A SPATIAL AUDIO USER INTERFACE FOR GENERATING MUSIC PLAYLISTS

Jarmo Hiipakka and Gaëtan Lorho

Nokia Research Center
Speech and Audio Systems Laboratory
P.O. Box 407, FIN-00045 NOKIA GROUP, Finland
[jarmo.hiipakka, gaetan.lorho]@nokia.com

ABSTRACT

This paper presents a user interface (UI) designed for non-visual interaction with a music collection. The UI system can be utilized to navigate a large list of musical items organized in a hierarchical structure, and to generate personal playlists. The interface only relies on stereo audio output and tactile input using a limited small of keys. This interaction scheme enables the use of the UI in situations where visual feedback is limited on the device, or when visual attention from the user is not possible or desired. However, the interface was also designed to be compatible with and complementary to visual UI concepts found in existing audio players.

The design and implementation of this spatial audio UI is described in the paper. Preliminary subjective tests are also reported regarding the immediate usability of this interface, when in 'eyes-free' mode on a handheld computer.

1. INTRODUCTION

Music playlist creation from a large music collection is becoming an important task for music listeners using digital audio storage formats. As a result of large storage capacity and efficient coding methods, it is common to find audio devices that contain a few thousand audio files. A visual user interface is usually used to handle such a large library of songs. Pauws *et al.* [1] also proposed a multimodal solution for interaction with a music collection. However, the problem has not been addressed for the auditory modality only, which could be of interest for devices with a very limited visual display size, or in the cases where 'eyes-free' mode of interaction is desired, e.g., when stereo headphones and a small remote control are available.

In this paper, the idea of spatial sound separation is applied to the presentation of a hierarchically structured collection of music files. The number of sound items used to represent the structure is intentionally limited to few positions for simplicity, but an efficient spatial organization of the items is used to provide navigation cues to the user.

2. UI DESIGN

This UI design is based on the idea of mapping between auditory and visual presentation of a hierarchical menu structure. Spatial sound separation offers a way to inform the listener about the current context, i.e., the relations between menu levels, and the interaction possibilities currently at hand. The sound reproduction system uses a simple technique to externalize sounds for listening comfort. In the following subsections, we review the approaches to non-visual presentation of UI first. Then, we describe the design of this spatial audio interface in detail, and discuss the potential problems seen with this UI.

2.1. Non-visual presentation of user interfaces

User interface design for auditory display has been considered in several research reports on auditory mapping of graphical UI and design for non-visual UI. The GUIB project [2], the Mercator project [3], and the audio window system by Ludwig *et al.* [4] illustrate the two different strategies available for auditory presentation of a UI. In the audio window and GUIB systems, a direct analogy to the spatial presentation in visual interfaces is proposed. The Mercator project utilizes another approach focusing on the presentation of object hierarchies in a UI, rather than a pixel-by-pixel level of the interface.

An auditory interface merging these two approaches has also been proposed with the idea of a hierarchical navigation system based on direct manipulation with 3-D Audio [5]. The use of a generic tree-based menu, addresses the problem of complex UI presentation, while the spatial audio presentation extends the scope of the interaction with the possibility of UI object focus and the concept of 3D pointing on a spatial ring structure.

The hierarchical menu presentation approach is well suited to the user interface considered in this study. Indeed, a music database can easily be organized as a folder structure, including, e.g., the menu levels *genre*, *artist*, *album* and *song*. The use of 3-D audio to support user interaction in this type of menu is also an attractive solution in our case. However, a simpler approach to spatial audio display was preferred.

2.2. Spatial auditory displays

3-D sound has been utilized in different types of auditory displays where spatial sound separation is desired. Sounds processed with head-related transfer functions (HRTF) are normally heard outside the head when replayed over headphones. They can also be placed virtually at any position around the listener. However, the reproduction of 3-D sound sources around a listener in an accurate, controlled and efficient way is technically challenging. For instance, the audio window system proposed by Ludwig *et al.* [4] relies on dynamic 3-D audio processing, which is computationally expensive and requires head tracking.

The concept of 3D audio ring presented around the listener has also been exploited in auditory displays such as the Nomadic Radio [6] project. More recently, Walker *et al.* [7] reported a similar approach for spatial mapping of temporal information.

However, when head-tracking is not available for 3D audio rendering, limitations such as front-back confusions and elevation problems are common with non-individual HRTFs. Goose and Möller [8] proposed an Audio Web Browser system, in which the display is restricted to a frontal arc, in order to avoid the potential problems existing with 3D audio rendering techniques.

The interface considered in this study is limited to the presentation of discrete items representing the nodes of the menu structure, which can easily be organized spatially with 3D audio. However, due to the limitations in sound localization with non-individual HRTFs, the auditory display uses a small number of sound positions presented along a line from left to right as shown in Figure 1. This ensures that the spatial sound presentation will work reliably for any user with normal hearing. An earlier study on the discrimination of multiple non-speech sounds showed that non-trained listeners could make a distinction between three or five positions [9]. For the current UI design, the auditory display was reduced to three sound items for simplicity. In the implementation, stereo amplitude panning is used combined with a simple stereo widening technique [10] to externalize the left- and right-most spatial locations.

2.3. User interaction: navigation vs. selection

The process of selecting songs in a music collection requires two separate tasks: navigating in the collection to search songs, and selecting songs. A music database is usually organized as a hierarchical menu, which includes, e.g., the levels music style, artist, album, and song. When creating a personal playlist, the user should be able to select any item from any level of the hierarchy. This motivates the separation between navigation in the menu and item selection in the interface. It should be noted that this distinction might not be intuitive to users. Indeed, simple menu systems are often based on the principle of item selection to go deeper in the menu structure, which is not compatible with the UI menu considered in this application.

2.4. Vertical tree and horizontal tree display approaches

In our interface design, the goal was to “emulate” the presentation of a two-dimensional menu structure with few spatially separated sound sources, based on a conceptual mapping between the visual and auditory representations of a tree structure. Restricting the spatial sound presentation to few items along the horizontal plane of the ears, allowed us to display only one dimension of the hierarchical menu structure. This limitation left us with two strategies for menu navigation, which we called the *item display* and the *level display* approach in our previous publication [11].

When the tree structure is considered vertically, as depicted in *Figure 2* a), three menu items from the same level can be presented at once, but only one level of the tree is audible. Spatial mapping of the menu navigation implies left/right keys for scrolling items of the current level, and up/down keys to move to another menu level. Alternatively, when a horizontal presentation of the tree structure is used, as illustrated in Figure 2 b), only one item is audible for each menu level, but three menu levels can be displayed at once. For spatial consistency, the left/right keys are used to change level now.

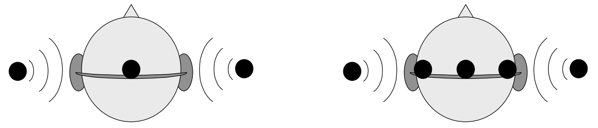


Figure 1. *Spatial presentation of three and five items along the horizontal axis.*

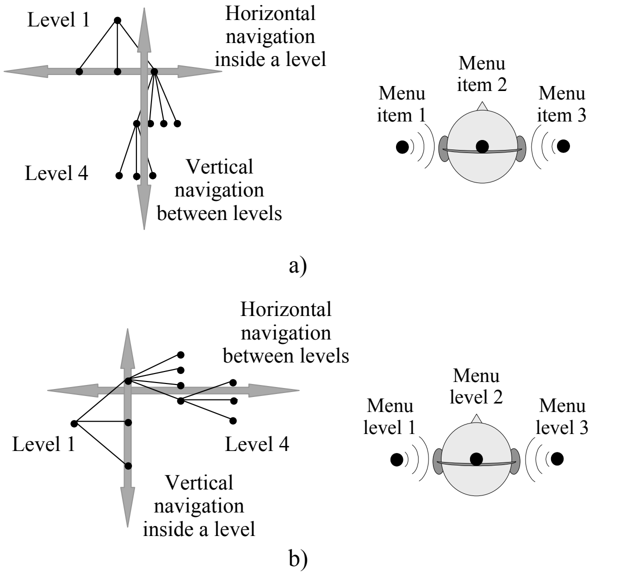
From these two strategies, the vertical tree approach is the most efficient for the display of simple lists of items. However, since visual menus are usually presented as vertical lists of items, we selected the horizontal tree approach in order to make this auditory interface compatible with a visual menu structure.

Another advantage of the horizontal tree presentation is the possibility to provide feedback on changes in several menu layers at once. This feature was exploited in our UI design, as we allowed continuous scrolling of items from different sub-menus of the same level, i.e. the user can always go directly to the next item (e.g., a song) on the same level, even if this item belongs to a different group (e.g., to a different album). This eliminates the need to move back and forth between levels when scrolling all the songs from one artist for instance.

2.5. Sound item playback

The spatial display does not limit the choice of the sound items in any way. Interaction feedback can be presented to the user with speech, non-speech sound, and music. Most of the information in the current application is textual. Therefore, speech is an important output means in this interface. However, an audio preview function of the songs is also implemented, and non-speech sounds can be used to enhance navigation feedback.

The spatially separated sound items can be played either consecutively or with a temporal overlap. For maximum instant usability, the non-overlapped replay was chosen. For advanced users, however, overlap between sounds could be a natural way to increase the interaction speed of the interface.


Level 1
Horizontal navigation inside a level
Menu item 2
Menu item 1
Menu item 3
Level 4
Vertical navigation between levels
Horizontal navigation between levels
Menu level 2
Menu level 1
Menu level 3
Level 1
Level 4
Vertical navigation inside a level

a)

b)

Figure 2. *a) Vertical tree, and b) horizontal tree approach for menu structure presentation using three sound positions.*

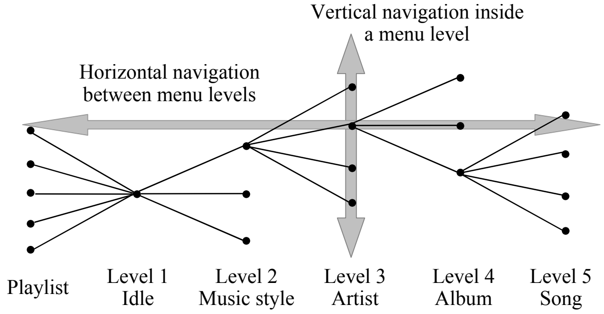

Vertical navigation inside a menu level
Horizontal navigation between menu levels
Playlist
Level 1
Idle
Level 2
Music style
Level 3
Artist
Level 4
Album
Level 5
Song

Figure 3. *Mapping horizontal and vertical movement to navigation in the hierarchy.*

2.6. Playlist creation

The approach employed for this audio user interface is shown in Figure 3. The user starts the playlist creation from the idle state. In this state, the personal playlist is on the user's left side, and the music collection is on the right. The user will hear "Playlist" on the left and "Music collection" on the right.

Moving right from the idle position, the user will arrive at the music style column, with the artist menu on the right. The user will hear "Music collection" and "Classical" (the name of the first style) from the middle position, and "Artist" from the right. The item presented at the right side provides feedback about the next level in the menu. Moving up and down, the user can now scroll through the list of items of the level positioned at the center, i.e. the musical styles available in the collection. By pressing the right key, the user can now move to the artist level, where "Artist" and the name of the first artist of the current style will be heard from the center and "Album" from the right. If the user crosses a border between two musical styles while scrolling through the artists, this change in the context is presented to the user from the left side. The same is true on all the levels of the music collection. At the song level, a preview of the current song is played on the right.

By pressing the middle key, the user may select any item positioned at the center, while moving in the collection. After selection, the item is added to the playlist and "Item added to playlist" is heard. When navigating in the playlist, the selection of an item will remove the corresponding song from the list, and "Item removed from playlist" will be heard.

2.7. Potential problems

As can be noted from the description above, the user has indication on neither the length of the current menu, nor the fact that the current menu is vertical. Therefore, it can be anticipated that users may not intuitively scroll the menu up and down. This should only be a problem in the purely auditory interface, though. In combination with a vertical visual menu system, the need for up and down scrolling would be evident.

Two other potential risks can also be identified in the interface: 1) The idea of direct breadth search and notification of the changing context may be confusing to the first-time users. 2) The feedback when selecting items to the playlist may be insufficient. Indeed, as there is no feedback mentioning that the playlist is on the left side the idle position, users may not easily find his/her way back to the playlist.

3. IMPLEMENTATION

The user interface design was implemented on a handheld Sharp Zaurus PDA running Linux. Prerecorded synthetic speech samples were used to guide the user in the hierarchical structure. The music collection consisted of CD-quality samples. The standard five-way key of the selected device was used for user input.

3.1. Real-time audio processing

We implemented an audio engine capable of mixing and processing several audio streams in real-time. The engine keeps a small portion of each relevant audio file in the main memory of the device so that the necessary files are always ready to be played. The rest of any of the longer files is streamed from the mass storage on demand. This makes audio feedback almost immediate, restricted only by the inherent scheduling latencies of the operating system.

In the engine, the audio processing is given its own higher priority thread, so that no dropouts in the output audio can be heard. All the parameters are also interpolated gracefully to produce a signal free of clicks and pops.

3.2. Text-to-speech synthesis

We used a low-complexity text-to-speech synthesizer to generate output for the hierarchical structure: the names of the artists, albums, and songs. In the implementation, the speech samples were prerecorded and stored in the device for playback on demand. The speech samples were generated with a sampling rate of 8 kHz.

For future implementations, the speech synthesis can be done in real-time even on small devices, and CD-Text information or the ID3 tags [12] commonly used in MP3 files can be used as speech prompts automatically.

3.3. Music collection

For the first implementation, we used a music collection of ca. ninety song excerpts. The collection contained songs in classical, jazz, movie soundtrack, and pop genres. The songs were stored in CD quality.

3.4. Interaction

The UI control was designed for four cursor keys for navigation and a separate key for selection. The selected device has a five-way key, so it was a natural choice to use that for user interaction. The touch-screen on the device was not used at all. All interaction events were recorded to a log file for later analysis.

4. PRELIMINARY SUBJECTIVE TESTING

We tested the proposed interface in a qualitative subjective test. At this point, the relative efficiency of the interface was not the primary concern, but we wanted to know, if the interface was intuitive enough on its own. Therefore, we tested the instant usability of the interface similarly to Pauws *et al.* [1].

The interface was designed to be compatible with a visual navigation system. However, in the first test we evaluated the interface in an eyes-free usage context. This way, we could better see, if the anticipated problems described in section 2.7 would surface in the test.

4.1. Test procedure

As an introduction to the test, the subjects were given a written general level description on the task of composing a playlist from a collection of music files. Additionally, the description stated that the feedback will be purely auditory, and that the input method used will be the five-way key of the device. The users were also encouraged to explore the interface as much as they liked.

The users were given five minutes to familiarize themselves with the interface. After the period had elapsed, the users were asked for a description of the features of the interface. The questions concerned the organization of the music collection, the audio feedback, and the functions of the five keys used in the UI.

After the familiarization phase, the users were given a more detailed description of the UI and were asked to perform three specific tasks. In the first two tasks, they had to answer questions related to the structure of the music collection, and in the third task, the users had to compose a playlist of six songs, from three different musical styles.

4.2. Results

The results from the immediate usability test are encouraging. Eight of our ten subjects could fully describe the structure of the music collection after the five-minute familiarization period. The other two could browse the collection, but they felt slightly lost, and afterwards could not describe the hierarchy of the collection.

All our subjects understood the key mapping used for navigation, and the difference between selection and navigation. Eight subjects can be said to fully understand the playlist creation process consisting of adding songs to the playlist, removing songs from the list, and navigating the list.

The overall left–right spatial organization of the audio items was described correctly by seven subjects. The changing context playback on the left, however, was fully understood only by one user. Three people commented on the missing guidance in up–down direction. However, these three were the ones who felt the most comfortable using the UI.

In the three tasks performed after the familiarization, the subjects were very successful. One of the subjects missed the correct answers on the collection structure related questions by one; the others made no mistakes. The playlist creation task was successfully completed by all the users. However, some users first accidentally added extra songs in the playlist. This was because in the tested UI, there was no confirmation required to add items to the playlist.

The biggest problem that was found in the user test was the missing confirmation on the addition of an item to the playlist. Having the confirmation in place would also help to clarify the difference between selection and navigation. At least the confirmation is needed before adding an extensive number of songs to the playlist at once. Now it was all too easy to add a complete musical style to the list.

The problems anticipated before the test (see section 2.7) did indeed show up in the test. Three users commented on the missing feedback on the vertical menu, but still found the interface intuitive. The idea of direct breadth search and notification of the changing context was confusing to some users; only one subject fully understood this feature of the UI. The feedback on selecting items to the playlist was less of a problem. Only two users were slightly confused on the location of the playlist.

5. CONCLUSIONS

In this paper, we described a new UI designed for non-visual interaction with a music collection. The UI can be used for navigating a large list of songs organized hierarchically, and for generation of personal playlists. The interface utilizes stereo audio output with headphones and a five-way key for input.

We also described the implementation of the UI on a PDA, and presented the qualitative results of the first subjective test. The results of the test are encouraging: most of our subjects could use and understand the interface after a five-minute period. However, the test also revealed some problems in the interface to be solved in the future. After the corrections, the interface should be more thoroughly tested with and without a complementary visual user interface.

6. REFERENCES

[1] S. Pauws, D. Bouwhuis, E. Eggen, “Programming and Enjoying Music with Your Eyes Closed,” in *Proc. of CHI2000*, The Hague: ACM Press Addison-Wesley, 2000, pp. 369-376.

[2] Crispien, K., Petrie, H., “Providing access to GUIs for blind people using a multimedia system based on spatial audio presentation,” in *Proc. of the 95th AES Convention*, New-York (1993).

[3] Mynatt, E. and Edwards, W. K., “Mapping GUIs to Auditory Interfaces,” in *Proc. of ACM Symposium on User Interface Software and Technology (UIST)*, 1992.

[4] L. Ludwig, N. Pincever, M. Cohen, “Extending the Notion of a Window System to Audio," *IEEE Computer*, 1990, pp. 66-72.

[5] Savadis A., Stephanidis, C., Korte, A., Crispien, K., Fellbaum, K., “A Generic Direct-Manipulation 3D-Auditory Environment for Hierarchical Navigation in Non-visual Interaction”, in *Proc. of Assets’96*, New York: ACM, pp. 117-123.

[6] Sawhney, N. and Schmandt, C., “Nomadic Radio: Speech and Audio Interaction for Contextual Messaging in Nomadic Environments”, *ACM Transactions on Computer-Human Interaction*, Vol.7, No. 3, September 2000, pp 353-383.

[7] Walker, A., Brewster, S. A., McGookin, D. and Ng, A., “Diary in the sky: A spatial audio display for a mobile calendar,” In *Proc. of BCS IHM-HCI 2001*, Lille, France, Springer, pp. 531–540.

[8] Goose, S. and Möller C., “A 3D Audio Only Interface Web Browser: Using Spatialization to Convey Hypermedia Document Structure”, *ACM Multimedia* (1) 1999: 363-371.

[9] G. Lorho, J. Marila, J. Hiipakka, “Feasibility of Multiple Non-Speech Sounds Presentation Using Headphones,” in *Proc. of ICAD’01*, Espoo, Finland, July 29–August 1, 2001, pp. 32–37.

[10] O. Kirkeby, “A Balanced Stereo Widening Network for Headphones,” in *Proc. AES22nd Int. Conf. on Virtual, Synthetic and Entertainment Audio*, Espoo, Finland, June 15–17, 2002, pp. 117–120.

[11] G. Lorho, J. Hiipakka, and J. Marila, “Structured Menu Presentation Using Spatial Sound Separation,” in Proc. Mobile HCI 2002, Pisa, Italy, September, 18–20, 2002, pp. 419–424.

[12] M. Nilsson, “ID3 tag version 2.4.0,” November 1, 2000. Available from <http://www.id3.org/develop.html>.