MULTIMODAL DIALOG IN A MOBILE PEDESTRIAN NAVIGATION SYSTEM

Christian Müller
Department of Computer Science
University of the Saarland
D-66041 Saarbrücken
cmueller@cs.uni-sb.de

Abstract This paper describes the outlines of a multimodal interface to a mobile pedestrian navigation system. First it provides a short description of the system REAL that consists of an indoor component (IRREAL) and an outdoor component (AR-REAL). Currently the interaction of both is limited to unimodality (manual input and graphical output). The remainder of the paper describes an approach on how to obtain a multimodal interface. On the input side we use a combination of pen and voice. On the output side the graphical path descriptions are augmented by synthesized speech. When planning the multimodal dialog we discriminate between two different user groups: average aged adults and the elderly. Our investigations are focused on how to optimize the system's behaviour to the special needs of these user groups. Finally, the paper briefly describes two approaches that are investigated within this project: using speaker clustering techniques to build up a user model and 3D audio spatialization as an additional navigation cue.

Keywords: mobile pedestrian navigation, multimodal interface, speaker clustering, 3D audio spatialization

1. The REAL Pedestrian Navigation System

The REAL mobile pedestrian navigation system (Baus et al., 2001) consists of two parts [1]. The first component is an indoor navigation system that uses small PDAs to display simple sketches of the environment transmitted via infrared. The second component is an outdoor navigation system that uses a small laptop in combination with a head-mounted display. A GPS system determines the user's current position and an electronic compass tracks their orientation. All graphical path descriptions are tailored to the cognitive limitations of the user (e.g. the walking speed, spatial familiarity and time pressures) and the technical constraints of the output device (e.g. display and computational

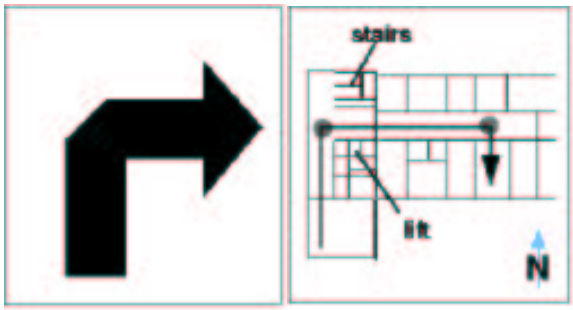

Figure 1. Graphical path descriptions in IRREAL

power). Adaptive services include the choice of camera perspective and path as well as the decision to include landmarks and interactive areas in the graphics. The limitations are also considered during the pathfinding process. Instead of choosing the shortest route, REAL tries to avoid complex redirections along the way at costs of a slightly longer route, and thus minimizes the additional cognitive load on the user.

1.1 IRREAL: The indoor navigation component

IRREAL uses handheld computers running PalmOS. A presentation server constantly provides the handheld device with data over strong infrared transmitters that are mounted on the ceiling. A unidirectional transmitting protocol was built where all the information is transmitted again and again in the form of cycles. IRREAL transmits interactive text and graphics, very much like hypertext documents. The presentations are tailored to the user's walking speed: the users will receive only information with high priority when they are staying for a short time in the transmitting area, because it is transmitted more often. When a user stays there for a longer time, more complex information about the environment will become available.

Beside the walking speed, the quality of the positional information also influences the system's output behavior. The less the system knows about the actual position and orientation of the user the more details of the environment are shown to lead the user in the right direction. The range goes from a simple arrow (when exact information is available) to a complete map of the environment (see figure 1).

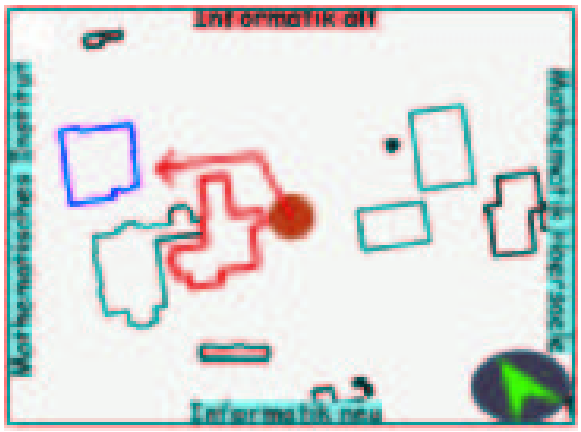

Figure 2. Graphical path description in ARREAL

1.2 ARREAL: The outdoor navigation component

The outdoor component of the pedestrian navigation system, ARREAL, consists of a subnotebook for the relevant computations, a clip-on set of glasses for graphical or textual output, and a small GPS and magnetic tracker to determine the users' position as well as their orientation in the environment. The magnetic tracker was modified and equipped with two additional buttons, so that it could be used to interact with the system analogously to a standard two button computer mouse that is also used as a 3D-pointing device. The user can for example retrieve additional information by pointing on a building.
The system supports different perspectives (birds-eye or egocentric-perspective) and different levels of detail in the visualization. ARREAL reacts to the changing quality of positional and orientational information in different ways. The egocentric-perspective for example is only chosen if adequate positional and orientational information is available. In other cases the system prefers a birds-eye-perspective, whereby the size of the dot that indicates the user's current position varies depending on the precision of the positional information. The system also takes the user's current walking speed into account. If they move fast, the system presents a greater portion of the map in order to help the user orientate themself and at the same time to reduce the amount of information about buildings on the edges of the display. Figure 2 depicts the output of the system when exact positional information is available (small dot) and the user is moving relatively slowly. See Wahlster et al., 2001 for a comprehensive description of the system.

Figure 3. Integrated version of the REAL pedestrian navigation system

2. Current limitations of the system

Recently, the indoor and outdoor components have been integrated into a single system. It consists of a wearable subnotebook equipped with an electronic compass, GPS and infrared receiver, and a portable touch screen for interaction. But the components are still quite big and uncomfortable to wear (see figure 3). The interaction is restricted to a single modality. The user points on the touch screen and receives various kinds of graphical representations together with textual annotations. Furthermore the system does not take into account the needs of special groups of users, for example elderly people. We intend to develop a mobile pedestrian navigation system that runs on a handheld device and provides a multimodal interface. In addition, we try to optimize the system's behaviour for two different user groups: average aged adults and the elderly.
The remainder of this document is concerned with the issues of a multimodal interaction in the context of a mobile pedestrian navigation system with a special focus on elderly people as a potential user group.

3. The Input Side

We try to achieve a mulit-modal input, where pen-based and spoken input is combined. When talking about spoken input, the first question that arises is how to run a robust speech recognition component on a handheld computer. Todays PDA's are limited in computational power. Particularly, CPUs li-

ke the StrongARM processor used in the Compaq iPAQ lack floating point computation-power, needed when running a speech recognition engine. Several current aproaches like Mipad (Huang et al., 2001) avoid this problem by distributing the speech recognition in a client-server architecture. The handheld device preprocesses the signal while the actual recognition is made by a server that is connected through a wireless network. A pedestrian navigation system should not however be dependent on a persistent network connection to a server. There exist some speech recognition systems like Speechworks' Speech2Go Embedded Recognizer that run stand-alone on a handheld device. However these are mostly limited to relatively simple voice input commands like those used in command and control appplications. What is needed is the full functionality of systems like CMU's SPHINX where different language models can be dynamically loaded (see section 6).
Currently we are evaluating several speech recognition systems. Among them are HTK, IBM Viavoice embedded, and Logox.

4. The output side

The graphical output of the system will be augmented with audio (speech as well as non-speech sounds). With respect to the speech output we follow the approach of a limited domain synthesis (Black and Lenzo, 2000), that tries to make use of the relatively small amount of potential utterances to optimize the quality of the synthesized speech. The most common utterances of the system are fully recorded prompts whereas rather rare cases are diphone-synthesized. In this way we combine the high quality of prerecorded prompts with the flexibility of a diphone synthesis. A key aspect of building a limited domain synthesizer is the design of a prompt list that adequately covers the domain. Ideally the information about the frequency of use should also be available. Before building up voices we need to undertake studies about what utterances the system will output in this special domain.

5. Issues of multimodality

The before mentioned application raises some interesting questions on multimodal dialog planning. Beside the existing adaptation strategies, we will have some additional aspects that raise the need for an adaptive system behaviour.
There will be situations when the interaction is solely spoken, e.g. when the user doesn't have the possibility to look at the device. Consider a scenario, when the users tries to achieve a path description while they are carrying heavy luggage. They may have to interact with the system while the PDA is in their pocket. Even when they are carrying the PDA in their hands, they may not be able to use the pen or look at the graphics, because they are running too fast or they have to look at the ground. A further reason to provide solely

spoken output may be a strong preference of the user, e.g. due to a visual impairment.
On the other side, there will be some occasions when no spoken input and output is used. When the environment is too noisy for example, audio output is of no use. We should also consider that a navigation system will be used in a public place. Speech should be avoided when the information is confidential, or simply out of courtesy to others. Similarly, the user may prefer solely graphical output. Adapting the interaction modality to the users preferences will be an important aspect for the acceptance of such a system.
The combination of graphical and auditory modalities can be achieved in several ways: alternative, complementary, or redundant. While planning the multimodal dialogue we will have a special focus on optimzing the interaction for two different user groups: average aged adults and the elderly. In cooperation with gerontologists we are developing adequate strategies to optimze the dialog for the different cognitive resource limitations of these two groups. While redundancy of graphical and auditory output may seem adequate to cope with resource limitations, there is evidence that it may confuse people, especially the elderly.

6. Speaker Clustering for user modelling

Some approaches of robust speech recognition try to analyze the voice of the speaker to rapidly adapt the recognition process accordingly (Hazen, 2000). The underlying idea is to improve the quality of the recognition by loading specific acoustic models. One method that has been proven successful is hierarchal speaker clustering. Similar training speakers are clustered to create models which represent specific speaker types. The construction can be performed using unsupervised bottom-up clustering based on acoustic similarities, or a supervised top-down clustering. In our case, a very simple cluster tree is created in a supervised top-down fashion (see figure 4). Depending on the amount of training data that is available, the models can be more or less specific. Large clusters are more general but can be trained more robustly. Smaller clusters can represent more specific speaker types but may lack a sufficient amount of training data. In addition to a better recognition quality, we expect to be able to use this information for acquiring a user model. Once the speaker is recognized as belonging to a specific cluster, lets say an ELDERLY WOMAN, the system can adapt it's multimodal dialogue strategies respectively. Although further sources of information about the user may be taken into account, this is an elegant way of solving a well known problem in user modeling: users don't like to provide the system with information about themselves before the actual dialog begins. With the speaker clustering approach, the system could adapt to the user after the first instances of the natural occuring dialog.

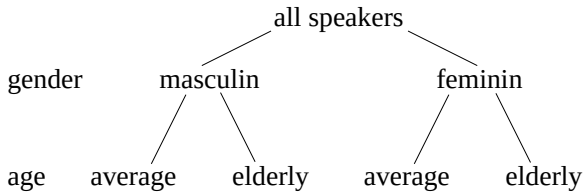

Figure 4. Hierarchical cluster tree that could be used for user modeling

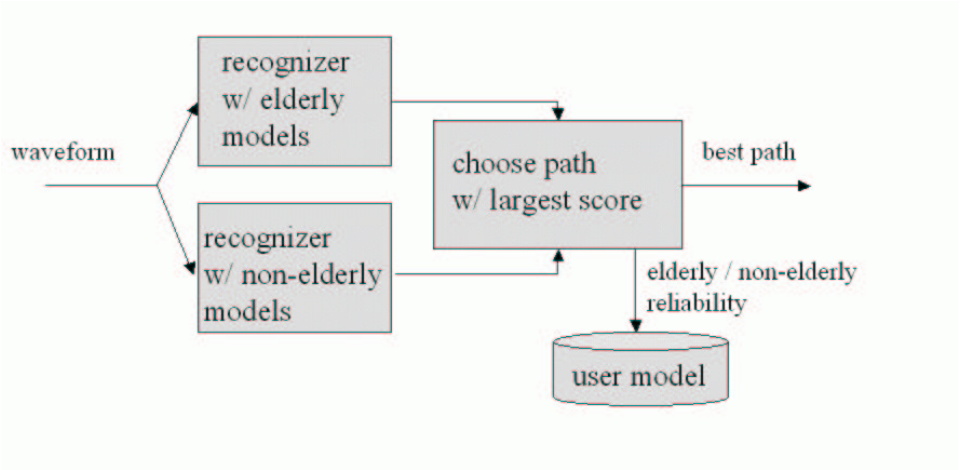

Figure 5. Diagram of a parallel age dependent recognition system

The possibilty to gain information about the user on the basis of their speech was empirically investigated previously. Müller et al., 2001 describe a set of indicators of cognitive load and time pressure that may serve as an information source for acquiring a user model.

Anderson et al., 1999 collected 78 hours of speech from 297 elderly speakers, with a average age of 79. They found that acoustic models built using the speech from elderly people provide much better recognition than that of those using non-elderly acoustic models models. The word error rates were 42,1% with the elderly models and 54,6models. Figure 5 depicts how this information could be used to classify the speaker. We can assume that a recognizer with the appropriate acoustic model will find a path with a higher score in most cases. Depending on the differences between the two *best paths* we can estimate the reliability of the classification.

Due to the limitations of computational power, such a parallel approach however is not appropriate for a standalone mobile setting. In this case it is ad-

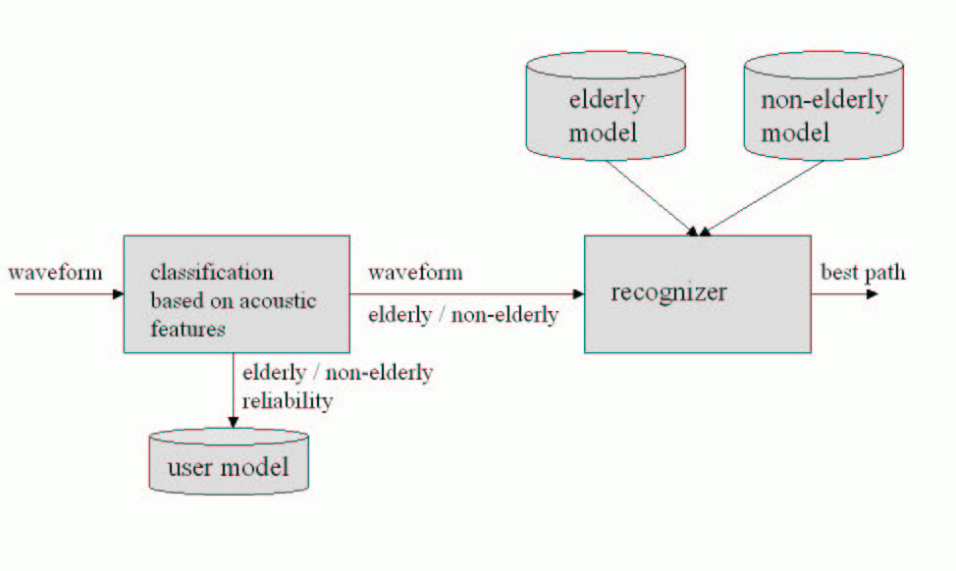

Figure 6. Diagram of a serial age dependent recognition system

visable to prepend the clustering from the actual recognition process. Figure 6 depicts an approach, where the classification into elderly/non-elderly is done on the basis of the acoustic features of the speech. The information is then stored in the user model and the recognizer loads the appropriate acoustic model to achieve a better recognition quality. The question that has to be addressed in this case is what kind of acoustic features may be significant for the classification. Speech rate may one of those features, because elderly people tend tend to speak slowlier than younger people. In addition, age related changes in the vocal tract lead to different vowel format frequencies (Xue et al., 1999).

7. 3D Spatialization as an additional navigation cue

The auditory output will consist of speech as well as nonspeech sounds. The latter can serve as navigation cues, especially when they 'point' to a certain direction. We are investigating the possibility of using 3D spatialization techniques to provide additional navigation cues to the user. In 3D spatialization, filters are used to place a sound within a virtual 3D environment. A regular stereo headphone can be used as an output device. The underlying idea is, that when the user perceives a sound as coming from a certain point in their environment, the information consumes less attention than a respective

description of the direction. The same technique can also be used for speech output (Goose and Möller, 1999). To place the sound at the right point in the user's environment high quality information about their position and orientation must be available. The real environment should also not be too noisy and the user must not be auditory impaired.

8. Summary

The outlines of a multimodal interface for a mobile pedestrian navigation system have been described. The work will focus on solving the described problems on the speech input side as well as developing a limited domain speech synthesis to combine with the graphical output. As mentioned before, we will also undertake empirical studies to develop adequate adaptation stategies for the multimodal dialog.

Notes

1. Actually it consists of three parts. The third part is a 3D-graphics workstation, where a virtual walk through the environment is shown by a virtual presenter. It is not considered here because we focus on the mobile parts.

References

Anderson, S., Liberman, N., Bernstein, E., Foster, S., Cate, E., Levin, B., and 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.

Baus, J., Kray, C., Krüger, A., and Wahlster, W. (2001). REAL: A resource-adaptive mobile navigation system. In *Proceedings of the International Workshop on Information Presentation and Natural Multimodal Dialog*, Verona, Italy.

Black, A. and Lenzo, K. A. (2000). Limited domain synthesis. In *ICSLP2000, Proceedings of the 6th International Conference on Spoken Language Processing*, Beijing, China.

Goose, S. and Möller, C. (1999). A 3d audio only interactive web browser: Using spatialization to convey hypermedia document structure. In *Proceedings of the 7th ACM International Conference on Multimedia*, Orlando, Florida.

Hazen, T. J. (2000). A comparison of novel techniques for rapid speaker adaptation. *Speech Communication*, 31:15–33.

Huang, X., Acero, A., Chelba, C., Deng, L., Droppo, J., Duchene, D., Goodman, J., Hon, H., Jacoby, D., Jiang, L., Loynd, R., Mahajan, M., Mau, P., Meredith, S., Mughal, S., Neto, S., Plumpe, M., Steury, K., Venolia, G., Wang, K., and Wang, Y. (2001). MIPAD: A multimodal interaction prototype. In *icassp2001, Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, Utah.

Müller, C., Großmann-Hutter, B., Jameson, A., Rummer, R., and Wittig, F. (2001). Recognizing time pressure and cognitive load on the basis of speech: An experimental study. In Bauer, M., Vassileva, J., and Gmytrasiewicz, P., editors, *UM2001, User Modeling: Proceedings of the Eighth International Conference*. Springer, Berlin.

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

Xue, S., Jiang, J., Lin, E., and Glassenberg, R. (1999). Age-related changes in human vocal tract configurations and the effects on speakers' vowel formant frequencies: A pilot study. *Logopedics Phoniatrics and Vocology*, 24:132–137.