A 3D-Auditory Environment for Hierarchical Navigation in Non-visual Interaction

Kai Crispien and Klaus Fellbaum
Communication Engineering, Technical University of Cottbus—Germany
Anthony Savidis and Constantine Stephanidis
Foundation for Research and Technology Hellas

Abstract: This report describes the development of a generic, reusable spatial auditory environment, which combines a spatial auditory display derived by processing sound sources with head-related transfer functions, with gesture and speech-based input techniques. The spatial auditory environment is structured in a "ring topology", where the user is surrounded by a virtual "ring" containing auditory interaction objects. Auditory objects can be reviewed and selected by using 3D-pointing, hand gestures or speech recognition input, providing a multi-modal direct-manipulation method for the exploration of auditory interaction objects. Interaction processes are supported by different classes of spatial non-speech and speech audio cues that provide an immediate auditory feedback on different user activities. This auditory environment is used to implement a hierarchical navigation dialogue in a multimedia non-visual interaction toolkit that is currently under development.

Introduction

In the context of developing tools and methods to support non-visual interaction, recent research on auditory displays had a strong impact on the development of auditory-based non-visual interfaces, where audio is the primary direct perception channel. Aside from the improvement of synthesized speech quality and new findings in non-speech audio dialogue techniques using "auditory icons" and "earcons", the use of affordable spatial audio synthesis techniques derived by HRTF-processing has opened up new perspectives in non-visual interface design.

Spatial audio output can enable blind users to review spatially organized information and to navigate and explore spatially structured user interfaces. In a few cases, 3D-audio output techniques have been employed in implementations of non-visual interfaces. The MERCATOR (Mynatt et.al, 1994) and the GUIB (Textual and Graphical Interface for Blind People) project used 3D-auditory display techniques, though for different purposes: the MERCATOR project aimed at organizing the presentation of object hierarchies, while the GUIB project aimed at an immediate auditory transformation of the MS-Windows GUI, including spatial audio presentation of text (Crispien et. al, 1994) and spatial auditory mappings of graphical interaction objects and dialogue structures.

In the case of the GUIB interface, we have identified some problems and requirements which are in the focus of this work: (i) due to perceptual distortions, mainly caused by the use of non-individual HRTFs, the localization accuracy achieved with the spatial auditory display (especially for elevated locations) was far too imprecise to enable an immediate auditory transformation of the visual interface or to provide auditory-based direct manipulation with conventional pointing devices (e.g. a mouse), (ii) novel input interaction techniques—like hand gestures and speech input—have not been exploited, and (iii) the existing developments have been too highly specialized to support reusability of the implemented approaches and techniques in different application contexts.

The Spatial Auditory Environment

Addressing these problems, a multimedia toolkit for non-visual interaction is currently being developed which provides a 3D-auditory navigation environment. This environment will enable blind users to review a hierarchical organization of auditory interaction objects by using direct manipulation techniques through 3D-pointing, hand gestures and speech recognition input.

To design a spatial auditory user interface, appropriate to support a hierarchical navigation scheme, a "ring" metaphor was chosen as the conceptual model underlying the auditory interaction environment. According to this model, the items comprising a particular selection set are structured in a three-dimensional "ring topology", which surrounds the user. This topology is provided by structuring auditory interaction objects in virtual locations on a circular track in the horizontal plane around the user's head (Figure 1) through dedicated spatial audio processing. The locations of virtual auditory objects within the ring topology have been determined by considering perceptional constraints which arise, among other factors, from the use of non-individual HRTF's in the spatial audio processing system. Due to the fact that many psychoacoustic investigations of virtual auditory displays based on processing of non-individual HRTF's reported a decreased localization accuracy, especially for elevated and rear-horizontal locations, we chose a horizontal structure restricted to 12 objects with a spacing of 30°. Using this horizontal structure ensures a secure determination of auditory objects for the majority of users. Elevated locations are omitted in this navigation structure but might be used to spatially structure dedicated dialogue tasks which would require an additional selection space (e.g., copy-and-paste tasks).

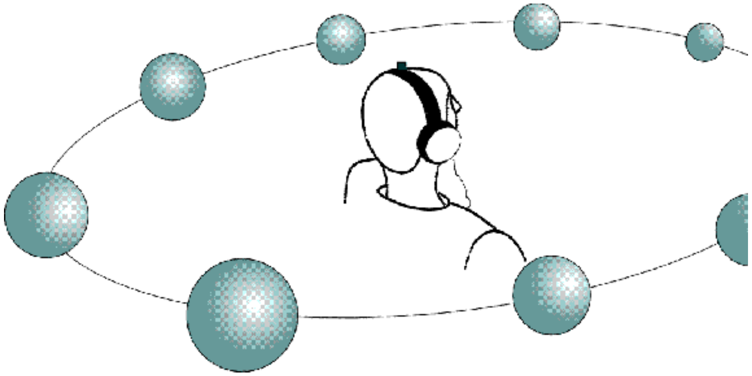

Figure 1: Illustration of the ring topology structure of the spatial auditory environment.

To avoid an overload of the user's auditory perceptive capabilities, we have restricted the number of simultaneously audible objects within a ring structure. Only those three objects are auditorily represented, which are parts of a virtual "auditory focus area" (Figure 2). This area covers an angle of 90°, derived from the actual frontal head direction of the user, which is dynamically scanned and updated by a magnetic position-tracking device residing on top of the user's head (headphone). If the user changes his "view point" or the ring structure is actively turned through dedicated input commands, objects which are no longer part of the focus area smoothly fade-out and the corresponding successors within the updated area fade-in. The use of the head-tracking device also provides the dynamic identification of a "focus object," by determining the collision of the head direction vector with a certain auditory object. Thus, the user becomes able to intuitively focus an object of interest, by simply changing his "view point." Input commands for the manipulation of an auditory object will automatically be dedicated to the selected focus object.

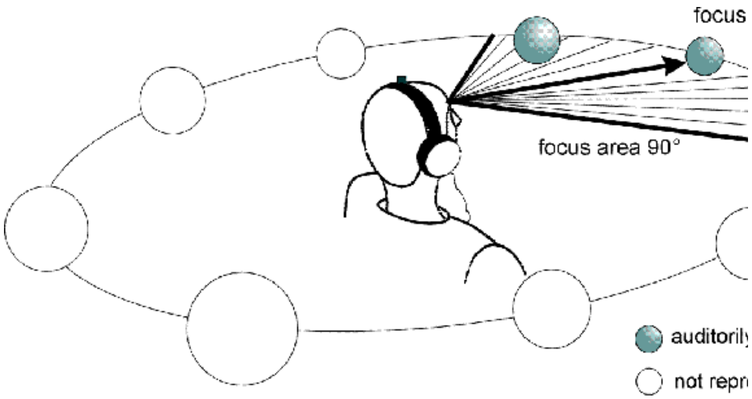

Figure 2: Conceptual illustration of the dynamically changeable auditory focus area, derived by processing head movements of the user.

Auditory objects can be selected and reviewed by dedicated input commands using 3D-pointing, hand gestures or speech input, which provides a multimodal direct-manipulation method for the hierarchical exploration of auditory interaction objects. All interaction processes are supported by different classes of spatial non-speech and speech audio cues, such as focus, modify, insert, delete, object and class cues, providing an immediate auditory feedback on different user activities. It must be noted that this report does not address the specific design of input and output commands (e.g. which sound cues, hand gestures or speech commands are appropriate for certain tasks) which have to be carried out by the user interface developer, using the toolkit under development.

Technically, the 3D-audio environment is realized with an "ACOUSTETRON 2" system from Crystal River Engineering Inc., connected to a SUN workstation in a client server topology. Currently two DSP boards are used, providing four simultaneously audible auditory objects (16 bit, 44.1 kHz). Speech recognition is carried out with a PC-based system, called "Speechmaster" from the German manufacturer Aspect. This system performs a speaker-dependent compound- word recognition that is highly reliable after performing several training sessions with a specific user. A "Cyberglove" pointing device, capable of recognizing hand gestures is used to provide pointing and hand gesture command input. The software architecture of the interaction environment consists of three main components: the ring server, the glove server and the voice server. These components operate asynchronously and provide the facilities summarized above. A communication protocol with these components has been designed and implemented, while user interface developers are provided with a logical programming interface for these components by hiding communication-specific details. The implementation of the spatial auditory output and speech input components was finished in October 1996. In order to finish the complete implementation, a harmonization with the glove server was carried out in November-December 1996.

References

Mynatt, E.D., & Weber, G. (1994). Nonvisual Presentation of Graphical User Interfaces: Contrasting Two Approaches. Proceedings of CHI'94, New York: ACM Press.

Textual and Graphical Interfaces for Blind People. The GUIB Project. Final report, publicly available.

Crispien, K., Würz, W., Weber, G. (1994). Using Spatial Audio for the Enhanced Presentation of Synthesised Speech within Screen Readers for Blind Computer Users. In Zagler, W., et al. (eds.), Computers for Handicapped Persons - Proceedings of ICCHP 94, Vienna: Springer Verlag.

Author Information

Kai Crispien and Klaus Fellbaum
Communication Engineering
Technical University of Cottbus
Germany

Anthony Savidis and Constantine Stephanidis
Institute for Computer Science
Foundation for Research and Technology
Hellas
Heraklion, Crete—Greece