PANPHONEN:
A SPATIAL ENHANCED AUDIO INTERFACE FOR BLIND USERS

F. Pittarello and A. Celentano

Dipartimento di Informatica
Università Ca' Foscari di Venezia
Via Torino 155, 30172 Mestre (Ve) Italia

ABSTRACT

In this paper we describe the model and the prototype of an auditory and haptic interface for blind users, based on human capabilities to memorize and retrieve positions in a 3D auditory space; the proposal takes advantage of recent development in hardware and software in a prototype system to test different paradigms for interaction suitable for wide application by a large number of users.

Keywords: Multimodal interfaces, haptics, speech recognition, virtual reality, visually impaired users, VRML.

1. INTRODUCTION

Recent advances in haptic and auditory devices have brought new opportunities for the evolution of personal computer interfaces; the new devices have been used mainly as an enhancement of the classic graphic interface for office automation tasks [3][5] or entertainment.

A parallel research line has involved the experimentation of different paradigms for visually impaired users or for people that for particular reasons can't use the classic graphic interface. In particular speech recognition technology is largely improved, even if the largest part of the experimentation is based on the model typical of the speech-automated phone services, where monophonic information is exchanged between the user and the system.

The main idea of this proposal is to experiment a different interaction model, based on the human faculty to memorize and retrieve the spatial location of information. We propose a 3D auditory information space controlled by users wearing a head mounted tracker for navigation and using a speech recognition engine to communicate requests and data to the system. We claim that the usability of this system is superior to the ordinary mono-dimensional audio interfaces, and that people using such a system should be facilitated to perform their tasks (read, communicate, etc.).

2. PREREQUISITES

There are a number of prerequisites we have considered in order to build a system really useful for blind users. We will briefly cite the most important before describing the architecture of the system:

- the system must be applicable to average personal computers, with modest integration of their basic configuration for hardware and software expansions, to grant a rapid diffusion of this interaction paradigm;
- the system must be made of components that use open or widely diffused standards to exchange information, to permit speed of development and to compare the usability of different configurations of components.

3. SYSTEM ARCHITECTURE

Figure 1 illustrates the basic architecture of the system. The core is the 3D audio environment that contains information and objects for interaction. The user navigates and retrieves information through the 3D world using a head mounted tracker. The feedback is a crucial part of the system and it is achieved using both a class of auditory non-verbal signs and dynamic verbal responses granted by a text-to-speech engine. The envisioned system architecture includes also a speech recognition engine as an additional means to communicate commands and information.

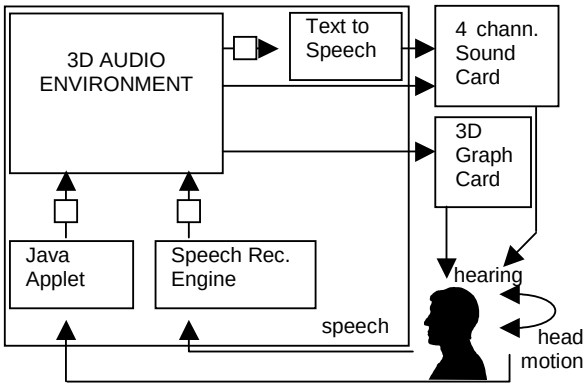

Figure 1. The *Panphonen* system architecture.

In particular the auditory feedback takes advantage of the auditory artifacts called *earcons* [2]; they were introduced by Meera Blattner et al. as an auditory counterpart of icons for graphic bi-dimensional interfaces; they were defined as *non-verbal audio messages used in the computer/user interfaces to provide information to users about some computer object, operation or interaction*.

Our work on usability and efficiency for 3D environments [7][8] [9] led to extend the use of these artifacts even in 3D scenes, as an additional support to identify the scene structure. The *earcons* used in our prototype system are derived from our previous research, even though in this case they had to be adapted being the main artifact for orientation in a 3D environment based only on audio messages.

A visual output of the 3D environment is given as an additional help for evaluators to control the intuitiveness of the system during the testing phase (for this purpose, each auditory object has also a geometric counterpart on the screen).

4. SPATIAL MODELS AND CONSTRAINED NAVIGATION

One of the main guidelines we followed for this project has been the simplification of the morphology of the spatial environment, being convinced by the experience of blind users in real world that a scene with too many features would have been an obstacle rather than a help to improve the efficiency in the task execution. So we decided to concentrate on three different basic typologies of spaces: the *wall*, the *corridor* and the *well* (see Figure 2). All these typologies share a fundamental feature: they can be easily represented using a low cost four-point speakers system.

The choice of modularizing the system and the adhesion to open and wide diffused standards helped us in the phase of building the prototypical environment. No manual coding or creations of custom tools were necessary; instead, common production tools for VRML (the language conceived to create interactive worlds on the net [10]) were used to design the spatial environments and the objects populating them. The result was exported to a VRML browser supporting a Java interface to exchange data with the rest of the system.

For what concerns navigation we chose not to take advantage of all the degrees of freedom offered by the chosen head mounted tracker. Instead, we decided to give the user the chance to perform only the movements necessary to reach all the available locations, in order to avoid disorientation and loss of vertical position in the virtual world [4].

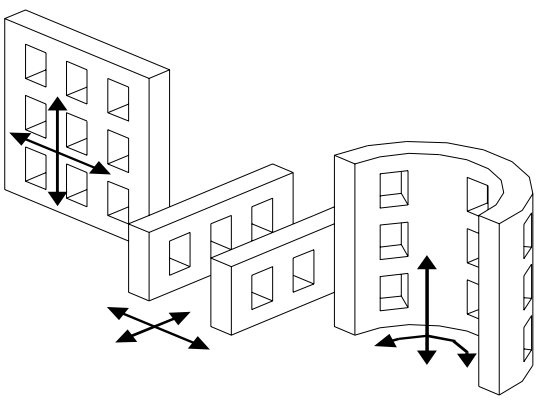

Figure 2. Three different spatial models for interaction.

We considered a number of different bindings between spatial configurations, head movements and navigation in the three typologies of virtual worlds.

The first morphology to be analyzed was the *corridor*; in this case the up and down movements of the user's head (monitored by the tracker) triggered, in the 3D world, movements along the main axis of the corridor; the left and right head movements triggered choices of items located on the two opposite walls of the environment. The main drawback of this association was related to the movement binding: e.g. the association of the horizontal rotation with the choice of items on the walls was not considered intuitive for the average user.

After that we considered the *wall* environment; in this case the user's head movements triggered movements along two orthogonal axis of the plane (vertical and horizontal); the analysis led us to discard this choice because bindings between rotations of the head and linear movements in the 3D world were not felt intuitive.

Therefore we considered circular environments (the *well* and the *sphere*) to have a more intuitive connection between the user's head motion and the 3D world. In particular the sphere environment gave us the opportunity to translate the rotational head movement into an analogous motion into the 3D world.

Even though the sphere morphology was felt as the most intuitive choice, there were additional issues to be considered; in particular, as will be evidenced in the following section, we wanted to give a special structure to the 3D environment, separating clearly different classes of activities.

That is the reason why we chose the well environment. In this morphology the rotational horizontal head movement were translated into an analogous rotation in the 3D worlds, but the up and down head movements were linked to shifts between different vertical levels of the well. We considered that this hybrid solution had the advantage to let the user to think explicitly to level shifts.

As stated before, the goal of the project was to take control of a 3D space in order to execute tasks; that is the reason why we called this prototype *Panphonen*, in memory of an analogous spatial configuration conceived in 1787 by the father of the Utilitarianism, Jeremy Bentham, as a means to take the full control of a circular real space (the *Panopticon* [1]).

5. A SOCIAL AUDITORY ENVIRONMENT

Another important choice for the project was to give a social connotation to the environment, trying to reflect the user need for privacy or social communication in the space organization.

So, in the case of the *Panphonen* prototype, we decided to divide the environment in levels (see Figure 3), corresponding to different classes of activities. In particular, thinking of a final system connected to the network, the prototype had three different plans:

- the highest level, conceived for social tasks such as collaborative real-time discussions, chat, phone calls;

- the intermediate level, reserved for intermediary forms of communication, such as e-mail management or similar tasks involving non real-time communication with other people;
- the lowest level, reserved for private needs such as private diary, repository for personal thoughts, etc.

Naturally the division in levels could have been done accordingly to other principles, for example sorting activities in alphabetic order or in reference to execution priorities.

Nevertheless, the main idea that convinced us to divide the *Panphonen* environment according to a social order, was that privacy is one of the most sensible factors to be considered when any user performs a certain activity. In general, this factor can be hardly controlled by blind people causing often a sense of insecurity during the interaction; this is the reason why we decided to give a priority to this issue, supporting privacy even through the organization of the 3D environment in *social levels.*

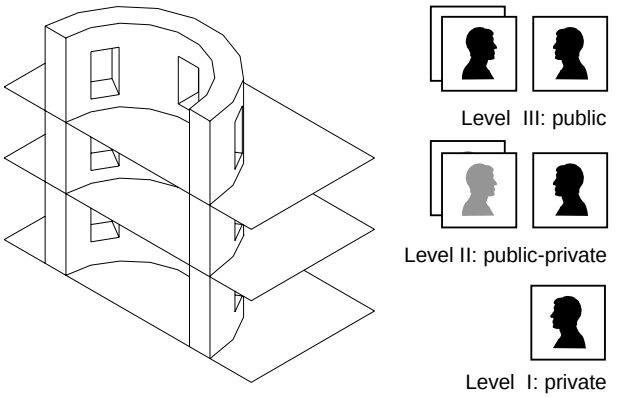

Figure 3. Socialization and spatial configuration.

6. USING EARCONS FOR NAVIGATION

In our previous works we defined the concept of *interaction locus* as a means to help the user to recognize the main parts of a 3D scene characterized by a specific content or a certain category of interaction mechanisms. The concept was introduced using a multimodal approach, useful to overcome the limits of vision for orientation in 3D worlds [7][9].

The basic idea of the *interaction locus* concept was to identify a specific part of the scene as a coordinated summa of 3D representation, auditory signs and hypertextual information. In particular, for what concerns the audio component of the *locus* concept, we proposed to use an audio artifact (*earcon*, see above) composed by a looping audio fragment. In some implementations we used *earcons* that were aesthetically pleasant [9], while in other occasions we used very raw audio loops. In spite of these differences, the main point is that in all the situations *earcons* worked as position markers, audio artifacts that reminded the visitor of the places he was walking in.

Users testing gave us a confirmation that these audio artifacts were very useful as a means to let the user to orientate in digital spaces; we are currently working at enhanced versions of these components in order to denote not only the individuality of particular places but also the category [8].

The positive response of the users was the reason why we decided to experiment these artifacts in a 3D audio-only environment and to use the *locus* audio component as the main feature to let the user to perceive the scene structure.

In particular, in the *Panphonen* prototype we associated a specific background loop to each social level, as defined in the previous section. In this way the user, that uses audio as the main means to communicate with the outer world, is constantly informed about his position and consequently about the social level of his current verbal communication. This is extremely useful not only for user orientation itself, but also, for example, for preventing the user from communicating involuntarily private thoughts in the public level.

Our previous works on *earcons* was not limited to the definition of the category of audio artifacts we have just examined. What we tried to do was to define specific classes of audio components useful for different situations [7].

In particular in our scheme we defined a class of artifacts called *Audio Arrow*; a sound fragment that has a spatial position and, when activated by the user's proximity, moves along two spatial coordinates to suggest or underline the user's direction characterizes this class.

In general an additional feedback about the user motion can be useful in 3D worlds. In the *Panphonen* environment, communication from the system to the user is achieved mainly through the audio channel. Therefore we felt that using an additional cue for the user could have been extremely useful in this specific context. That was the reason why we decided to implement an instance of this class also in the *Panphonen* prototype.

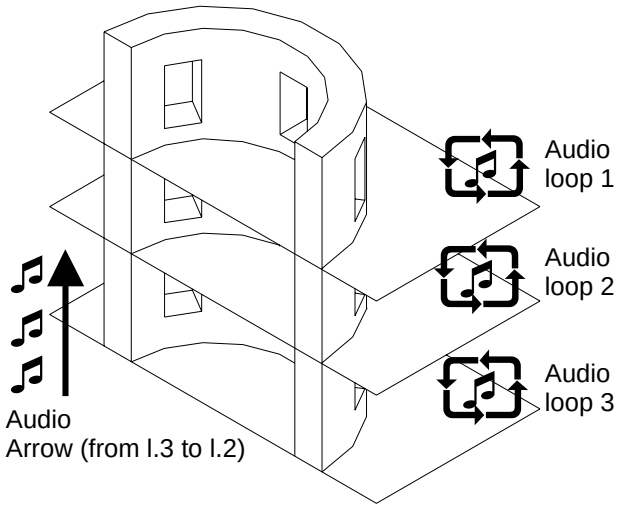

Figure 4. Audio artifacts for the 3D audio environment.

In our implementation, the user wearing the head mounted tracker moves from one level to another turning his/her head up and down; the confirmation of the level shifting is given not only by the change of the audio loop, but also by the *Audio Arrow* artifact that plays repeatedly the same note along the direction of the user movement.

Figure 4 shows the prototypical *Panphonen* environment with three social levels denoted by different audio loops; each loop can be heard only when the user enters a specific level. The icons evidenced on the left emphasize that each time the user shifts from level three to level two, the audio arrow artifact is activated, playing a sound fragment along the same direction followed by the user.

The work we performed so far on *earcons* is related to the task of navigating in the 3D audio environment; the resolution of this issue is crucial, being a prerequisite for the execution of more sophisticated behaviors such as the vocal management of an e-mail client and so on.

Our work on audio artifacts includes other classes that can be useful when the user is involved in activities different from navigation; the implementation of these classes in the *Panphonen* context will be part of our future work.

7. CONCLUSION

We have started a preliminary evaluation of the system described with a small group of users, acting as testers of this approach. A preliminary training was necessary to teach the users the basic movements to move in the 3D space and to recognize the meaning of the different *earcon*s; after this training the users were asked to reach particular positions in the 3D auditory well and generally this task was easily accomplished by the them. The association of the auditory stimulation to navigation and events improved gradually, in parallel with time spent in training.

A number of testers suggested to take advantage of the sense of touch to improve the usability of the system; the sense of touch is generally greatly developed in blind people and it is widely used by them to move and to perform tasks. Even if the speech recognition engine would be useful to perform certain tasks, the use of the sense of touch would allow them to perform tasks even in room populated by other people without disturbing or being disturbed. Besides, the combined use of touch and speech seems in certain cases less laborious than the continuous use of the voice.

Work performed so far has concerned the definition of the system architecture and a first set of tests to measure the usefulness of the different spatial configurations and the best bindings between devices and 3D environment; the *Panphonen* prototype will serve as a basis to experiment different models of speech interaction (the technical feasibility to compose speech engines with the current system has already been tested) and the interfacing with real world applications (e-mail systems, chat systems, etc.) in order to obtain a really useful system.

For what concerns the training period of the *Panphonen* prototype, an automated guided tour to the system functionality is of great importance, and will be one of the next developments. This tool will be useful not only for the preliminary training period but also as an auditory help to be invoked by the user any time he/she needs to be informed about the use of the system.

Besides, the use of complementary haptic devices will be experimented, as suggested by a number of users. In particular, as stated at the beginning of this paper, there will be a preferential choice for low cost devices that can be easily inserted in average personal computers, with modest integration of their basic configuration. In particular, the use of the mouse with tactile feedback [6], a widely available low cost device, seems one of the most promising artifacts to experiment in this context.

8. REFERENCES

[1] Bentham J., *Panopticon*, in a series of letters written in the year 1787, T. Payne, London, 1791

[2] Blattner M., Sumikawa D. and Greenberg R. “Earcons and icons: their structure and common design principles”. *Human Computer Interaction*, 4:11-44, 1989

[3] DragonSystems - speech recognition technology http://www.dragonsys.com

[4] Hanson, J. and Hughes S. “Constrained navigation interfaces”. Hans Hagen, editor, *Scientific Visualization*. IEEE Computer Society Press, 1999

[5] IBM speech recognition technology http://www.ibm.com/viavoice

[6] Logitech – IFeel Mouse http://www.logitech.com

[7] Pittarello F. “Desktop 3D Interfaces for the Internet Users: Efficiency and Usability Issues”. *Phd thesis*, Università di Bologna, 2001

[8] Pittarello F. and Celentano A. “Interaction locus: a multimodal approach for the structuring of virtual spaces”. To be published in *Proceedings of the VII SIE National Congress*, Florence, September, 2001

[9] Pittarello F., M. Pittarello M. and Italiano G.F. “Architecture and digital exhibitions - the Einstein Tower world”. *Virtual Environments 1998*, Springer Verlag, Wien/New York, 1998

[10] Web3D Consortium http://www.web3D.org