A Human Centered Approach to Speech Interfaces in Mobile and Ubiquitous Computing Environments

Botond Pakucs

CTT, Centre for Speech Technology, Dept. of Speech, Music and Hearing
KTH, Royal Institute of Technology,
Drottning Kristinas väg 31, 100 44 Stockholm, Sweden
botte@speech.kth.se

Abstract. Currently, speech interfaces are about to be integrated in consumer appliances and embedded systems and are expected to be used in mobile and ubiquitous computing environments. New usability and human computer interaction related problems may be introduced in these environments. We argue that a user-centered approach, with support for personalization, user modeling and context awareness should be used when designing speech interfaces for these new environments. Further, we are proposing a novel user-centered and application independent architecture for speech interfaces. Finally, we present SesaME, a generic dialogue manager built according to the suggested architecture.

DRAFT!!!

1 Introduction

Recently, the possibility to use speech interfaces in embedded products and consumer appliances in mobile and ubiquitous computing (UC) environments has begun to attract interest in the speech technology community. If the growth of speech interfaces will be as large as expected, users will be surrounded by a multitude of speech controlled services and appliances. However, in the new environments the requirements on speech based interfaces may increase and new, human computer interaction (HCI) and usability related, problems may be introduced.

In this paper we will discuss some major usability and HCI related issues, which should be considered when designing speech based interfaces for mobile and UC environments. We argue that a user-centered approach, with support for personalization, user modeling, and context awareness is required for speech interfaces in these environments. Further, we propose a novel user-centered and application independent architecture for speech interfaces. We suggest that each user should use a single personalized speech interface for accessing a multitude of appliances and services in these new environments. Finally, we present SesaME, a new generic and task-oriented dialogue manager built according to the proposed architecture and for facilitating adaptivity, user modeling and context awareness.

2 Background

In UC the user's interaction with the distributed services and embedded appliances should be a not-annoying interaction, also called *calm computing* [19]. Intuitiveness, ease of use and seamless access are some of the major desired features. The user should be able to concentrate upon the task to be performed and not be forced to cope with interface issues.

For these requirements, speech based natural language interfaces appear to be a desirable choice. Speech interfaces are also advantageous in mobile situations and/or when hands and eyes are busy.

The suitability of speech interfaces for mobile and UC environments was for instance discussed by Sawhney and Schmandt [16]. However, due to the transitory and serial nature of speech and due to the fact that the current speech technology solutions have not reached maturity yet, speech interfaces are usually considered unreliable and hard to employ. Speech interfaces in experimental UC environments are used in a limited way [12] or not at all. The research on speech interfaces for UC environments is also rather sparse.

On the other hand, the speech technology industry has already recognized the potentials of the new emerging market. Some speech controlled appliances may be just around the corner and about to hit the market.[1]

Application-centered design and speech processing functionality embedded in, and performed by, the appliances themselves [11], are the main common features of the proposed industry solutions. However, in environments with diverse and concurrent speech interfaces, new usability and HCI related problems may be introduced.

3 Usability issues

Speech based interaction with ubiquitous services differs from accessing speech services through telephones or interacting with desktop based computers. Being on the move, and with hands, eyes and sometimes even with the mind busy, the user's demands on the HCI are significantly increased. The user will not have the time and patience to correct missrecognitions and misunderstandings.

Designing and building user-friendly speech based interfaces with excellent usability properties in their own right just might not be enough. The scenario with a multitude of different and concurrent speech interfaces, may still result in environments with unsatisfactory usability characteristics and poor user satisfaction.

There is a need for a shift in how we think about speech based interactions in mobile and UC environments. The usability and HCI issues should be considered for whole environments rather than for isolated services and appliances. We will discuss below each of the major factors which may cause new usability and HCI related problems.

[1] For example Telespree has released a voice operated mobile-phone. (http://www.telespree.com/products/main_handset.html)

3.1 Diverse Speech Technology Solutions

We may expect, in the near future, a multitude of speech interfaces with various complexity and based on different speech technology solutions employing from simple voice triggered commands to complex mixed-initiative, conversational dialogue systems.

Interface consistency is a central and well understood concept in the HCI and usability community [7, 14]. The importance of the consistency among different speech interfaces has also been recognized [8]. When encountering diverse speech interfaces, it will be hard for the users to identify the currently available voice commands, dialogue management solutions and vocabularies. It might even be hard for the users to know which services and appliances can be speech controlled. The lack of consistency among different speech interfaces may cause therefore usability problems.

The same user may be an expert user of some speech interfaces but still a novice user of other systems. Diverse speech technology solutions will require different interaction strategies from the user and thus, the use of several different cognitive models [9]. When encountering unfamiliar services, the user will be forced to find out which cognitive model is the most appropriate one to use. An incorrect cognitive model may still be chosen resulting in unexpected and problematic user behaviors.

Exposing the users to environments with diverse speech interfaces will result in increased cognitive load and thus the opposite of "calm computing".

3.2 Usability Requirements

In mobile and dynamically changing UC environments the user's intentions and needs may rapidly change. The user should be able to initiate a new task while waiting for some other specific service to be completed. It should also be possible to easily cancel a previously issued command or change the parameters of some previously initiated service. Furthermore, the system itself should be able to interrupt an ongoing dialogue and direct the user's attention to higher priority events taking place in the user's immediate environment. Thus, *asynchronous dialogue management* [1] should be supported.

For supporting a wide range of domains within one and the same dialogue and for allowing the user to transparently and seamlessly switch between several topic domains and services a *multi-domain approach* [4] is also necessary. However, the support for these features in current industry solutions is limited.

3.3 Multiple Concurrent Speech Interfaces

As far as we know, the effects of using several concurrent speech interfaces at the same time have never been studied. This situation may actually occur in UC environments, when several embedded interfaces are listening for user commands, or even taking initiative pro-actively. Due to missrecognitions, it is possible that several speech interfaces may be triggered by a single user utterance.

Using straightforward solutions, such as naming, will inevitably result in an increased cognitive load on the user, who must remember and use the proper names. Finding out the right names might also be a problem in unfamiliar, or seldom frequented, environments.

Consequently, to provide unobtrusive and user friendly speech interfaces in mobile and UC environments and to avoid the introduction of new usability related problems we need means to coordinate and control the various speech interfaces.

4 User-Centered Architecture Model for Speech Interfaces

The currently successfully employed speech interface architectures for desktop or telephony based interaction all share an application-centered multi user system design, see Fig. 1A. In the fast growing telephony and VoiceXML based voice portal applications[2], one single speech interface is used for accessing a multitude of services, see Fig. 1B. These architecture models cannot easily support solutions for the above discussed usability problems and thus are inappropriate for the mobile and UC environments.

With VoiceXML a new way of developing speech applications has been introduced. The standardized markup language shields the developers from low-level implementation details and thus facilitates rapid application development. However, the dialogue management is still application-centered and the desired "calm computing" is not supported. The voice portal based solutions are not appropriate neither for controlling consumer appliances nor the user's environments. Interacting with the coffee-pot in the kitchen through a telephone call is not the most natural solution.

We suggest a user-centered, application independent architecture for speech interfaces, see Fig. 1C, for mobile and UC environments. Thus, every user is expected to use a *SINGLE*, highly personalized speech interface to access all services and appliances. We believe that a user-centered *single user and multiple application* speech interface would be an appropriate solution to remediate the usability problems discussed in the previous section and will also support personalization, context awareness and user modeling.

It is possible to implement the user-centered speech interface as a back-end server based solution. However, it would be preferable if the user-centered speech interface could be integrated into some personal, wearable appliance such as a mobile phone or a PDA. In that case, the speech interface would always be accessible with all user-dependent data activated and ready to use.

The domain dependent data, such as knowledge-models and dialogue descriptions could be developed in a similar way as in the voice portal solution and they

[2] See: speechWeb from PipeBeach or Voice Web from Nuance:
(http://www.pipebeach.se/speechweb_product_description.pdf)
(http://www.nuance.com/partners/voiceweb.html)

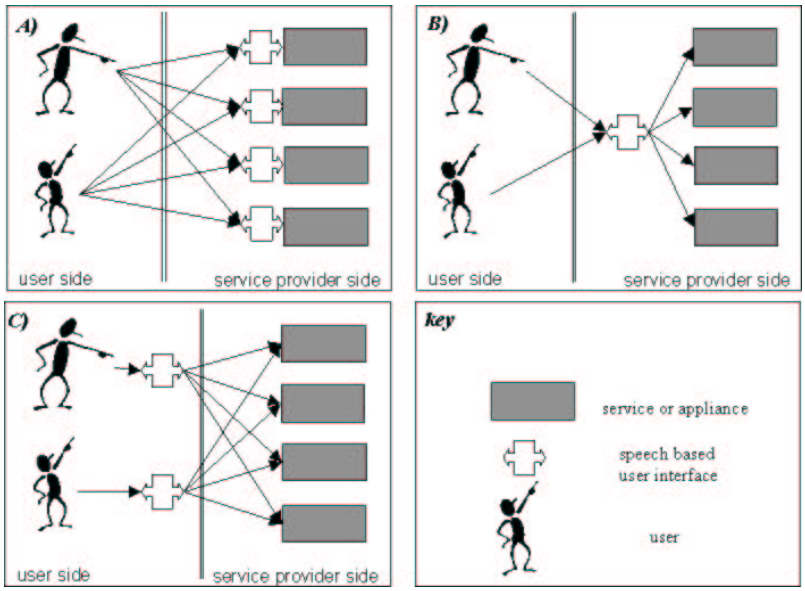

Fig. 1. Speech interface architecture models: *A)* Embedded and application-centered speech interfaces. *B)* Voice portals - application-centered, centralized speech interfaces. *C)* User-centered and application independent speech interfaces.

should be stored locally on the service provider side, in the appliances. These service descriptions could be loaded dynamically into the personalized speech interface through some ad-hoc and wireless communication solution, such as, e.g. Bluetooth[3], when they are needed in the appropriate context and environment.

In addition to the service descriptions some storing capacity is all that is needed together with some wireless communication capability on the service provider side. Thus, the user-centered model would offer a more cost effective, simple to develop and maintain and computationally less demanding solution for the industry.

4.1 Speech Processing Advantages

By applying the user-centered architecture, the impact of some major challenges for spoken dialogue systems [6] can be reduced. Through the possibility to use speaker dependent and adaptive speech recognition the *speaker variation* can be reduced significantly. In this way the amount of speech recognition errors could be decreased substantially and unnecessary dialogue management problems could be avoided. Due to the adaptation of the speech recognizer to individual users, all services and appliances can be made available to all users, in spite of dialects, non-native accents etc. The time necessary to train the system will

[3] http://www.ericsson.com/bluetooth/

also be minimized. The variability in *channel conditions* could also be reduced through consistent use of a personal headset for audio I/O.

The variability in spontaneously spoken language led to an increased interest for *adaptive solutions* and frameworks [18] in spoken dialogue systems. However, for successful adaptation, extensive data collections are required. In UC environments some interactions may be very short, actually so short that recognition of the users characteristics and the adaptation to them may be impossible. On the other hand, adaptation under such conditions would not be a problem if a user-centered speech interface was used for accessing all available services.

4.2 Ubiquitous Computing Related Advantages

A user-centered solution would also be ideal for gathering data on the user's behavior and speech for building *user models*. It is suggested [10] that user models would be a great help in dialogue management to predict user intent and tailor the interaction to an individual user. Extensive usage of user and domain knowledge models could provide support for *context awareness* [5] and thus facilitate the prevention of misconceptions by the user and would increase the system's cooperativeness.

Due to the user's familiarity with the personalized speech interface, the problems with diverse and concurrent speech interfaces could be avoided and the user satisfaction would be quite reasonable. It would also be unnecessary for the users to adapt to several different interfaces. Accordingly, *calm computing* is facilitated by the user-centered approach.

Collecting data on the user's behavior, speech patterns etc. is a delicate issue. We believe that a single user-centered interface, because it is controlled by the user, provides better *security and integrity* features than a multitude of different embedded and distributed systems, which are outside the user's control.

Due to an expected frequent use, and almost continuous availability of the speech interface, it is even possible to use the interface to perform text independent and continuous speaker verification [3]. This feature would also increase security and integrity.

4.3 Open Questions

The variability of the acoustic environment may be disadvantageous for the accuracy and robustness of speech recognition, and thus, for the dialogue management. The current acoustic environment has to be identified and a suitable robust speech recognition solution applied. Robust speech recognition is a very active research area.

Designing a generic, user-centered, adaptive spoken language dialogue system gives rise to some major research questions. In the next section we will present a dialogue manager, which is our first attempt to solve at least some of them.

5 SesaME: a Generic, User-Centered Dialogue Manager

We assume, that most of the tasks users may want to perform through speech interfaces in mobile and UC environments are simple and concrete tasks. The length of the interactions is expected to be relatively short. Based on these assumptions, we believe that a task-oriented dialogue manager is an appropriate implementation for the user-centered speech interface architecture model.

Taking into consideration the usability issues discussed above, we have designed and implemented SesaME, a new generic, task-oriented dialogue manager. It is designed to be used in mobile and UC environments. Special attention has been given to support adaptive interaction methods and user modeling.

The architecture of SesaME is modular and extensible. Thus, the functionality of SesaME may be gradually extended. SesaME relies on, but is not dependent of, the Atlas generic software platform for speech technology based applications [13]. The Atlas platform provides high-level primitives for basic speech I/O, but access to low-level data is also facilitated.

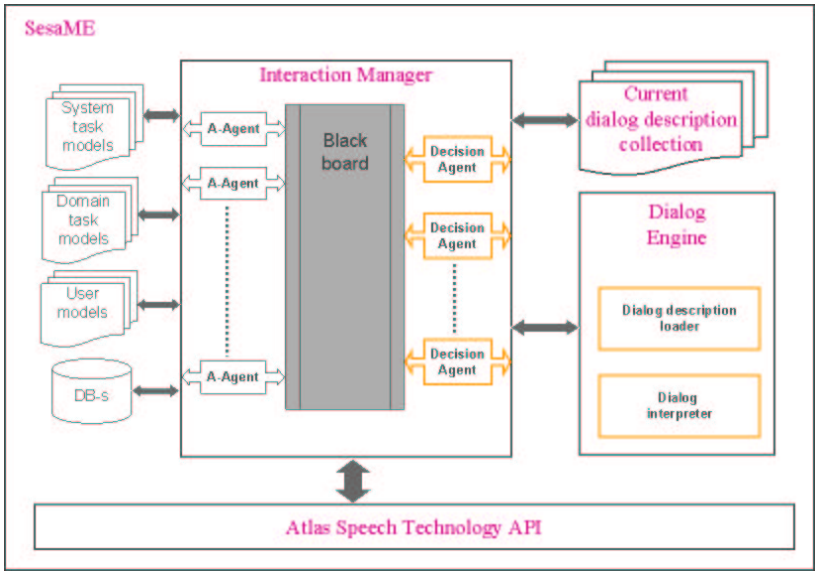

Fig. 2. SesaME system architecture.

The main components of SesaME, see Fig. 2, is the *Interaction Manager* (IM) and the *Dialogue Engine* (DE). They are described in detail below. The division of dialogue management into two parts is a major feature of SesaME. The aim is to separate application dependent dialogue management from the application independent general dialogue feature management. Application dependent data and domain task knowledge is handled within the IM, but in a general and application independent way.

5.1 The Dialogue Engine

The main task of DE is the interpretation of the user's last utterance, and planning and performing the system's next utterance. This task is performed using a straightforward slot-filling method. Based on the results from the linguistic analysis, several slots may be filled during one single turn. The IM may suggest some eligible values for one or more appropriate slots. These suggested values may be directly filled in the appropriate slots or used to ask a better constrained question to the user.

The choice of the dialogue descriptions formalism has a crucial effect on the dialogue management. However, at the first development stage we have chosen not to focus on these issues. While designing and developing SesaME, we have chosen, mainly for practical reasons, to use a subset of the VoiceXML standard for dialogue descriptions.

5.2 The Interaction Manager

The interaction manager takes care of the application independent tasks of the dialogue manager. The most important duties performed by the IM is to supervise the interaction with the user and to manage various knowledge models. The major tasks carried out by the IM are:

- **Identifying the appropriate dialogue descriptions** that should be activated.
- **Supervising the DE**. Such as managing time-outs etc.
- **Suggesting appropriate values for the DE to use.** This is done based on an extensive use of the context and the available user models and domain knowledge models.
- **Supervising the interaction with the user.** The aim is to detect end evaluate general dialogue management phenomena such as "no new information supplied by the user during last turn", etc.
- **Performing error management and error prevention** is one of the major task of the IM. The currently ongoing dialogue flow may be interrupted and some error handling dialogue descriptions may be activated.

The IM is composed of a collection of autonomous agents, each one taking care of one well-defined atomic task. The communication between the agents is done through a central blackboard where a *subscribe and notify* mechanism is used. Agents may perform some plain information refinement or they may generate some hypothesis about some specific detail in the current dialogue. The evaluation of the delivered hypotheses and the coordination of the IM's overall work is carried out by the *decision agents*.

5.3 System Evaluation

The SesaME dialogue manager is still under development and it will be tested and evaluated within the framework of the PER (Prototype Entrance Receptionist) project [15] as a traditional stand-alone, multi-user and application dependent dialogue manager.

Evaluating SesaME as a personalized and generic dialogue manager for mobile and UC environments is difficult due to the lack of the infrastructure for mobile, wireless and context aware services. Two alternative solutions are: to either evaluate SesaME within some experimental framework such as the *Locust Swarm* [17], or to use a simulated environment, such as *QuakeSim* [2].

6 Conclusions and Future Work

In this paper we have discussed some usability and HCI related problems which may arise when speech interfaces are integrated in mobile and UC environments. Based on these issues a novel user-centered architecture was proposed for speech interfaces. We suggested that each user should use a single personalized speech interface for accessing a multitude of appliances and services in these new environments. Finally, SesaME, a new generic and task-oriented, dialogue manager, built according to the user-centered model, was introduced and described.

One of the issues not discussed here is multimodality. Users may prefer to combine voice input with direct manipulation and graphical user interfaces. How speech should be mixed with other modalities and seamless switching across modalities require further research.

The suggested architecture creates novel possibilities for supporting personalization, context awareness and user modeling in dialogue management. We believe that the usability and HCI related advantages are interesting enough to make the proposed architecture and thus the presented dialogue manager worthy of further test and evaluation.

7 Acknowledgements

This research was carried out at the CTT, Centre for Speech Technology, a competence center at KTH, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organizations.

References

[1] J. Boye, B. A. Hockey, and M. Rayner. Asynchronous dialogue management: Two case studies. In D. Traum and M. Poesio, editors, *Proceedings of GÖTALOG 2000*, Gothenburg, Sweden, June 2000.

[2] M. Bylund and F. Espinoza. Using Quake III Arena to simulate sensors and actuators when evaluating and testing mobile services. In *Proceedings of CHI 2001*, Seattle, USA, Apr. 2001.

[3] M. J. Carey and R. Auckenthaler. User validation for mobile telephones. In *Proceedings of 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, 2000.

[4] G. Chung, S. Seneff, and L. Hetherington. Towards multi-domain speech understanding using a two-stage recognizer. In *Proceedings of Eurospeech '99*, pages 2655–2658, Budapest, Hungary, Sept. 1999.

[5] A. K. Dey and G. D. Abowd. Towards a better understanding of context and context-awareness. In *Proceedings of the CHI 2000 Workshop on The What, Who, Where, When, and How of Context-Awareness*, The Hague, The Netherlands, Apr. 2000.

[6] J. R. Glass. Challenges for spoken dialogue systems. In *Proceedings of 1999 IEEE ASRU Workshop*, Keystone, CO, USA, Dec. 1999.

[7] J. Grudin. The case against user interface consistency. *Communications of the ACM*, 32(10), Oct. 1989.

[8] F. James, M. Rayner, and B. A. Hockey. Accuracy, coverage, and speed: What do they mean to users? In *CHI 2000 Workshop on Natural Language Interfaces*, The Hague, The Netherlands, Apr. 2000. http://www.cs.utep.edu/novick/nlchi/.

[9] L. Karsenty. Shifting the design philosophy of spoken natural language dialogue: From invisible to transparent systems. In *CHI 2000 Workshop on Natural Language Interfaces*, The Hague, The Netherlands, Apr. 2000. http://www.cs.utep.edu/novick/nlchi/.

[10] R. Kass and T. Finin. Modeling the user in natural language systems. *Computational Linguistics*, 14(3):5–22, Sept. 1988.

[11] J. A. Larson. Speech-enabled appliances. *Speech Technology Magazine*, November/December 2000. http://www.speechtek.com/.

[12] N. Marmassee and C. Schmandt. Location-aware information delivery with ComMotion. In *Proceedings of HUC 2000, Handheld and Ubiquitous Computing*, pages 157–171, Bristol, UK, Sept. 2000.

[13] H. Melin. ATLAS: A generic software platform for speech technology based applications. *TMH-QPRS,Quarterly Progress and Status Report*, 2001(1), 2001. To be published.

[14] J. Nielsen. *Usability Engineering*, chapter 5.4. Morgan Kaufmann, Inc., San Francisco, CA, USA, 1994.

[15] B. Pakucs and H. Melin. PER: A speech based automated entrance receptionist. In *NoDaLiDa 2001, Proceedings from the 13th Nordic Computational Linguistic Conference*, Uppsala, Sweden, May 2001. To be published.

[16] N. Sawhney and C. Schmandt. Speaking and listening on the run: Design for wearable audio computing. In *Proceedings of ISWC'98, International Symposium on Wearable Computing*, Pittsburgh, Pennsylvania, USA, Oct. 1998.

[17] T. Starner, D. Kirsh, and S. Assefa. The locust swarm: An environmentally-powered, networkless location and messaging system. In *Proceedings of the First International Symposium on Wearable Computers*, Cambridge, MA, USA, Oct. 1997. IEEE Computer Society Press.

[18] M. Turunen and J. Hakulinen. Jaspis - a framework for multilingual adaptive speech applications. In *Proceedings of 6th International Conference of Spoken Language Processing (ICSLP 2000)*, Peking, China, 2000.

[19] M. Weiser and J. S. Brown. Designing calm technology. *PowerGrid Journal*, July 1996. Available from: http://www.ubiq.com/hypertext/weiser/UbiHome.html.