The Disc Approach to Spoken Language Systems Development and Evaluation

Dybkjær, L. (1), Bernsen, N.O. (1), Carlson, R. (2), Chase, L. (3), Dahlbäck, N. (7), Failenschmid, K. (4), Heid, U. (5), Heisterkamp, P. (6), Jönsson, A. (7), Kamp, H. (5), Karlsson, I. (2), Kuppevelt, J.v. (5), Lamel, L. (3), Paroubek, P. (3), Williams, D. (4)

(1) MIP, Odense University, Forskerparken 10, 5230 Odense M, Denmark

(2) KTH, Department of Speech, Music and Hearing, Drottning Kristinas Väg 31, S-100 44 Stockholm, Sweden

(3) LIMSI/CNRS, Human-Machine Communication Department, B.P. 133, F-91403 Orsay Cedex, France

(4) Vocalis Ltd., Chaston House, Mill Court, Station Road, Great Shelford, Cambridge CB2 5LD, UK

(5) IMS, Universität Stuttgart, Azenbergstraße 12, D-70174 Stuttgart, Germany

(6) Daimler-Benz, F3M/S, Wilhelm-Runge-Straße 11, Postfach 2360, D-89013 Ulm, Germany

(7) Linköping University, Dept. of Computer and Information Science, S-581 83 Linköping, Sweden

Abstract

The paper reviews recent progress in the DISC project on developing a systematic and general scheme for in-depth characterisation of current practice in the development and evaluation of spoken language dialogue systems and their components. This scheme consists of a 'grid' which serves to characterise the properties of any particular spoken language dialogue system or component, and a life cycle model which accounts for how that system or component was developed and evaluated. The work reported forms part of the wider DISC agenda of developing and testing a best practice methodology which will contribute to the establishment of dialogue engineering as a sub-discipline of software engineering.

1. Introduction

The development and commercialisation of integrated spoken language dialogue systems (SLDSs) is a recent phenomenon. Only within the last few years have SLDSs matured to the point of attracting broad industrial interest. Simple, speaker-independent, telephone-based SLDSs using continuous and spontaneous speech have now become commercially available. However, despite accelerating progress, SLDSs development and evaluation is replete with unknowns and steps that are undersupported in terms of procedures, concepts, theory, methods and software tools. At this time there are no accepted standards or even widely understood benchmarks for assuring potential customers or users of SLDSs of the quality of systems. Neither are there any reliable methods for comparing the quality of two SLDSs before selecting one for deployment in the field. This situation continues to generate uncertainty about the potential of SLDSs technologies, their proper domains of application, their usability, the cost of producing them, their development time and the quality of products in both absolute and comparative terms. In an increasingly competitive marketplace, the ability to state that some system has been developed following a carefully designed and validated dialogue engineering methodology, along with the ability to report evaluation results in a standardised framework, will give products developed in this way a competitive advantage. That in turn is likely further to stimulate take-up of the methodology by other organisations.

The DISC project (http://www.elsnet.org/disc/) is an Esprit Long-Term Research Concerted Action which started on 1 June 1997 and runs until 30 November 1998. The aim of DISC is to contribute towards establishing dialogue engineering as a sub-discipline of software engineering through developing a detailed and integrated set of development and evaluation methods and procedures (guidelines, checklists, heuristics), constituting a first dialogue engineering best practice model, as well as a range of support concepts and software tools.

DISC draws together actors from the national and European SLDSs development projects that have been executed during the last decade, i.e.: The Maersk Institute (MIP), Odense University, Denmark (co-ordinator); Human-Machine Communication Department, LIMSI/CNRS, France; Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart, Germany; Department of Speech, Music and Hearing, KTH, Sweden; Vocalis Ltd, UK; Daimler-Benz, Germany. Stichting Elsnet, The Netherlands, is tasked with DISC information collection and dissemination. The Department of Computer and Information Science, Linköping University, Sweden is a sub-contractor to KTH.

DISC pursues its goals through three work tasks addressing (a) current practice in the development and evaluation of SLDSs and their components, (b) best practice in the development and evaluation of SLDSs and their components, and (c) novel concepts, guidelines and software tools, respectively. The tasks of investigating current practice in dialogue engineering and of defining best practice are closely related. The DISC approach is to advance towards a first definition of best practice through a thorough investigation of current practice in the development and evaluation of SLDSs and their components. Both tasks focus on key aspects of SLDSs including speech recognition, speech generation, language understanding and generation, dialogue management, human factors, and systems integration.

For the initial investigation of those aspects, and in order to achieve a well-founded view of current practice, the DISC partners contribute full access to a wide range of products, running prototypes and prototypes under development. Each current practice aspect review in DISC is based on at least three significantly different exemplars of SLDSs or SLDS components.

This paper describes and illustrates ongoing DISC work towards achieving a well-founded view of current practice. The work originated with a first skeleton dialogue engineering model of SLDSs and their components as well as of how SLDSs and their components are currently being developed and evaluated (Section 2). The model was subsequently revised and refined through the study of the DISC exemplars and is now getting to the stage of providing a solid basis for in-depth characterisation of individual systems and components as well as for making comparisons across systems and components (Section 3). The model will continue to be refined throughout the life time of DISC, eventually leading to the final DISC best practice model (Section 4).

2. The First Skeleton DISC Dialogue Engineering Model

Current documentation on SLDSs and their components tends to provide only selective information on their properties as well as on their development and evaluation process (Fraser, 1995). Moreover, the documentation does not have any standard conventions to follow. It is therefore often difficult or impossible to (i) discover deficiencies of a system or component, other than those explicitly (and infrequently) reported; (ii) compare test results, because of insufficient context; and (iii) compare and evaluate systems, because of insufficient and incomparable information. One approach to removing these very real problems is to develop a standardised scheme which can provide the information required for characterising an interactive speech system or component and its development and evaluation process in a way which includes sufficient contextual information and allows comparison with other systems. The DISC dialogue engineering model is being developed for this purpose.

The first skeleton dialogue engineering model was based on work presented in (Bernsen et al., 1998) and consists of a 'grid' and a life-cycle model. The *grid* is a refinement of the task-oriented theory of spoken human-computer interaction illustrated in Figure 1. Figure 1 reflects analysis of a number of interactive speech systems as well as concepts derived from a range of theories of spoken human-human interaction. The theory illustrated in the figure provides a set of elements of interaction which are important to the design and construction of SLDSs. The theory takes a software-oriented approach, focusing on elements that may be used to construct high-level models of SLDSs and explain their behaviour. Hardware, such as telephones and microphones, is not included and the same holds for the user's physical work environment.

The model in Figure 1 exhibits two modes of organisation. First, the elements have been organised into five *layers*. At the bottom of the figure, the *context* layer includes history of interaction, domain model and user model. At the level above the context layer, the interaction *control* layer includes states of attention as well as the structures defined by the interlocutors' intentions and structural aspects of their linguistic exchanges. System control is largely based on structures at this level. The following *language* layer describes issues of linguistic interaction**.** Then follows the *speech* layer which includes the transformations between speech signals and the symbolic expressions of language. The final *performance* layer is a function of the other layers taken together and includes some general issues of system behaviour. Secondly, the grey band in Figure 1 indicates the overall processing *flow* among the various types of element - input, intention and attention, output and performance - in a context defined by contextual elements. Note, however, that the rounded boxes in Figure 1 do not represent system modules but, rather, logical processing steps. Developers often refer to the elements in Figure 1 in terms of the corresponding system modules, such as the recogniser, parser, dialogue manager, inference engine, text generator and player, system performance being replaced by an abstraction of the (physical) user. Whilst related to system modules, the logical processing steps in Figure 1 could in principle be performed by systems with considerably different modular architectures.

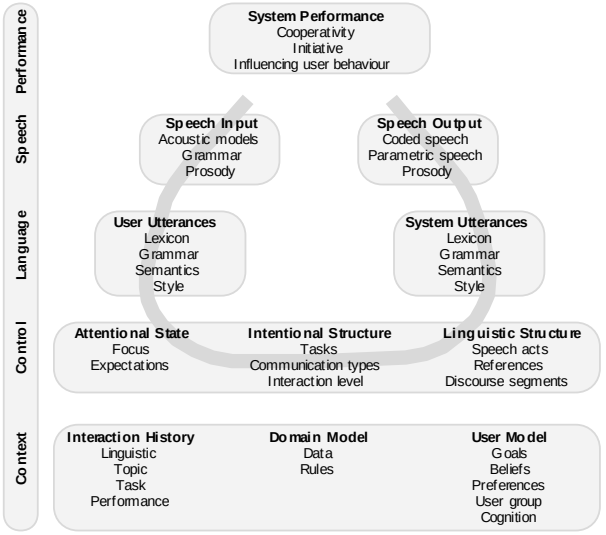

Figure 1. Elements of an interactive speech theory. The grey band and grey boxes reflect the logical architecture of SLDSs.

The DISC 'grid' takes the form of a series of "checklist" entries which should enable a comprehensive characterisation of the properties of any SLDS or SLDS component. The first DISC grid was heavily based on Figure 1 and was somewhat biased towards dialogue, i.e. the control and context layers, and system cooperativity. Figure 2 shows the representation of the control layer in the Danish Dialogue System in (Bernsen et al. 1998). Subsequent versions of the grid have been very substantially expanded, particularly as regards the language and speech layers.

In the DISC dialogue engineering model, the 'grid' is complemented by a *life-cycle model* which aims to capture the development and evaluation process of

particular SLDSs or SLDS components. The life cycle model departed from work presented in (Bernsen et al., 1998, Chapter 3) and has subsequently been extended based on discussions in DISC. Figure 3 shows a draft summary of five different speech recognition systems that have been analysed in DISC.

Attentional state	
Focus	Current sub-task plus meta-communication tasks.
Expectations	Predictions sent to recogniser and parser; task dependent parsing.
Intentional structure	
Tasks	Danish domestic flight ticket reservation; well-structured task.
Communication	System-directed domain communication. Mixed initiative meta-communication; users may initiate meta-communication through keywords. System-directed other communication, such as the opening and closing of a dialogue.
Interaction level	Some questions are yes/no or multiple choice, most are general and focused.
Linguistic structure	
Speech acts	Primitive distinction between commands (meta-communication) and statements (answers) in user input; use of commands (questions), and statements for providing feedback, error messages and other information in output.
References	No anaphora resolution; ellipses are being handled.
Segments	-

Figure 2. Grid representation of the control layer in the Danish Dialogue System.

Overall design goal(s): *What is the general purpose(s) of the design process?*

The speech recognizers we investigated were all designed as continuous speech, speaker-independent medium vocabulary systems, to be embedded in a spoken language system. They were also designed to run in real-time (or close enough so as to be perceived as real-time) while minimizing word error rate to the extent possible. For many of the systems, robustness was also sought in order to retain an adequate level of performance in the presence of background noise (e.g. telephone channels or public halls).

Hardware constraints: *Were there any a priori constraints on the hardware to be used in the design process?*

Hardware constraints were not typically a priority, as the systems were designed as prototypes. High-end Unix workstations are generally used to run the recognizers. These machines typically have fast processors and are heavily loaded with memory - they are top-of-the-line scientific workstations.

Software constraints: *Were there any a priori constraints on the software to be used in the design process?*

The systems are generally developed in C/C++ to enhance portability across various platforms.

Designer preferences: *Did the designers impose any constraints on the design which were not dictated from elsewhere?*

Certain sites in the speech recognition research community regularly host organized performance evaluations. Comparative evaluation on common tasks has influenced international progress in large vocabulary continuous speech recognition, and the development of the speech recognizers analyzed here has been influenced by this process.

Design process type: *What is the nature of the design process?*

Most of the systems we looked at were exploratory research systems.

Development process type: *How was the system/component developed?*

They were all tested under laboratory conditions and some have been deployed in real operational environments. Generally the design and development process was iterative and not well documented.

Realism criteria: *Will the system/component meet real user needs, will it meet them better, in some sense to be explained (cheaper, more efficiently, faster, other), than known alternatives, is the system/component "just" meant for exploring specific possibilities (explain), other (explain)?*

The realism of the speech recognition components varied from system to system. The need to provide real-time speech recognition able to handle speech from unknown speakers indicates at least a minimal attempt to meet real user needs. In some cases, real users use the systems regularly.

Functionality criteria: *Which functionalities should the system/component have (this entry expands the overall design goals)?*

In terms of functionality, real-time speaker-independent recognition of a variety of languages are supported by the systems. Most systems operate in only a single language, although some are available in multiple languages. For some of the systems telephone quality speech was supported.

Customers: *Who is the customer for the system/component (if any)?*

As most of the systems we looked at were developed under European or national contracts, there was typically no predefined customer. However, the recognition components were often developed with other projects in mind, and often the same core software is used in other systems or products.

Users: *What are the intended users of the system/component?*

The target users are generally native speakers of the language in question, with no *a priori* knowledge of the service or system. No speaker-specific training is assumed.

Figure 3. Part of draft life cycle model summing up five different analyses of the speech recognition aspect, done by the LIMSI group. This model was based on life cycle models of each of the five speech recognition components summarised in the figure.

3. The DISC Current Practice Review

Our initial approach to current practice was to analyse a series of systems and components with respect to the six aspects of speech recognition, speech generation, language understanding and generation, dialogue management, human factors, and systems integration. The exemplars that were analysed with respect to one or more aspects were: The French LE Arise system on telephone accessed train time-table information systems (http://www2.echo.
lu/langeng/en/le3/arise/arise.html), the the CMU Phoenix parser (Ward and Issar, 1995), the Daimler-Benz dialogue manager (Heisterkamp and McGlashan, 1996), the Daimler-Benz parser (Mecklenburg et al., 1995), the Danish Dialogue System for flight ticket reservation (Bernsen et al., 1998), the Vocalis Operetta automated call routing system (http://www.vocalis.com/products/
operetta/infoframe.html), the Vocalis Voice Activated Dialling system (http://www.vocalis.com/products/speech
tel/infoframe.html), the Verbmobil spoken language dialogue translation system (http://www.dfki.de/
verbmobil/), and the multimodal Waxholm tourist boat information system (http://www.speech.kth.se/waxholm/
waxholm.html).

From the point of view of methodology, each aspect is being analysed by at least two different sites. For each aspect at least three significantly different exemplars are being investigated. No aspect of a system or component is being analysed by a site that has been involved in its development and evaluation. Every analysis of an aspect of an SLDS or component is being verified by the developers of that particular SLDS or component.

Analysis of an aspect of a particular system or component consists in applying the 'grid' and the life-cycle model to the description of that particular exemplar. During this process, it often happens that the grid or life cycle model has to be expanded, or otherwise revised, in order to appropriately characterise the exemplar in question. Grid and life cycle model application requires large amounts of information on systems and components. Typically, first versions of the 'grid' and the life-cycle model were completed on the basis of available papers and reports describing a certain system or component. This first iteration always produces a - sometimes quite large - number of questions which cannot be answered with sufficient certainty, or not at all, based on the initially collected information. Answers to such open questions are then being sought through, i.a., email or telephone interaction with the colleagues who were involved in the development and evaluation of the particular aspect of the system or component in question, access to additional data, such as transcriptions and recordings of user-system interactions, and site visits, interviews and demonstrations. In fact, site visits have proved necessary to the satisfactory analysis of most DISC exemplars. The final step in the analysis of an aspect of a system or component is to invite verification from that system or component's developers in order to remove any misconceptions from the grid and life-cycle representations.

The results of individual system and component aspect analyses have been thoroughly discussed at the second DISC workshop in March 1998, which led to many additions to, and revisions of, in particular, the grid. For example, it was necessary to add a series of entries on multimodality to the grid because some of the systems under investigation include modalities other than speech.

The revised grid turned out to become so detailed and comprehensive as regards the entries concerning each aspect that it was decided to develop a two-level grid hierarchy consisting of (i) a general grid capturing the key properties for each aspect and meant to provide an overview of each exemplar analysed, and (ii) a detailed version which will capture the properties of any of the six aspects in much greater detail. The idea is that, when focusing on a particular aspect of a system, such as dialogue management, it is convenient to have an overview of the system in which the dialogue manager is embedded whereas much greater detail is desirable for the aspect in focus. Figure 4 shows a general-grid description of a speech recogniser.

Speech input	
Nature	Continuous, spontaneous speech
Device(s)	Telephone, microphone
Phone server	Yes
Acoustic models	SCHMM
Search	A*
Vocabulary	5000; good performance for up to 10000 words but not used in any application yet
Barge-in	No
Word hypotheses	Yes, word hypothesis graph
Grammar	Statistic language model or finite state model over categories
Prosody	-

Figure 4. A completed general grid concerning speech input, done by the MIP group. The grid contents provide an overview of a particular speech recognition component.

Once all the individual aspect analyses have been completed, the next step is to integrate the results from different partner sites concerning each of the six aspects under investigation, thereby arriving at a common DISC representation of current practice. DISC is in this phase right now, which has proved to lead to fruitful and sometimes quite comprehensive discussions of the theoretical foundations for systematically and unambiguously characterising current interactive speech technologies. Expectedly, this phase will lead to further revisions of the grid and the life cycle model. Early examples of this process of integration are shown in Figures 3, 5 and 6.

4. From Current Practice to Best Practice

In the work done in DISC so far, emphasis has been on developing, clarifying, sharpening, and agreeing upon the contents (entries) of the grid and the life cycle model. We have concentrated on identifying the questions which should be asked in the grid and the life cycle model, and on distinguishing important issues from less important ones. Essentially, the theoretical basis for this work has been the knowledge resources that were available in the DISC consortium from the outset, strongly enriched, of course, through the experiences gained from having to analyse in

Speech output	
Sound generation technique	System A Formant synthesis by rule. Based on smoothed square waves for most parameters. Reflects the model that speech production is created by step functions smoothed by articulators. System B Concatenation with PSOLA, the stored units are di-phones, syllables, consonant clusters that are combined in accordance with the transcription given by the text-to-speech rules or from the lexicon. 2200 units are stored for each voice. System C Concatenation, the stored units are at least one word and up to one sentence long. In all less than 2000 units for one voice.

Figure 5. There are sometimes big differences among the findings for an aspect as shown in this comparison of three systems as regards the grid question on sound generation technique, done by the KTH group.

Early results show that although SLDSs are tightly integrated software systems with numerous (semi-) autonomous functional modules, they tend to make use of proprietary standards and protocols. This makes modification and adaptation of the systems to a new target domain time and cost extensive. Furthermore, the systems integration life-cycles for research systems differ from the ones for commercial systems. The individual stages in the life-cycle are identical for the two types of systems, however systems integration for research systems tends to be driven by the need for integration of existing functional modules. By contrast, systems integration for commercial system tends to be driven by the need of achieving certain functionality as described by the client.

Figure 6. Example of early summary evaluation of DISC findings on current practice in systems integration, done by the Vocalis group.

depth a series of often unfamiliar SLDSs and components.

Once a consolidated representation of the original DISC exemplars has been arrived at, the next step will be to invite colleagues from outside the DISC consortium to comment on the grid and the life cycle model. The DISC Advisory Panel has been established for this purpose (http://www.elsnet.org/disc/ap/. We hope that they will do so not only by letting us benefit from their wide experience in terms of remarks on the grid and the life cycle model and their individual entries, but also by applying the grid and life cycle model to exemplars that they are familiar with. Having incorporated their input, the DISC consortium will take the grid and the life cycle model one step further: from being purely *descriptive* models of current practice to becoming first draft best practice *prescriptive* models for spoken language dialogue systems development and evaluation. Again, this step will initially be performed on the basis of the expertise available in the DISC consortium.

The best practice draft will then enter an iterative test phase in which the testing of the grid and the life cycle model on novel systems and components from projects inside as well as outside the consortium, will alternate with model revisions. It is the aim to involve as many developers as possible in using and testing the best practice drafts. At some point during the test phase, the grid and life cycle model focus will shift from *contents-only* to *contents-and-form*. For the time being, it is not known how to package the DISC dialogue engineering model for optimum usability by dialogue systems developers. Several approaches have been proposed for packaging the grid and the life cycle to this effect. Also these ideas, as well as others that may emerge in months ahead, will need to be tested in DISC-external development practice.

References

ARISE: http://www2.echo.lu/langeng/en/le3/arise/arise.html

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

DISC: http://www.elsnet.org/disc/

Fraser, N. M. (1995). Quality standards for spoken language dialogue systems: a report on progress in EAGLES. In Proceedings of the ESCA Conference on Spoken Dialogue Systems, Theories and Applications, Vigsø, 157-160.

Heisterkamp, P. and McGlashan, S. (1996). Units of dialogue management: an example. In Proceedings of ICSLP'96, Philadelphia, 200-203.

Mecklenburg, K., Hanrieder, G. and Heisterkamp, P. (1995). A Robust parser for continuous spoken language using PROLOG. In Proceedings of Natural Language Understanding and Logic Programming 1995, Lisbon, Portugal, 127-141.

Operetta: http://www.vocalis.com/products/operetta/infoframe.html

Verbmobil: http://www.dfki.de/verbmobil/

Voice Activated Dialling: http://www.vocalis.com/products/speechtel/infoframe.html

Ward, W. and Issar, S. (1995). The CMU ATIS System. In the proceedings of the ARPA Workshop on Spoken Language Technology, January, 249-251.
Waxholm: http://www.speech.kth.se/waxholm/waxholm.html