusion	0.78	9%
Friendliness	0.82	6%
Voice	0.92	5%

Table 2 Identified Subscales from the OVID experiment

[a] The proportion of the explained variance of the PCA components

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)

Evaluation of Spoken Dialogue Systems using Objective and Subjective Measures

Lars Bo Larsen

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

Abstract

This paper presents results and conclusions about the current evaluation methodologies for Spoken DIalogue Systems (SDS). The PARADISE paradigm, used for evaluation in the DARPA Communicator project is briefly introduced and discussed through the application to the OVID home banking dialogue system. It is shown to provide results consistent with those obtained by the DARPA community, but a number of problems and limitations are pointed out.

The issue of user attitude measures through questionnaires is discussed. This is an area that have not received much attention from the speech technology community, but is important in order to obtain valid results and conclusions about usability.

1. Introduction

This paper investigates the reasons why speech based interfaces still - despite many predictions of "near-future breakthroughs" and substantial technological advancements - have not yet achieved this status. 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 Section [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* Section [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). One possible explanation is that the **usability** of speech based systems have not yet reached a sufficient level to be acceptable to the general public.

The aim of this paper is to analyse how the usability of speech systems currently is evaluated in order to set focus on the applied methods' strengths and weaknesses. In particular, the PARADISE scheme, proposed by Walker and colleagues from AT&T Section [2] will be in focus. PARADISE has been used in a number of evaluations, e.g. the recent DARPA Communicator project Section [4] 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. PARADISE is described and discussed in Section 3 below and illustrated by the application of the method to the OVID home banking corpus [5][6][7].

One issue that has been largely neglected by the speech research community is the methods for elicitation of the user's attitudes [6],[7],[8]. Many researchers put a set of Likert-like statements together, addressing topics of interest and collect the user's responses. However, this does not in any way ensure that the outcome is a 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 drawn from it is on very thin ice.

Hone and Graham review a number of such 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"* [8].

The issue of ensuring valid user attitudes for evaluation of speech based systems is the second focus point of this paper. However, before addressing these two issues, a brief description of the OVID experiments is provided. This is followed by a discussion of the methods for questionnaire design and validation. The PARADISE scheme is briefly introduced and illustrated by applying it to the OVID corpus. Finally some conclusions are drawn up and discussed.

2. The OVID Home banking Application

The OVID project addresses 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. ([5],[6],[7],[8],[9]).

The Danish OVID dialogue corpus comprises 700 transcribed and annotated dialogues by more than 300 users calling the system. Each user returned a questionnaire with information about their attitudes towards the system. Performance data for dialogue and task turns, -duration, task completion rates, user initiatives, etc. were collected. However, the combined analysis of the subjective and objective data was not originally performed.

PARADISE was created for this purpose, and hence it would be of interest to investigate if additional new information can be extracted from the corpus by applying PARADISE. It is important to note that the OVID experiments were not planned or carried out with Paradise in mind. Therefore, it is also of interest to examine whether it is possible to apply the scheme, and especially whether for example a total re-annotation of the corpus is necessary.

The Paradise paradigm is well-known and has been published elsewhere, so the following introduction is kept very brief.

3. The Paradise Evaluation Scheme

Originally, Paradise was conceived to enable comparison across different tasks and dialogue management strategies. It has been reported in numerous occasions, but perhaps the most comprehensive description is given in [2], on which this introduction is based. The basic principle is to apply Multiple Linear Regression (MLR) to maximise user satisfaction while minimising costs and maximising task success. The dialogue "costs" are divided into measures of efficiency (e.g. system turns, elapsed time) and qualitative measures (e.g. task completion, barge-ins and help requests). However, as the authors note, the model is general and does not require some particular measures. The structure of the Paradise model is shown in Figure 1.

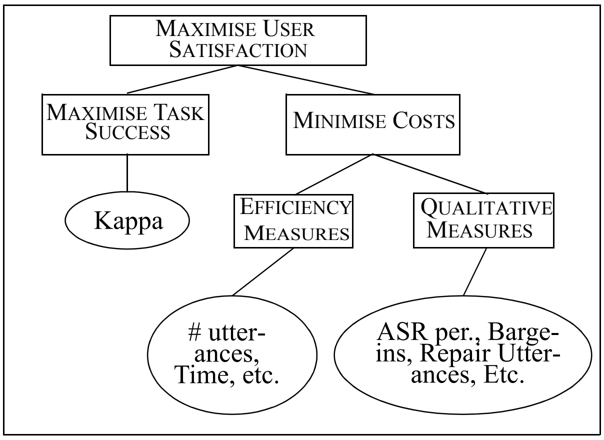


Figure 1. Paradise Structure [2]

In order to compare across different domains and tasks, it is desirable to compensate for task complexity. This is done by representing task success by the Kappa statistic [2],[15]. Kappa will compensate the actual achieved task success with the probability of obtaining the correct information "by chance".

Another important concept in the Paradise model is the Attribute-Value Matrix (AVM), which is used to measure the degree of task success. Since the model attempts to decouple how some piece of information was obtained (i.e. the dialogue agent) from the information, only the information itself is of interest when considering task success. This information is contained in the AVM, which in turn is used to compute Kappa.

An example of an AVM for the OVID home banking application is shown in Section Figure 2.

Attribute	**Value**
Id-number	**973625**
Account	**Cash Credit**
Action	**Statement**

Figure 2. Example of OVID AVM

User satisfaction is represented by the accumulated sum of the users' attitudes towards a number of statements in a post-test questionnaire. These covered ASR- and TTS performance, task ease, expected behaviour, etc. An important issue is the question of the user's perception of task success. Clearly, the perceived task success may have an impact on the users' attitude towards the system and be reflected in the questionnaire.

Therefore the user is asked whether s/he completed the assigned tasks, and this information is used as an qualitative measure in the model.

3.1. Results from applying PARADISE to the OVID corpus

In short, the observed or derived performance measures, which can potentially be used in a Paradise evaluation of the OVID experiments are:

Objective (performance) measures:

- Number of turns (overall and for each sub task)
- Proportion of user-initiated turns
- Number of repair turns in access sub tasks
- Time spent (overall and for each sub task)
- ASR (speech concept) rates
- Overall and subtask success rates (expressed as the Kappa coefficient), and a more traditional ratio between desired and achieved (sub)goals.

Subjective (user attitude) measures:

- Average user satisfaction
- Perceived task success

However, since virtually all (96%) of the users reported to successfully having completed the required scenarios, this last measure does not provide any information and is not used in the further analysis.

The OVID field trial was scenario-driven and as the OVID dialogue model is fairly simple and well-structured, AVMs for the scenarios can be formulated and calculated for each dialogue, see Figure 1. A closer description of the measures can be found in e.g. [7].

The first step is to investigate the correlation between the parameters to get an impression of the relationship between the variables. This resulting covariance matrix, for a subset of 35 users (and 105 dialogues) is shown in Table 1 below.

Covariance Matrix	Kappa	Task Success	**ASR**	Total Turns	Total Time	**User Satisfaction**
Kappa	1.0					
Task Success	0.6	1.0				
REC	0.2	0.4	1.0			
Total Turns	0.1	0.0	-0.4	1.0		
Total Time	0.0	-0.1	-0.6	0.9	1.0	
User Satisfaction	**0.5**	**0.3**	**0.6**	**-0.3**	**-0.4**	**1.0**

Table 1 . Covariance Matrix for selected dialogue measures. REC is the speech concept recognition score. Due to a Z-normalisation of the variables, the values in the diagonal (the variances) all equal 1 and the (absolute) off-diagonal values lie between 0 and 1.

From Table 1 it can be observed that “Total time” and “Total Turns” are heavily correlated (0.9), which is expected. It can also be seen that none of the other parameters are highly correlated. “Task Success” is a more conventional measure for task completion, roughly equivalent to the proportion of achieved (sub)goals. As expected, it correlates with kappa. For comparison, “User Satisfaction” is also included in the table and all parameters show some correlation with it. As described in [7], user satisfaction is calculated as the averaged score for each user for the 20 statements used in the usability questionnaire used in the OVID experiment.

Turns and Time correlate negatively, whereas the task success and SR measures correlate positively with the user satisfaction parameter. This means that the longer a dialogue takes in terms if turns or time, the less satisfied the users are. A positive correlation between the task completion rates and recognition accuracy implies that users get more satisfied the better the performance is.

Having established that a correspondence between user satisfaction and the performance parameters indeed exists, the next step is to perform a MLR to derived the exact coefficients. However, it turns out that only Kappa and Rec. are significant predictors of usability. Table 2 below shows a comparison between the resulting performance functions for OVID and a number of SDS reported by AT&T [3]

SDS (Domain)	Performance Function	Var
OVID (Home banking)	Perf = 0.41*κ + 0.47*Rec	51%
TOOT$_1$(Train travel)[a]	Perf = 0.45*Comp + 0.35*Rec -0.42*B.I	47%
TOOT$_2$$^{(}$Train travel)	Perf = 0.33*Comp + 0.45*Rec -0.14*Time	55%
Annie (Voice Dialling)	Perf = 0.25*Comp + 0.33*Rec -0.33*Helps	41%
ELVIS (Email access)	Perf = 0.21*Comp + 0.47*Rec -0.15*Time	38%

Table 2 Comparison of results from OVID and three SDS from AT&T ([3]). Perf is the average usability score, Rec the speech recognition score, Comp is the perceived task completion rate, B.I. is Barge-Ins, Time is the duration of the dialogue and Helps is the number of help requests. Var is the proportion of variance explained by the model

[a] The two Toot systems address the same domain, but employ different dialogue management strategies

Several interesting observations can be made from the table. As mentioned above, only kappa and Rec turned out to be statistically significant predictors of user satisfaction (Perf) for the OVID system. Although “number of turns” and “elapsed time” (Time) also correlate with Perf (see Table 1) they are not significant predictors, and a model including one of these measures does not produce a better fit of the independent variable Perf.

Comparing the results from OVID with similar PARADISE analyses of the three SDS built by AT&T researchers and reported in (Kamm et al 1999), a notable correspondence between the findings is found.

Except for one case, speech recognition is found to be the most important contributor, which is to be expected. The influence of speech recognition performance for OVID is identical to those found for Toot$_2$ and Elvis and close to the ones found for Toot$_1$ and Annie. Although the AT&T experiments applied the users’ perceived task success (Comp) instead of kappa, close to identical results are found for Toot$_1$ and Toot$_2$. The AT&T experiments also found significant predictors, although of lesser importance, for Help, Barge-Ins and elapsed time. Except for elapsed time, these measures were not available for the OVID corpus.

A comparison of the ‘goodness of fit’ of the estimate (expressed as the percentage of the variance explained by the model) also shows very similar results, (between 38% and 55%) with 51% explained variance for the OVID SDS. However, the confidence intervals are quite large, as can also be seen in Figure 3 below.

.

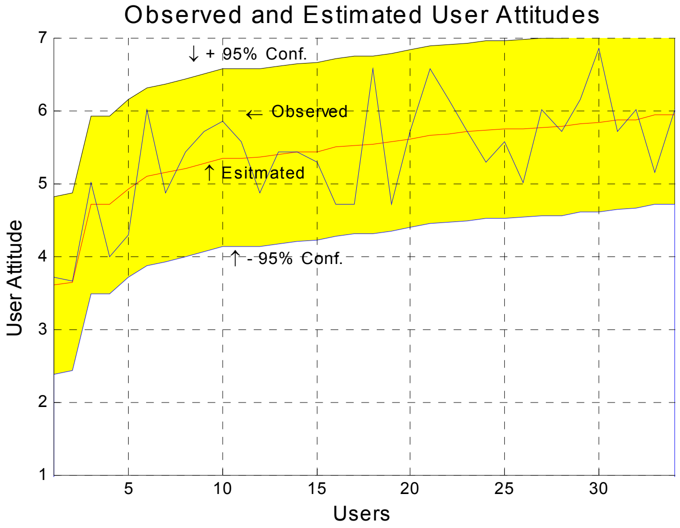


Figure 3 Observed and estimated user attitudes, using PARADISE. The red line represents the estimated user attitudes, and the blue line the observed values.

Figure 3 shows a plot of the values predicted by the model together with the recorded values and 95% confidence band. Although the observed values generally fall well within the confidence band of the estimated (one outlier was identified and removed before the regression was carried out) of the predicted ones, it is obvious that a large proportion of the variability has not been captured by the model.

3.2. Implications of the experiment

As shown above, it is possible to apply the Paradise scheme to a subset of the OVID corpus and obtain results comparable to those published by Walker and colleagues. The important question is of course whether it revealed any new information about the corpus. It is hardly surprising that a relationship between ASR performance and user satisfaction can be observed. This is a well-known fact and has published numerous times, also for the OVID corpus, although only a weak one [5]. Obviously there are other important factors influencing user satisfaction.

Interestingly, kappa proved to be a better predictor than a more traditional measure for task completion based on a simple ratio between desired and obtained goals. The main function of kappa is to normalise for task complexity. This indicates that the two scenarios in the OVID experiment turned out to differ in complexity and that kappa captured this fact.

The estimated model only explains approximately half of the observed variability and the coefficients are estimated with some uncertainty. One likely reason for this is that the users unfortunately did not have a very precise perception of their actual task completion, as mentioned above. This will of course be reflected in an unclear or muddled relationship between their attitudes towards the system (as expressed in the questionnaire) and the hard facts (obtained from the logfiles).

An interesting implication of the results shown in Table 2 concerns the questionnaires used to obtain the user attitudes. The AT&T user attitude questionnaire only comprise 6-9 statements (on a five-point Likert scale, see [2],[3]). In contrast the OVID questionnaire consists of 20 statements (on a seven-point scale), quite different from those of AT&T. Regardless, PARADISE produces quite similar results both regarding the combination of measures and the fit of the model. Two explanations come to mind: Either PARADISE is not particularly sensitive to the user attitude elicitation questionnaire, or both questionnaires essentially capture identical measures from the users. If the latter is correct, the PARADISE analysis can be regarded as a supplementary proof of validation of the questionnaires.

3.3. Discussion

Although the application of the PARADISE scheme on the OVID corpus worked out quite well, there are two matters that are problematic in a wider perspective: One concerns the requirement for specific scenarios with clearly defined goals and has been briefly touched above. This poses a serious threat to the generality and scalability of PARADISE to e.g. multi modal systems. The other concerns the calculation of some of the parameters. A number of assumptions are made, e.g. about linear relationships between performance and subjective measures. There are really no hard evidence that this is the case. As mentioned previously, many studies have shown a relationship between speech recognition performance and user attitudes. Indeed, such curves are seldom straight lines, and sometime even have a “threshold”, where the slope changes abruptly.

Furthermore, the parameters, most notably the AVM and κ measure used to represent task success cause problems. Unless the test scenarios are very structured and well-defined the definition of these become ambiguous, as indeed a some experiments have revealed, e.g. [10] and in an adaption of PARADISE for evaluation of multi modal systems in the German SmartKom project [11]. Some of the problems can be solved by a more specialised and dynamic generation of AVMs, but this will in turn reduce the value of PARADISE as a comparative tool across tasks and domains.

However, when applications of similar complexity and outlook are to be evaluated, as in the case of the DARPA Communicator project, or when e.g. successive versions of a system is tested, PARADISE is a powerful tool. Furthermore, the existence of a widely used and (although with limitations, as discussed above) standardised evaluation paradigm can only be a positive element. At least it will stimulate research and development of other, perhaps better paradigms. The efforts by the SmartKom project [11] to extend and modify PARADISE to multi modal dialogues is a good example of this.

4. Obtaining User Attitude Measures

As mentioned in the introduction, an often overlooked problem is that of obtaining the user’s attitude towards the system being evaluated. A reason for this could be that the techniques used for this belongs to the field of experimental psychology denoted psychometrics, which may be unknown to most speech technology scientists. Another likely reason is probably that it is a difficult and resource-demanding process to develop and validate a usability questionnaire.

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 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 [6],[7],[8]. 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, to the authors knowledge.

4.1. The CCIR questionnaire

One exception to this is the questionnaire developed at CCIR at Edinburgh University through an iterative process of testing, reformulation and validation [12]. This was used in the OVID project. It has the form of a set 20 of Likert statements [13], to which the subject expressed his/her attitude on a seven-point scale (from "Strongly Agree" to "Strongly Disagree"). While the validity of a scale is a complicated and time-consuming process to establish, the reliability can be estimated by checking the internal consistency of the answers. The most common test is Cronbachs Alpha [13]. For the OVID test, Cronbachs Alpha is 0.92 (for the raw items). This is quite satisfactory and indicates that there is a high degree of consistency between the items in the scale.

4.2. Factor Analysis

One way to investigate the degree to which a questionnaire is valid is to look closely at the underlying relationships between the individual statements. This will reveal if any inconsistencies, redundancies and ambiguities of the statements exists. A commonly used technique for this is Factor Analysis (FA) [13],[14], which will uncover the common factors among the statements. Table 3 below shows the FA for the OVID questionnaire.

Factor	Label / Statements	Var.%
F_1	**Quality of Interface, Performance**	10.7
	Efficiency, Ease of Use, Frustration, Need of Improvement, Reliability	
F_2	**Control/Confusion**	10.4
	Out of Control, Too Complicated, Flustered, Remember Too Much, Knew What To Do	
F_3	**Convenience**	9.7
	Use Again, Good Value, Convenient, Enjoyment, Preference for Human	
F_4	**Personality**	9.3
	Friendliness, Like Voice, Politeness, Voice Clear	
F_5	**Confidence**	8.1
	Security, Confidentiality, Reliability	
F_6	**Cognitive Load**	7.5
	Under Stress, Too Fast, Concentration	

Table 3 Six-Factor structure with all statements included. The total explained variance is 56%.

Six common factors has been identified and labelled according to the statements loading on them (loading is the term for correlation, when doing FA). The column at the right shows how much of the statement variance each Factor represents. More Factors could have been included, but would have generated a less interpretable structure. FA has a close resemblance to Principal Components analysis (PCA), but as indicated above there is an element of interpretation in FA, where PCA is purely data-driven. For comparison, a PCA carried out on the OVID data with 5 components captured 68% of the variance, but the resulting components can not be interpreted in an analytical way.

The FA uncovered only one statement, (the degree of confusion felt by the users), which did not turn out to fit any of the clusters. This is a strong indication that the questionnaire is consistent and reliable.

5. Discussion

The OVID corpus did – with some extra effort – provide the necessary information needed to apply Paradise: Both quantitative (log/performance) and qualitative (user attitude measurements) information were available. The dialogue tasks are goal-directed, hence an AVM could be formulated and the Kappa statistic computed for each dialogue. The Kappa statistic was found to be a superior predictor of user satisfaction when compared to a more traditionally derived task success measure.

The important question in this study is: “Did the Paradise analysis uncover significant, new information in the corpus?” The answer to this is that Paradise did not produce any unexpected insights as such. It can hardly be surprising that task success and speech recognition performance are decisive factors for the users preferences. However, it was interesting in itself to see the high degree of correspondence between the results obtained by OVID corpus and the AT&T results. Especially considering the very different measurements of the user attitudes.

The discussion about the usability questionnaire revealed that this is an area where more work needs to be done. The questionnaire used in the OVID experiments were analysed and shown to have a high internal consistency.

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] 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.

[3] Candace Kamm, Marilyn A. Walker, and Diane Litman. “Evaluating Spoken Language Systems” In Proceedings of American Voice Input/Output Society, AVIOS, 1999.

[4] The DARPA Communicator Evaluation Committee: http://

www.research.att.com/~walker/eval/

[5] 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

[6] Lars Bo Larsen: “Assessment of Spoken Dialogue System Usability - What are We really Measuring?” Proc. of Eurospeech’03, Geneva Switzerland, September 2003

[7] Lars Bo Larsen: “On the Evaluation of Spoken Dialogue Systems” Ph.D. Thesis, Aalborg University, July 2003.

[8] 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.

[9] Keith Edwards, K. Quinn, P.B. Dalziel, M. A. Jack: “Evaluating Commercial Speech Recognition and DTMF Technology for Automated Telephone Banking Services”. in IEE Colloquium on Advances in Interactive Voice Technologies for Telecommunication Services. 1997.

[10]Any Hjalmarsson: “Evaluating AdApt, a multi-modal conversational, dialogue system using PARADISE”. Masters thesis in Speech Technology, KTH, Stockholm November 2002.

[11]Nicole Beringer, U. Kartal, K. Louka, F. Schiel, U. Türk: “PROMISE - A Procedure for Multimodal Interactive System Evaluation”, in Proc. LREC Workshop on Multimodal Resource and Multimodal Systems Evaluation, 2002

[12]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

[13]A.N. Oppenheim “Questionnaire Design, Interviewing and Attitude Measurement”. New Edition. Continuum, London, 1992 (1966).ISBN 0-8264-5176 4

[14]Barbara Tabachnick and L. Fidell: “Using Multivariate Statistics”, 4th. Edition, Allyn and Bacon, 1996 and 2001. Printed in MA, U.S. ISBN 0-321-05677-9.

[15]Sidney Siegel, N.John Castellan: “Non parametric Statistics for the Behavioural Sciences” (second edition). New York, 1988, 1956, McGraw-Hill. ISBN 0-07-057357-3

References

This list contains all the literature referenced in the present work, except from those referred in the articles in Appendix C, which each includes a reference list.

The format of the references are:

Author(s), year
Author names, "Title" in reference year

- where "(Author(s), year)" is the text that appears when referring to it.

It is followed by the full reference text. All references are listed alphabetically after first author. Otherwise identical references are denoted e.g." Larsen 1997b", "Larsen 1997b" etc. Web links have been provided whenever applicable. A note is made by the links to indicated when it was last visited. If no date is present, June 2003 is assumed. If a reference only refers to a link, the date indicated on the website is used as the publication year (e.g. "last updated"). If no date is supplied, the present year is used instead. For example:

Darlington 1997
Richard B. Darlington: "Factor Analysis". January 1997. http://comp9.psych.cornell.edu/Darlington/factor.htm (last visited April 2003)

1. List of References

Aiken 1996
Lewis R. Aiken: "Rating Scales and Checklists - Evaluating Behaviours, Personality and Attitudes" John Wiley and Sons, New York 1996, ISBN 0-471--12787-6

Anderson et al 2002
David Anderson, D. Sweeney, T. Williams: "Statistics for Business and Economics", 8th Edition, U.S. 2002. Publish by: South-Western. ISBN 0-324-06672-4

Beringer et al 2002
Nicole Beringer, U. Kartal, K. Louka, F. Schiel, U. Türk: "PROMISE - A Procedure for Multimodal Interactive System Evaluation", in Proc. LREC Workshop on Multimodal Resource and Multimodal Systems Evaluation, 2002

Bernsen et al 1998
Niels Ole Bernsen, H. Dybkjær, L. Dybkjær: "Designing Interactive Speech Systems: From First Ideas to User Testing". Springer Verlag, Berlin 1998.

Bouwman and Hulstijn 1998
“Dialog Strategy (Re-)Design with Reliability Measures Information”. *First International Conference on Language Resources and Evaluation 1998* Also available as: http://lands.let.kun.nl/literature/bouwman.1998.1.ps

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

Carletta 1996
Jean Carletta. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics, 22(2):249-254,* June 1996.

CCIR 2003
CCIR “Intelligent dialogues project - Research Reports on automated telephony”
http://www.ccir.ed.ac.uk/doc/ccir_dialogues_reports.htm (visited April 2003)

CCIR 2003a
“Attitudes to voices in automated telephone services” - Research Reports on automated telephony. Report number 3:
http://spotlight.ccir.ed.ac.uk/public_documents/technology_reports/No.3%20Voices.pdf (visited April 2003)

CCIR 2003b
“The Effects of Speaker State (Tone of Voice) and Speaker Style (Fast Track) in Dialogue Prompts” - Research Reports on automated telephony. Report number 9:
http://spotlight.ccir.ed.ac.uk/public_documents/technology_reports/No.9%20Speakstate.pdf (visited April 2003)

Chin et al 1988
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

Dalsgaard and Bækgaard 1994
Paul Dalsgaard, A. Bækgaard: “Spoken Language Dialogue Systems”. pp. 178-191 in Proceedings in Artificial Intelligence, Volume “*Prospects and Perspective in Speech Technology*”, C. Freksa (ed.), Infix, München 1994.

Danieli and Gerbino 1995
M. Danieli and E. Gerbino “Metrics for evaluating dialogue strategies in a spoken language system”, in *Proc. of the 1995 AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation, pp 34-39,* 1995

Danmarks Statistik 2003
Nyt fra Danmarks Statistik: Emnegruppe: Serviceerhverv Befolkningens brug af internet 1. halvår 2003 Nr. 231 • 26. maj 2003. http://www.dst.dk/pukora/view/pdf.asp?id=4416 (June 2003)

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

DISC 2000
The website for the DISC1 and DISC2 projects. Last Revised: 11 March, 2001: http://www.disc2.dk. Visited March 2003

Dutton et al 1993
R.T. Dutton, M.A. Jack, F.W. Stentiford: “Identifying Usability Attributes of Automated Telephone Services”, In Proc of EUROSPEECH ‘93, Berlin 1993.

Dybkjær et al 1996
L. Dybkjær, N.O. Bernsen, H. Dybkjær: “Evaluation of spoken dialogues: user test with a simulated speech recogniser” (two volumes), Spoken language dialogue systems; reports 9a and 9b (February 1996) CPK - Center for PersonKommunikation, Aalborg University ISBN: 87-7349-301-5

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

Eckert et al 1995
Wieland Eckert, E. Nöth, H. Niemann, E. Schukat-Talamazzini: “Real Users Behave Weird - Experiences made when collecting large Human-Machine Copora”, in Proc. Esca ETWR on Spoken Dialogue Systems - Theories and Applications”. Vigsø, Denmark 1995.

Edwards et al 1997
K. Edwards, K. Quinn, P.B. Dalziel, M. A. Jack: “Evaluating Commercial Speech Recognition and DTMF Technology for Automated Telephone Banking Services”. in IEE Colloquium on Advances in Interactive Voice Technologies for Telecommunication Services. 1997.

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

Gibbon et al 1997
Dafydd Gibbon, Roger Moore, Richard Winski (eds.): “Handbook of Standards for Spoken Language Systems”, Mouton de Gruyter, Berlin 1997. ISBN 3-11-015366-1

Grice 1975
H.P. Grice: “Logic and Conversation”. In “*Syntax and Semantics, Vol. 3, Speech Acts”, pp. 41-58.* Academic Press, NY, 1975.

Hjalmarsson 2002
Anna Hjalmarsson: “Evaluating AdApt, a multi-modal conversational, dialogue system using PARADISE”. Masters thesis in Speech Technology, KTH, Stockholm November 2002.

Hone et al 1998
K. S. Hone, R. Graham, M. C. Maguire, C. Baber, G. I. Johnson: “Speech technology for automatic teller machines: an investigation of user attitude and performance”. In Ergonomics, 1998, Vol. 41 No. 7 pp 962-981. Taylor-Francis 1998

Hone and Graham 2000
K. S. Hone, R. Graham. Towards a tool for the subjective assessment of speech system interfaces (SASSI). In Natural Language Engineering, 6(3/

4), 287-305. (2000) Also: http://www.brunel.ac.uk/~csstksh/ Hone_and_Graham_NLE_2000.ps (March 2003)

Hone and Graham 2001
K. S. Hone, R. Graham. "Subjective assessment of speech-system interface usability". Proceedings of Eurospeech 2001, 7th European Conference on Speech Communication and Technology, Aalborg, Denmark, 3-7 September, 2001, Volume 3, 2083-2086.

ISO 1998a
International Standardisation Organisation (ISO): "ISO 9241: Ergonomic requirements for office work with visual display terminals (VDTs) -- Part 10" http://www.iso.org

ISO 1998b
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

Jack et al 1993
Mervyn A. Jack, J. C. Foster, F.W. Stentiford "Usability Analysis of Intelligent Dialogues for Telephone Services" Proc of *ESCA/NATO RSGIO workshop on applications of Speech Technology*, Irsee Italy, Sep. 1993

Jurafsky and Martin 2000
Daniel Jurafsky and "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition" Prentice-Hall, 2000. ISBN: 0-13-095069-6

Kamm and Walker 1997
Kamm, C. and Walker, M. A., Design and evaluation of spoken dialog systems. *Proc. 1997 IEEE Workshop on Speech Recognition and Understanding, (ASRU)* 1997.

Kamm et al 1999
Candace Kamm, Marilyn A. Walker, and Diane Litman. "Evaluating Spoken Language Systems" In *Proceedings of American Voice Input/Output Society, AVIOS*, 1999.

Kirakowski 2003a
Jurek Kirakowski: "Questionnaires in Usability Engineering - A List of Frequently Asked Questions (3rd Ed.)" Human Factors Research Group, University College Cork, Ireland: http://www.ucc.ie/hfrg/resources/ qfaq1.html. Feb. 2003

Kirakowski 2003b
Jurek Kirakowski: "Estimating the Cost of Quantitative Evaluation" Human Factors Research Group University College Cork, Ireland. Originally 1994. WWW version http://www.ucc.ie/hfrg/resources/powerp.pdf. Feb. 2003

Kirakowski 1994/2003c
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

Kline 1986
Paul Kline: "A Handbook of Test Construction - introduction to psychometric design" Meuthen & Co. London 1986, ISBN 0-416-39430-2

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

Larsen 1997a
Lars Bo Larsen, “A Strategy for Mixed-initiative Dialogue Control “, In the proceedings of Eurospeech '97, September, 1997

Larsen 1997b
Lars Bo Larsen: “Investigating a Mixed-Initiative Dialogue Management Strategy”, Proceedings of IEEE Workshop on Speech Recognition and Understanding, ASRU 1997, December, 1997

Larsen 1998
Lars Bo Larsen: “The OVID Project Objectives and Results” Technical Report 98-0201 CPK Aalborg University, March 1998

Larsen 1999
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

Larsen 2003a
Lars Bo Larsen: “Applying The PARADISE Evaluation Scheme to an Existing Dialogue Corpus”. Submitted to ASRU’03, December 2003

Larsen 2003b
Lars Bo Larsen: “Assessment of Spoken Dialogue System Usability - What are We really Measuring?” Proc. of Eurospeech’03, Geneva Switzerland, September 2003

Litman et al 2000
Diane Litman, M. Kearns, S. Singh, M. Walker: “Automatic Optimization of Dialogue Management”, In Proceedings of COLING 2000.

Love et al 1994
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.

Mathworks 1999
Mathworks: “Statistics Toolbox. For Use with MATLAB - User’s Guide version 2”. The MathWorks Inc. MA, US, 1999. http://www.mathworks.com

Meko 2003
Dave Meko “Lecture notes in Time Series Analysis, Module 3" Laboratory of Tree-Ring Research, Univ. of Arizona, Tucson, Arizona. http://www.ltrr.arizona.edu/~dmeko/geos595e.html

Melin et al 2001
Melin H, Sandell A & Ihse M. “CTT-bank: A speech controlled telephone banking system - an initial evaluation”. TMH-QPSR, KTH, 1:1-27. KTH Stockholm 2001. http://www.speech.kth.se/qpsr/tmh/01-1-001-027.pdf

Montgomery 1997
Douglas Montgomery: “Design and Analysis of Experiments” 4th Ed. John Wiley and Sons 1997, Printed in the U.S. ISBN: 0-471-15746-5

Möller 2002
Sebastian Möller: “A New Taxonomy for the Quality of Telephone Services Based on Spoken Dialogue Systems” in proc. of the SIGDIAL Spoken Dialogue workshop, Philidelphia, 2002

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

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).

Oppenheim 1966
A.N. Oppenheim “Questionnaire Design and Attitude Measurement”. Heinemann Educational Books Ltd, London, 1979 (1966).ISBN 0-435-82676 X

OVID 1995
ESPRIT project No. 20171: OVID - Trial application of Voice Processing in Automated Telephone Banking Services: “Technical Annex”, July 1995, EU commission.

Perlman 2003
Gary Perlman: “Web-Based User Interface Evaluation with Questionnaires“. http://www.acm.org/~perlman/question.html (visited April 2003)

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

Polifroni et al 1992
J. Polifroni, L. Hirschman, S. Seneff and V. Zue: “Experiments in Evaluating Interactive Spoken Language Systems“. In Proc. of the DARPA Speech and Natural Language Workshop. pp. 28-33, Morgan Kauffman 1992.

Price et al 1992
P. Price, L. Hirschman, E. Shriberg, E. Wade: “Subject-based Evaluation Measures for Interactive Spoken Dialogue Systems“. In Proc. of the DARPA Speech and Natural Language Workshop. pp. 34-39, Morgan Kauffman 1992.

Rubin 1994
Jeffrey Rubin: “Handbook of Usability Testing - how to plan, design and conduct effective tests” Wiley Technical Communication Library, 1994, USA. ISBN 0-471-59403-2. 330 pages.

Scansoft 2003
X|mode Multimodal System datasheet: ftp://ftp.scansoft.com/pub/datasheets/xmode_datasheet.pdf

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

Siegel & Castellan 1988

Sidney Siegel, N.John Castellan: “Non parametric Statistics for the Behavioural Sciences” (second edition). New York, 1988, 1956, McGraw-Hill. ISBN 0-07-057357-3

Simpson and Fraser 1993

A. Simpson and N.A. Fraser: “Black Box and Glass Box Evaluation of the SUNDIAL System”. In *EUROSPEECH: European Conference on Speech Processing, Berlin, 1993, pp. 1423-1426*

Singh et al 2002

Satinder Singh, Diane Litman, Michael Kearns and Marilyn Walker. Optimizing Dialogue Management with Reinforcement Learning: Experiments with the NJFun System. *Journal of Artificial Intelligence Research(JAIR)*, 2002.

StatSoft 1999

StatSoft, Inc. (1999) Electronic Statistics Textbook. Tulsa, OK: StatSoft. WEB: http://www.statsoft.com/textbook/stathome.html.

SUMI 2003

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

Tabachnick and Fidell 2001

Barbara Tabachnick and L. Fidell: “Using Multivariate Statistics”, 4th. Edition, Allyn and Bacon, 1996 and 2001. Printed in MA, U.S. ISBN 0-321-05677-9.

Walker 2000

Marilyn A. Walker. An Application of Reinforcement Learning to Dialogue Strategy Selection in a Spoken Dialogue System for Email. *Journal of Artificial Intelligence Research, Vol. 12., pp. 387-416*, 2000.

Walker et al 1997

Marilyn Walker, Diane Litman, Candace Kamm and Alicia Abella. “PARADISE: A Framework for Evaluating Spoken Dialogue Agents”. *In Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics, ACL 97,* 1997.

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 et al 2000a

Marilyn. A. Walker, Candace. A. Kamm and Diane J. Litman. “Towards Developing General Models of Usability with PARADISE”. *Natural Language Engineering, 2000.*

Walker et al 2000b

Marilyn A. Walker, Lynette Hirschman and John Aberdeen. “Evaluation For Darpa Communicator Spoken Dialogue Systems”. In *Language Resources and Evaluation Conference, LREC*. 2000.

Walker et al 2001a
Marilyn A. Walker, Irene Langkilde-Geary, Helen Wright Hastie, Jerry Wright, and Allen Gorin. “Automatically Training A Problematic Dialogue Predictor for a Spoken Dialogue System” *Journal of Artificial Intelligence Research 16 (2002), pp. 293-319*. 2002 AI Access Foundation and Morgan Kaufmann Publishers.

Walker et al 2001b
Marilyn A. Walker, Rebecca Passonneau and Julie E. Boland: “Quantitative and Qualitative Evaluation of Darpa Communicator Spoken Dialogue Systems”. In *Meeting of the Association of Computational Linguistics, 2001*.

Walker et al 2001c
M. Walker, J. Aberdeen, J. Boland, E. Bratt, J. Garofolo, L. Hirschman, A. Le, S. Lee, S. Narayanan, K. Papineni, B. Pellom, J. Polifroni, A. Potamianos, P. Prabhu, A. Rudnicky, G. Sanders, S. Seneff, D. Stallard, S. Whittaker. “DARPA Communicator Dialog Travel Planning Systems: The June 2000 Data Collection”. In *EUROSPEECH: European Conference on Speech Processing, 2001*.

List of Abbreviations

ATIS	Air Traffic Information System
AVM	Attribute-Value Matrix
BT	British Telecom
CCIR	Centre for Communication Interface Research, Edinburgh University, Great Britain
CPK	The Center for PersonKommunikation, Aalborg University (since January First 2003 CPK is fully integrated into the Dept. of Communication Technology)
DARPA	(U.S.) Defence Advanced Research Projects Agency.
DM	Dialogue Management / Dialogue Manager
DSR	Distributed Speech Recognition
EAGLES	Expert Advisory Group on Language Engineering Standards
ETSI	European Telecommunications Standards Institute
FA	Factor Analysis
GPRS	General Packet Radio Service
MLR	Multiple Linear Regression
MUMMS	Measuring the Usability of Multi-Media Systems.
HFRG	Human Factors Research Group, University College Cork, Ireland
ISO	International Standardisation Organisation
PARADISE	Paradigm for Dialogue System Evaluation
PROMISE	Procedure for Multimodal Interactive System Evaluation
PCA	Principal Components Analysis
PDA	Personal Digital Assistant
QUIS	Questionnaire for User Interaction Satisfaction
SDS/SLDS	Spoken (Language) Dialogue System(s)
SMC	The Speech and Multimedia Communication Division at the Dept. of Communication Technology, Aalborg University
SR	Speech (concept) recognition Rate
SUMI	Software Usability Measurement Inventory
WAMMI	Website Analysis and MeasureMent Inventory.
WAP	Wireless Application Protocol
WOZ	Wizard-of-Oz