END-USER DRIVEN DIALOGUE SYSTEM DESIGN: THE REWARD EXPERIENCE

Klaus Failenschmid and J.H. Simon Thornton

Vocalis Ltd,
Chaston House, Mill Court,
Great Shelford, Cambridge CB2 5LD, UK
klaus.failenschmid/simon.thornton@vocalis.com

ABSTRACT

In the EC funded research project REWARD professionals in the target domain for a dialogue system were enabled to develop their own speech driven telephone service (dialogue system). Specialists in speech recognition technology provided them with graphical tools and technical advice. Guided by their in-depth understanding of the target domain and target user the four user partners developed four different dialogue systems. The shift of focus in the dialogue system design process from technology to end user driven design 'rationale' is influenced by research into the design process of dialogue systems and provides valuable data that informs user-centred-design processes of interactive systems.

In this paper we report on observations with regard to this end-user driven design process. Dialogue features, which are a direct result of this different approach to designing speech recognition applications, are presented.

1. INTRODUCTION

The increase in quality and robustness of speech recognition technology in recent years has helped to create a growing commercial demand for automated telephone services using speech recognition. Hence, the technology has moved from the research arena out into the wider public through the automation of telephone-based services (e.g. switch board, dialling by voice, train enquiries) [8, 1].

Such automated services tend to be highly domain and task specific in order to place sufficient constraints on the scope of the dialogue between the user and the system. This limits the size and complexity of the recognition vocabulary and the language model used by the system, a constraint imposed by the technology used. Because of this, the design and development process of Spoken Language Dialogue Systems (SLDSs) has often been guided by the constraints imposed upon a user-system dialogue by the technology.

Furthermore, because of the lack of simple design tools and relatively complex relationship between various components necessary in speech recognition based SLDSs, these have traditionally been designed by speech recognition and dialogue experts for clients. Apart from having a leading role in drawing up requirements and functional specifications, the clients (who are generally specialists in the target domain of the SLDS) have usually not had an active role in the practical design and development period.

Providing domain specialists with a more active role in the (practical) design and development process is likely to lead to SLDSs which have been designed guided by a thorough understanding of the end user and their requirements rather than the technological constraints. The resulting shift of focus in the dialogue system design process from technology to end user driven design 'rationale' can extend existing research into the design process of dialogue systems [2, 5] and provide valuable data that informs user-centred design processes of interactive systems [9, 11].

The EC Language Engineering project REWARD (REal World Applications of Robust Dialogue) [7] set out to introduce such a shift in focus through providing domain specialists with a more active role in the design process of a SLDS.

In the following sections of this paper we will report on the experience gained in the REWARD project. The paper is structured as follows. Section 2 introduces the aims of the REWARD project; section 3 describes the applications designed by the partners in the project. Lessons learned in (technically) supporting designers are discussed in section 4, which leads into a discussion of some aspects of the design process (section 5) before the REWARD user trials are considered in section 6.

2. OVERVIEW: THE REWARD PROJECT

REWARD was a three year EC Language Engineering project (December 1995 to November 1998) that brought together four user organisations and two technical suppliers. The user organisations were the Danish travel agency DanTransport Rejsebureau A/S (DTR), the British telemarketing and market research organisation Taylor Nelson AGB plc (TN), the Dutch market research organisation Nederlands Instituut voor de Publieke Opinie en het Marktonderzoek BV (NIPO) and the Spanish hardware maintenance company Manufacturing and Development S.A. (MADE). Vocalis Ltd (UK) and the Centre for PersonKommunikation, University of Aalborg (CPK, Denmark) acted as technology suppliers and specialists in speech recognition technology. All user partners had little or no speech recognition experience.

REWARD had two main aims,

1. the design of speech recognition enabled telephone services to be carried out by target domain specialists (design track), and

2. the development of an integrated dialogue creation and runtime environment (development track).

The way in which these two aims were going to be achieved was revised in spring 1997 which resulted in a twin-track approach for the second phase of the project The two tracks (design track and development track) ran in parallel.

The Design Track: One objective of REWARD was to give user partners (domain specialists) easy access to speech recognition technology to enable them to design and deploy their own telephone services utilising this technology. The four user partners were provided with Vocalis' existing dialogue design tools and runtime environment as well as technical support and training (in speech recognition technology). Using this technology, they developed SLDSs in Danish, British English, Dutch and Spanish.

The Development Track: The technology suppliers not only supplied the user partners with an initial set of graphical dialogue design tools and the dialogue runtime system but also embarked on developing a new suite of dialogue design tools [4]. Furthermore, the existing dialogue runtime system was re-developed. These two developments were based on advice from the non-expert designers. For more information about this aspect of the REWARD system refer to [3, 4].

Throughout the two-track practical design and implementation phase technology suppliers and user partners stayed in close contact making use of email and technical workshops.

3. TELESERVICES

Four quite different innovative teleservices were developed and tested with members target user groups. The main aim was the development of systems that are extremely robust and reliable so that a large number of calls can be automated.

The DanTransport (DTR) system is an automated domestic flight ticket reservation service for a selected group of regular (commercial) customers. The language of the service is Danish. It is fully integrated into DTR's on-line ticket booking system.

Taylor Nelson has developed a magazine subscription service in English. This service could be available for members of the general public who want to subscribe to a magazine and pay by credit card. The demonstration system can take orders for over 50 titles and provides full agent back-up at any state in the dialogue through its complete integration with an existing call centre.

NIPO's telephone service is a market research survey interviewing system which leads callers form a research panel through structured opinion-seeking interaction. The language used in the service is Dutch and the researched subject is recent holidays.

Key aspects of MADE's customer care service were automated in their application. Call-routing by name and product repair status information are the two main tasks the application supports. The language used is Spanish with key recognition items also recognisable in Castillian. It is fully integrated with their local repair status database as well as their company internal telephone network (for call routing).

The developed services not only differ in the language they use, but also in their complexity. While DTR's, TN's and NIPO's services require the collection of a large number of individual pieces of information, MADE's service only collects very few items (customer name and repair number). However, due to the wide variety of system outputs possible in MADE's service a speech synthesiser is used.

All services are turn-based, make use of isolated word, wordspotting and connected digit recognition. Some also offer talkover facilities.

4. THE DESIGN PROCESS

Technology suppliers were not actively involved in the (practical) development of the above-described automated telephone services. During the active design process in phase two of the project, their role was more passive through giving technical advice and assistance. However, they had a more active role during the first phase of the project. In this phase, the requirements and functional specifications for the telephone services were drawn up by the user partners in co-operation with the technology suppliers [10]. Along with giving advice on the feasibility of the implementation of proposed functionality, a core task was to provide 'consultancy' with respect to the capabilities of the available speech recognition technology. Since the user partners had either none or only very little experience with speech recognition technology, the technology suppliers offered a technical tutorial and a workshop that introduced the participants to the complex field of speech technology.

After this initial 'educational phase' user partners were supplied with the technology needed to build automated telephone services and could embark on their own designs. The technology consisted of Vocalis' runtime system including their proprietary speech recogniser plus a set of tools, namely

- a graphical dialogue design tool for finite state dialogues,
- a recognition vocabulary design tool,
- a set of programs to install applications on the runtime system,
- a prompt recording tool.

Up to this point, the design process was not dissimilar to a 'technology expert focussed' design process (see also [6]) in that the requirements for the service to be built were developed jointly between technology experts and domain experts. Differences occurred in the subsequent design phase since it was the *domain experts* who became active through doing all the design work. In this phase, technology suppliers were less engaged in 'application consultancy' work, but focussed on technical user support. This was provided through two main channels:

- email list
- technical workshops

User partners and technology suppliers subscribed to an email list in which technical problems were discussed. In addition to that, three technical workshops were held throughout the design phase. These brought together participants from all REWARD partners to discuss design progress and technical problems.

The nature of the user support work changed along with the users' growing experience with the technology at their disposal. While in the earlier days of the development process technical support was mainly given on technical issues such as the correct use of the provided tools, later on technical support was more sought on complex speech recognition and dialogue control questions.

For instance, it was found in the early days of development, that the toolset used by the designers did not adequately support application design in the four different languages represented in the REWARD consortium. Changes had to be made to the design tool to produce comprehensive support for the languages required. Also, since the tools had been developed as in-house development support tools by the same people who would normally design applications without such tools, non-expert designer often had problems understanding a certain behaviour of the tools. This was also true for error messages appearing on screen. Furthermore, the common use of numbers instead of names (for instance to identify words in the recognition vocabulary) was also difficult to relate to by the users.

The tools-related comments not only resulted in improvements of the existing tools where this was essential for a successful continuation of the design work, but, more importantly, were used in the design of the new REWARD tool set [4].

After this initial phase of technical user support enquiries related to the available tools, support was increasingly sought on speech recognition related issues such as recognising long phrases and extensive vocabularies (for more examples see below). This is because the technical tutorial and workshop at the beginning only served as an introduction to the field and a more thorough understanding could only be gained through hands-on design work.

Some of the technical problems encountered by the designers surprised the technology suppliers because they had never been observed when technology specialists designed applications. A possible explanation for this is that since the technology specialists who used the toolset were very aware of technical constraints they had interpreted the requirements and functional specifications in such a way that their designs would not challenge these constraints. In contrast to this, domain specialists interpreted the same specifications mainly guided by interactional requirements (e.g. dialogue structure, etc.).

5. DIALOGUE EXAMPLES

Through technically supporting user partners in their design process it became apparent that the domain specialists designed their speech enabled telephone service differently from the designers in the technology supplier's company. In this section we consider two typical examples where differences became apparent.

5.1 Dialogue Strategies

One user partner designed an application that includes the recognition of a credit card number. This is a subtask that has also been designed by a technical expert as part of a different application. We can therefore attempt a comparison of the two approaches that ultimately achieved the same goal. Such a comparison suggests that while the technology-expert designer made assumptions about (co-operative) caller behaviour that favoured a successful recognition, the domain expert designers favoured a view that erratic behaviour of callers is much more likely.

For instance, in order to optimise recognition performance the expert designer constrained the speech recogniser to the correct number of digits (or subset of digits) in a credit card recognition task assuming that the caller has the credit card ready and will utter the expected number of digits. Also, the repair of a mis-recognised credit card was attempted regardless of the number of mis-recognised digits.

In contrast to this - presumably based on experience 'in the field' - one user partner opted for constraining the recogniser to a range of digits that is slightly smaller and slightly larger than the number of digits which would be expected by looking at the credit card number. This allows the caller to utter a string of digits with the incorrect length. Furthermore, the repair of a mis-recognised credit card number would only be attempted if it were obvious that the recognition has resulted in only minor errors (such information is available through checking test-digits, etc. in the credit card code). In all other cases the caller would automatically transferred to a human operator who would try to resolve the problem.

While both designs for handling this subtask made use of algorithms that checked the validity of the recognised credit card number, only the technology-expert version actually attempted to repair an incorrect number. The domain-expert version used the result of the analysis to decide whether to continue in the dialogue or to transfer the caller to an operator.

Early limited analysis comparison of other similar subtasks and dialogues suggests that there is a difference in terms of error handling between domain expert designed dialogues and technology expert designed dialogues. While the former focus on dealing with interactional errors such as unexpected or incomplete input, the latter focus on technical errors such as dealing with mis-recognitions. Experience in different aspects of dialogue design is likely to be the reason for this; domain experts have a good understanding of the how people tend to carry out a similar task, while technology experts know what can lead to mis-recognitions.

5.2 Recognition Vocabulary

Domain experts experienced problems with designing for the available speech recogniser. The designers' limited knowledge

of speech recognition algorithms made it difficult for them to understand limitations and constraints imposed on recognition tasks. Recognition vocabularies were usually designed with the (planned) interaction with the system rather than the recogniser constraints in mind. This occasionally resulted in unexpected performances of the recogniser such as continued mis-recognitions, which did not seem to be logical.

An example for this is that the technology provided to the users only supported a limited size recognition vocabulary. In the beginning the designers constructed recognition vocabularies with numerous heterophones. It was thought that this would improve the recognition performance since a number of different ways of pronouncing a certain word were catered for. However, the negative effect of it was that besides only improving recognition performance very slightly it resulted in vocabularies that exceeded the existing limits of the recogniser.

This example highlights the changing role of technology constraints in the design process. While, in the beginning, the designers were not fully aware of these and their impact on the design process, they learned to better balance domain and technology constraints. This was clearly triggered by gaining experience in designing applications and encountering technical constraints. Technology suppliers helped by explaining the reason for problems occurring and suggesting ways to avoid them. As highlighted by the first example, the final design decision would often be based on a trade-off between desired usability functionality and technical capabilities.

6. FIELD TRIALS

REWARD is an innovative application project, which follows a commonly adopted development life cycle which is divided into three main stages: preparation, development and verification, and demonstration. The first two stages have been completed including the core design and development work as well as comprehensive design and technology testing with staff at each of the four user sites. The stage three (public) user trials will be conducted during September and October 1998. They are iterative so that the services may be adjusted based on user feedback and focus on three main issues:

- Equipment performance
- Dialogue and recognition performance
- User acceptability and performance.

It is expected that early results of these trials can be presented at the conference.

7. CONCLUSION

The experience in REWARD has shown that it is possible to get domain experts with no or only very little experience with speech technology to successfully design commercial speech recognition based telephone services. When designing applications, the REWARD dialogue system designers were largely guided by their thorough understanding of the application domain and the end users of the service. Detailed knowledge of how automatic speech recognition and natural language technology work and what constraints this technology exhibits had to be learned and reviewed during the design process. Hence, dialogue strategies were regularly reviewed based on the increased knowledge and designers learned to better balance domain and technology constraints.

5. REFERENCES

1. Aust, H., Oerder, M., Seide, F., and Steinbiss, V., “The Philips Automatic Train Timetable Information System”, *Speech Communication* 17: 249-262, 1995.

2. Bernsen, Niels Ole, Dybkjaer, Hans, and Dybkjaer, Laila, *Designing Interactive Speech Systems - From First Ideas to User Testing*, Springer Verlag, 1998.

3. Broendsted, Tom, “The Linguistic Components of the REWARD Dialogue Creation Environment and Run Time System”, *Proceedings of the 4th IEEE Workshop on Interactive Voice Technology for Telecommunications Applications*, Turino, September 1998.

4. Broendsted, Tom, Bai, Bo, and Olsen, J. “The REWAD Service Creation Environment. And Overview”, *these proceedings*.

5. Cheepen Christine, “Designing Advanced Voice Dialogues - What do Designers do and What does this Mean for the Future?” *http://www.soc.surrey.ac.uk/research/reports*.

6. Failenschmid, Klaus “Spoken Dialogue System Design - The Influence of the Organisational Context on the Design Process”, *Proceedings of the 4th IEEE Workshop on Interactive Voice Technology for Telecommunications Applications*, Turino, September 1998.

7. Fraser, Norman M., “Voice-based Dialogues in the Real World”, in: Varghese, K., and Pfleger, S., (Eds.), *Human Comfort and Security of Information Systems*, Springer Verlag, pp.75-86, 1997.

8. Fraser, Norman, M.., Salmon, Bob, and Trevor, Thomas “Call routing by name recognition: field trial results for the Operetta™ system”, *Proceedings of the 3rd IEEE Workshop on Interactive Voice Technology for Telecommunications Applications*, Basking Ridge, October 1996.

9. Norman, Donald, and Draper, Steve, *User-Centred Design*, Lawrence Erlbaum Associates New York, 1986.

10. Piper, Heather, and Fraser, Norman, (Eds.) “Requirements and Functional Specifications of Teleservices and Service Creation Tool”, *REWARD project deliverable tn004_4*, 1996.

11. Poltrock, Steven, and Grudin, Jonathan, “Organisational Obstacles to Interface Design and Development: Two Participant User Studies”, in: Rudisill, M., et al. *Human Computer Interface Design*, Morgan Kaufman San Francisco, pp.303-337, 1995.