Adapting Multimodal Dialog for the Elderly

Christian Müller
Department of Computer Science
University of the Saarland
cmueller@cs.uni-sb.de

Rainer Wasinger
DFKI GmbH
rainer.wasinger@dfki.de

ABSTRACT

In this paper, we outline the design of a multimodal interface for a mobile pedestrian navigation system developed within the project COLLATE. The aim of the interface is to adapt to different resource limitations of the user. It takes into account the cognitive load of the user as well as the age. We present an approach on how special acoustic models for elderly speakers can improve speech recognition quality and at the same time provide an information source for user modeling. Three different presentation strategies are presented: unimodal (speech only, graphics only), redundant (speech and graphics providing the same information), and concurrent (minimally overlapped speech and graphics).

1. INTRODUCTION

With systems getting more ubiquitous and mobile, designers are faced with the challenge of universal usability. This involves the need to accommodate for context and user diversity. The notion of context diversity covers areas such as different environments (indoor/outdoor) and different machines (desktop/pocket PC). User diversity refers to the problem that interfaces have to be designed to be usable by people with a wide range of needs and capabilities. One example of this diversity is the comparison between average aged adults and the elderly. Elderly people are one of the last groups to benefit from access to computers. What makes technology difficult for elderly people to use is that elderly people very often suffer from cognitive disabilities like age degenerative processes, motor impairments, short-term memory problems, and reduced visual and auditory capabilities (Jorge, 2001). These disabilities are often magnified by a person's unfamiliarity with the given technology and the different learning curves possessed by individuals. Although it is the elderly that have this increased load while performing everyday chores, average aged adults are often under the same conditions when they multi-task, for example driving a car and talking on a hands-free mobile phone. This provides us with a common ground between the two user groups. Users are also generally already overloaded when dealing with mobile devices (Oviatt & Cohen, 2000), and this again outlines the importance on designing interfaces that are simple and easy to use.

Speech recognition and speech synthesis can address this concern. Especially in a mobile context, speech is a more natural and intuitive way to control a system, bypassing pen and keyboard. On the other hand, speech is not always the optimal interaction modality, for example it is easier and more intuitive to referring to an object that is displayed on the screen than it is to describe it. The same holds for graphical output vs. speech. In case of a navigation system for example, the spoken output should be combined with maps, arrows, and similar graphical output.

In this paper we will outline the design of a multimodal interface for a mobile pedestrian navigation system. Based on the REAL mobile pedestrian navigation system (Wahlster, Baus, Kray & Krüger, 2001), (Mueller, 2002), COREAL (Collaborative Resource Adaptive Localization) combines both the indoor and outdoor components of this system to form a versatile and adaptive mobile pedestrian navigation system.

The system currently adapts to a user's limited cognitive resources, by modifying the displayed map information, according to both the user's walking speed and the GPS signal's location accuracy. When a user walks at a slow pace, they are presented with more information compared to when they walk fast. The GPS signal accuracy is denoted via a user's position marker, which decreases in size as the accuracy increases.

The system has grown with the technologies of its time, beginning its life under the REAL project on the PALM operating system and then being further developed on the more powerful Xybernaught. The current development platform is the Pocket PC, with the implementation concentrating on the Compaq IPAQ. A GPS receiver and a bluetooth capable mobile phone will be combined with the system. Outdoor localization is achieved through the GPS receiver, while indoor localization (where no satellite reception is available) is achieved through a number of Infra Red beacons fixed to the walls. The bluetooth mobile phone will provide the user with services available via the Internet such as bus timetables and the location of nearby taxi stands.

The system is currently aimed at average-aged adults. It is our objective to fulfill the needs of two user groups, young /average aged adults and the elderly. The system is currently also only unimodal. For the appropriate flexibility required in catering for these two user groups, a multi-modal interface will be designed. Output will take place in the form of speech and graphics, while input will take place through the use of a stylus, speech and a combination of both speech and stylus.Oviatt (2000) demonstrates that the processing of two types of inputs concurrently (speech and stylus) can substantially reduce the rate of user miss-interpretation.

2. ACOUSTIC MODELS FOR ELDERLY SPEAKERS

Current speech recognition systems have difficulty working with elderly voices (Wilpon & Jacobsen, 1996). To address this pro-

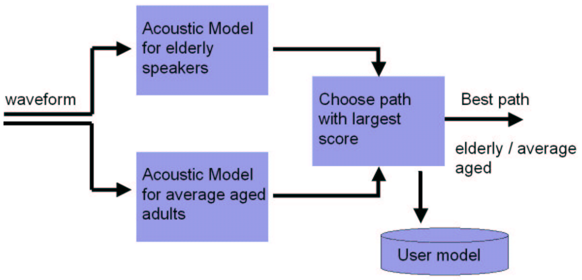
Acoustic Model for elderly speakers
waveform
Choose path with largest score
Best path
elderly / average aged
Acoustic Model for average aged adults
User model

Figure 1: Parallel account for speaker clustering

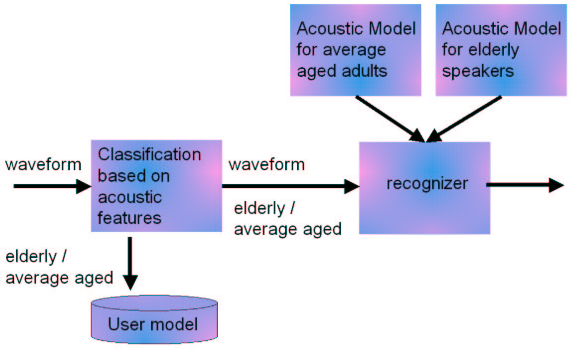
Acoustic Model for average aged adults
Acoustic Model for elderly speakers
waveform
Classification based on acoustic features
waveform
recognizer
elderly / average aged
elderly / average aged
User model

Figure 2: Serial account for speaker clustering

blem, specific acoustic models based on elderly speech are built within our project. The basis for the models is a large corpus of elderly speech that we will receive from ScanSoft (formerly Dragon Systems) for scientific purposes. Anderson et al. (1999) report that acoustic models built from elderly speech provide much better recognition than non-elderly models do (42.1 vs. 54.6 % Word Error Rate).

Besides the improvement of the speech recognition quality, the acoustic models could be useful for speaker clustering. Figure 1 depicts such an approach, where two acoustic models are used in parallel: a general non-elderly and a specific elderly model. On the basis of the findings in the literature, we expect the 'right' acoustic model to produce results with a higher accuracy (i.e. paths with a higher score). This information could be used as a hint that the speaker belongs to the elderly and non-elderly groups respectively.

Figure 2 depicts a serial approach, where relevant characteristics of the speech signal are extracted before the speech recognition process. After the group that the speaker belongs to is identified, the adequate acoustic model is used for speech recognition. To test this approach, further studies are necessary to identify a set of speech characteristics that are relevant for the discrimination of elderly and non-elderly and that are easy to extract from the signal.

In the context of a personal device, it is not crucial to infer the user characteristics from naturally occurring actions like speech input. The effort to provide this data by filling out a form is not too high, because the profile will be stored permanently on the device. The above described technology could nevertheless represent an elegant way of user modeling on public terminals, where the user is anonymous and interacts with the system only once. Consider for example a navigation system on the airport, where the user can receive path descriptions on a large screen and provide input via speech. In such a situation, user modeling should not take additional interaction steps. One of the aims of the system is to combine the personal device with public terminals.

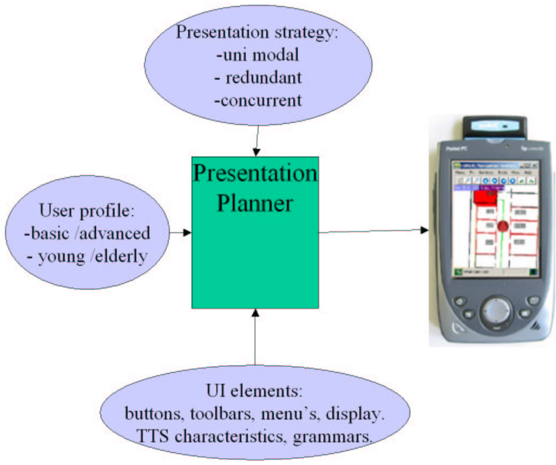

Figure 3: The presentation planner incorporates a given user profile, a presentation strategy and the user interface elements

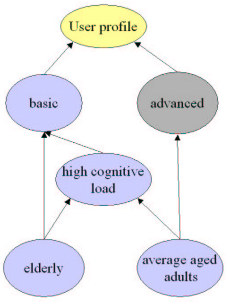

Figure 4: A stereotype graph of user profiles

3. A RESOURCE ADAPTIVE MULTIMODAL INTERFACE

The simplest and most elegant method of reducing the cognitive load of a person is to reduce the functionality of the overall system. To do this, a presentation planner will be introduced (see figure 3). It's aim will be to incorporate a given user profile, a presentation strategy and the user interface elements (speech and graphics) in an attempt to optimize the system to a user's cognitive load.

Two core profiles will exist, namely 'basic' and 'advanced'. The basic profile will provide greater emphasis on being robust and 'unbreakable', while the advanced profile will offer less frequently used functionality. Users will of course be able to change their profile. This will allow them to adapt the system to their own requirements and abilities as they become more adept with the software. Then, depending on the age and the cognitive load of the user, this profile will be extended through the presentation of the user interface components, both graphically and audibly as outlined below. A stereotype graph of user profiles is depicted in figure 4.

User interface elements can be categorized as being speech- or graphic-based. Speech-based elements include the acoustic models, the language models, and the synthesizer used for speech output. Graphic-based elements include the GUI components such as the menu, toolbars, buttons, the navigation map, and graphical navigation cues such as arrows.

The language models used in speech recognition will be optimized to the currently active functionality. This is determined by the base profile (basic or advanced). Speech recognition will also be used in an attempt to more easily befriend the system (Arafa, 1999). This will be achieved by affording the system with a personality, but only to the extent that commonly asked conversational statements specific to each user group be answered sensibly. Testing will be conducted to decide whether or not presenting an elderly user with vocabulary from the era that they grew up in or particularly enjoyed (for example the 1930's or perhaps the 1970's) would benefit the user in befriending the mobile device. These results will be incorporated into the systems synthesized speech output. An extension to this testing will see if the language used should also consider different social classes, for example the 'working' and 'upper' class as was common in Great Britain in colonial times, and the different forms of politeness as used in Germany today such as the 'du' and the 'Sie' forms. Elderly adults will receive speech output optimized to them in that it is spoken slower, clearer, louder and without elisions.

Directional commands such as left and right will be presented over the appropriate left and right audio channels, through an attached headset. This is considered to reduce cognitive load because directional information is obtained in an intuitive way. The channel selection will be switched off if the user is audibly impaired. Many non-speech audio events such as beeps and bleeps will then also be suppressed.

The GUI will leverage a 'what you see is what you get' interface. Much (if not all) of the menu system will be replaced with an additional toolbars, and similar to the speech adaptations for the elderly, the GUI will be clearer in that the toolbars, buttons, maps and text be displayed in a larger format.

The system will incorporate three different modes of presentation, namely a uni-modal mode, a redundant mode and a concurrent mode.

In the uni-modal mode, either speech or graphics can be selected as the input and output communication channel, but not both. It is expected that people with extremely reduced eyesight or hearing use this mode. Solely graphical output may also be chosen, if speech output is inconvenient, for example because of environmental noise. The speech only mode is optimal for hands-free-eyes-free situations, for example when the user is carrying luggage.

The redundant mode is the exact opposite of the uni-modal mode in that the same information is provided to the user over multiple channels, for example the system might say turn right in 100m and at the same time display an arrow pointing right with a 100m label attached to it. This would be used by people switching between modalities, for example navigating to a place while talking on the phone, or looking at a map while stopped at a set of traffic lights. Displaying redundant information has the disadvantage that it may increase cognitive load, and that it takes up more system resources. The advantage is that a user can choose the best way for them to receive information at any given time, reducing the effort needed to obtain the information in the first place.

The concurrent method is expected to be the default method. This method attempts to mould the graphical output with the speech output in the most natural way possible. In this case, information will be provided over both channels, but unlike the redundant strategy, there will be very minimal overlap between information. The audio will be used to present navigational directions such as 'Turn left on Mainzerstrasse after 15 meters'. The visual will present the navigation map, with the users path and position marked, but without large directional pointers. This method leverages the advantages of both modes, for example the ability for graphics to present a concise overview of information in a confined time frame and the ability for speech to provide for more flexibility when multitasking, while simultaneously minimizing each modes disadvantages.

4. SUMMARY

We outlined the design of a multimodal interface for a mobile pedestrian navigation system that adapts to different resource limitations of the user. It takes into account the cognitive load that differs depending on the complexity of the environment, walking speed, and tasks performed in parallel. In Addition, two user groups are discriminated between: average aged adults and the elderly. We presented an approach on how special acoustic models for elderly speakers can improve speech recognition quality and at the same time represent an information source for user modeling. Three different presentation strategies were presented: unimodal (speech only, graphics only), redundant (speech and graphics providing the same information), and concurrent (minimally overlapped speech and graphics). In this project phase, the acoustic models for elderly speakers are built, and the approaches for speaker clustering are evaluated. In addition, the basic interface with combined speech and graphical interaction is implemented. The first prototypes will be ready to demonstrate at the workshop.

References

Anderson, S., Liberman, N., Bernstein, E., Foster, S., Cate, E., Levin, B. & Hudsin, R. (1999). Recognition of elderly speech and voice-driven document retrieval. In *icassp1999, Proceedings of the International Conference on Acoustics, Speech, and Signal Processing.* Phoenix, Arizona.

Arafa, Y. (1999). Engineering personal service assistants. In *Proceedings of the EV/NSF Workshop in Universal Accessibility of Ubiquitous Computing: Providing for the Elderly.*

Jorge, J. (2001). Adaptive tools for the elderly, new devices to cope with age-induced cognitive disabilities. In *Proceedings of the EV/NSF Workshop in Universal Accessibility of Ubiquitous Computing: Providing for the Elderly.*

Mueller, C. (2002). Multimodal dialog in a multimodal pedestrian navigation system. In *Proceedings of the ISCA Tutotial and Research Workshop on Multi-Modal Dialogue in Mobile Environment.* Kloster Irsee.

Oviatt, S. (2000). Multimodal system processing in mobile environments. In *Proceedings of the 14th annual acm symposium on user interface software and technology.*

Oviatt, S. & Cohen, P. (2000). Multimodal systems that process what comes naturally. *Communications of the ACM*, *43*(3), 45–53.

Wahlster, W., Baus, J., Kray, C. & Krüger, A. (2001). REAL: Ein ressourcenadaptierendes mobiles navigationssystem. *Informatik Forschung und Entwicklung, 16*.

Wilpon, J. & Jacobsen, C. (1996). A study of speech recognition for children and the elderly. In *Proceedings of the IEEE Internation Conference on Acoustics, Speech, and Signal Processing* (S. 349). Atlanta.