Designing Low-Dimensional Interaction for Mobile Navigation in 3D Audio Spaces

Till Schäfers[1], Michael Rohs[1], Sascha Spors[1], Alexander Raake[1], Jens Ahrens[1]

[1] *Deutsche Telekom Laboratories, TU Berlin, Germany*

Correspondence should be addressed to Michael Rohs (`michael.rohs@telekom.de`)

ABSTRACT

In this paper we explore spatial audio as a new design space for applications like teleconferencing and audio stream management on mobile devices. Especially in conjunction with input techniques using motion-tracking, the interaction has to be thoroughly designed in order to allow low-dimensional input devices like gyroscopic sensors to be used for controlling the rather complex spatial setting of the virtual audio space. We propose a new interaction scheme that allows the mapping of low-dimensional input data to navigation of a listener within the spatial setting.

1. INTRODUCTION

Spatial audio has made its way into the consumer market. Surround sound systems for the living room are well established products. Also, sound card manufacturers have started enhancing their products by adding binaural effects which provide spatial sound impressions to listeners using standard stereo headphones. Standard APIs like OpenAL or DirectX Audio make it comparably easy to add spatial audio to desktop applications.

Mobile devices are less advanced with regard to spatial audio capabilities; to our knowledge, currently no API or standard software exists for building applications using 3D audio on mobile devices. However, OpenSL ES, a crossplatform standard for dealing with audio effects on mobile devices, authored by the Khronos Group [7], is on its way. Meanwhile, rendering 3D audio on a server, while using mobile devices for interaction and for visual representation, is a vital option for prototyping and evaluating mobile 3D audio applications.

Use-cases for enhancing mobile audio with 3D effects are manifold [4]: Mobile entertainment applications like games or movies will likely play an even greater role in the future than they do already today. Also, new application scenarios for managing multiple audio streams, e.g., teleconferencing or telepresence applications, as well as audio notification systems can greatly benefit from spatial distribution of multiple sound sources in a virtual audio environment. The potential constraint of having to carry a stereo headset for being able to use spatial audio is becoming less and less of a problem; the growing convergence makes stereo headsets a common accessory because users tend to use their mobile device for listening to music as well.

With spatial audio on mobile devices becoming a realistic scenario in the near future the question arises, how to incorporate this feature so users can benefit the most from it. In this paper, we investigate the new design space which is opened up by spatial audio and which will have to be thoroughly designed in order to maximize usability. Especially in mobile use-cases, with users possibly focusing on other tasks while interacting with the system, special care has to be taken when designing interaction. We focus on navigation within the 3D audio space and propose a new method for moving within the virtual soundscape, using low-dimensional input modalities like sensor-based motiontracking.

2. SPATIAL AUDIO – A NEW DESIGN SPACE ON MOBILE DEVICES

2.1. Design Space

Generally, spatial audio allows positioning of sound sources in a 3-dimensional scene, as well as specifying the listener's position within this scene. However, with the rendering techniques available at present, it is common practice to limit the area for source and listener positioning to two dimensions, thus using only positions on the horizontal plane around the listener's ears. This is done in order to avoid problems with perceiving exact locations of elevated sources and for being able to use 2-dimensional headrelated transfer functions (HRTFs) which are of substantially smaller size than their 3-dimensional counterparts. HRTFs filter the audio signal in a similar way as the head and ears do when receiving an audio signal from a specific position.

Within this soundscape, sound sources and listener can be regarded as objects in a 2D plane. Relevant parameters for spatial arrangement of objects in a plane are: Location, orientation, and their time-dependent counterparts, movement and rotation.

Location: The ability to arrange sound sources, as well as the listener's position, in a spatial setting is what makes the main difference between conventional audio and 3D audio. This makes positional information an important input to the audio rendering system and a powerful means of controlling the user's perception of the audio data, e.g., when regrouping participants in a teleconference.

Orientation: Binaural rendering techniques usually use the model of point sources which emit sound waves equally into all directions, so a sound source does not need to have an orientation parameter. The listener's orientation, however, is very important as it has a strong impact on how the audio is rendered and perceived: Although the so-called Cocktail Party Effect [1] enables us to concentrate on a sound source even when we are not facing it directly, turning towards a sound of interest is a very natural gesture.

Movement and Rotation: Changes to location and orientation over time add another dimension to the audio design space. Positional changes over time can be predefined as trajectories of sound sources, providing impressions of moving objects.

2.2. Generic Design Scenarios

When considering the use of spatial audio for interactive applications, it is worth taking a look at what types of generic use cases may emerge. In general, the main advantage of spatial distribution of sound sources is improved differentiation between sounds and, in particular, improved intelligibility in the case of speech from multiple simultaneous sources [3]. Also, identification of sounds can be enhanced when combining spatial audio with a visual interface. For example, in a teleconference this mapping between auditory and visual information could be used to identify previously unknown participants by direction of their voice mapped to additional visual information, e.g., their name, on the display. In general, this use case can be described as providing an optimized overview over the auditory scene; the spatial setting allows listening to all sound sources while enhancing differentiation.

As a second basic use case, the user might want to focus on one or more sound sources while still having awareness of the background sounds. This setting could be used in management of multiple audio streams; for example, the user might receive a phone call while listening to music. Instead of having to switch off the music in order to answer the call, the user might want to focus on the call while the music keeps playing in the background. This would resemble a living-room scenario, where one might simply turn down the volume before answering a phone call. Thus, the ability to focus on sound sources is another important ability that may have to be provided by the auditory design of an application.

Other use cases include changing positions of sound sources in order to re-arrange the scene to individual needs and preferences.

These use cases require the system to give some control over the spatial arrangement to the user. In order to focus on a source, the user might want to turn into its direction or navigate towards the source position. In order to keep an overview over the scene, the user might want to take a position in the middle of all sources.

When using a map-like representation for the auditory scene, this resembles the interaction used in desktop-based systems: The user has to be able to manipulate sources location-wise. For the representation of the listener, the orientation parameter also has to be adjustable. Thus, input modalities have to allow two degrees of freedom (x and y translation) for the sources and three degrees of freedom (x and y translation plus orientation) for the listener.

2.3. Dimensionality of Input Devices

When designing interaction for mobile devices, it has to be taken into account that available input devices can be restricted in terms of degrees of freedom they are capable of delivering. This is especially true when exploring new input modalities, using sensing techniques to capture gestures or device orientation. Some work has already been done to explore gesture-based interaction with spatial audio application on mobile devices: Brewster et. al. [5, 6] showed how input modalities like head tracking, nodding, or pointing can be used to control these types of applications. Billinghurst et. al. [2] evaluated several interaction techniques for spatial audio applications on mobile phones.

Promising and technically feasible solutions for one-handed interaction with spatial audio applications include:

- Head tracking by using a gyroscopic sensor attached to headphones.
- Tracking device orientation with compass heading or gyroscopic sensors.
- Using the built-in camera and optical flow algorithms to measure translational movements of the device.
- Sensing tilt by using accelerometers attached to or built into the device.
- Using the built-in joystick or touchpad, respectively.

Apart from the joystick interaction which provides true two-dimensional interaction capability, all types of sensing technologies mentioned above do not work well for more than one degree of freedom. Although some of them in principle allow more than one input dimension, they are hard to operate when more than one dimension is to be controlled at the same time. In part, this is due to a lack of sensor fidelity, but even when high fidelity sensing is available, navigation with full degree of freedom is difficult to handle. When multiple degrees of freedom are sensed simultaneously and mapped to different parameters, trying to control one parameter easily results in inadvertently changing another parameter as well; the task of freehandedly controlling multi-degree of freedom input generally requires very precise physical gestures.

Designing for mobile devices also means designing for mobile usage. The task of operating a device while walking along a street or standing in a crowded subway makes gesture-controlled interaction inherently difficult, which is another motivation to restrict sensor-controlled input to one degree of freedom.

3. DESIGNING FOR LOW-DIMENSIONAL INPUT

To tackle the aforementioned problems in interacting with rather complex scenes while using low-dimensional input devices, we explored positioning of audio sources within the spatial setting, as well as navigation techniques for the listener. The general strategy is to simplify interaction by using low-dimensional input and to arrange the sound sources and the listener's position relative to them such that desired states in the design space can be reached easily with low-dimensional input devices.

3.1. Overview and Focus

For providing an auditory overview to the user, sources should be positioned in a way that maximizes differentiation while keeping each source equally understandable. The most obvious spatial setup is the equidistant distribution of all sources on a circle, positioning the user in the center. However, due to common problems with front-back confusion in spatial audio, we restricted the source positions to a semi-circle.

Focus is provided by letting the user move freely within the semi-circle; sources close to the listener's position then appear louder due to spatial proximity. In our case, positioning is based on a polar coordinate system. The azimuthal parameter of the listener's coordinates is tied to the orientation of the listener. This reduces the degrees of freedom needed for navigation from three to two. Also, the two polar parameters allow a weighted input scheme as they are semantically disparate. In our case, the azimuthal parameter (along with the looking direction coupled to it) is used as the primary input parameter, while the distance from the circle center serves as secondary input parameter.

The main rationale behind this weighting scheme is the natural gesture of turning towards a sound of interest: For example, in a teleconferencing scenario with multiple participants one will likely change listening direction many times (if the tracking is done well), in order to turn towards the person that is currently talking. Setting the focus parameter will be done far less frequently. Decoupling the two input parameters opens up possibilities for new combined types of interaction; while the primary coordinate parameter can benefit from sensor-based input, the secondary parameter could be controlled by keypad interaction or by a different type of device sensing, e.g., tracking distance, height, or location. It could also be set to a constant value by the system, as for some spatial audio applications, changes to focus/overview during interaction might not make sense.

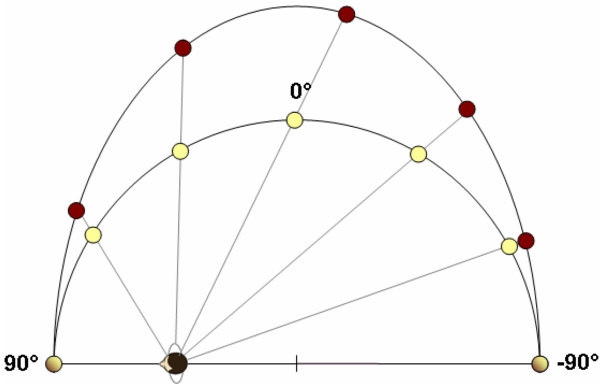

Fig. 1: Projection of source positions onto an elliptical path. The listener's offset from the circle center determines the length of the ellipse's major axis.

3.2. Dealing with Distraction

When distributing sources on a semi-circle, moving the listener closer to one source also means decreasing his or her distance to most other sources, which adds to auditory distraction. To tackle this problem, we designed the source positions to lie on an elliptical path (Figure 1). The minor axis of the ellipse equals the radius of the semi-circle and is oriented along the listener's looking direction. The length of the major axis depends on the distance between the circle center and the listener's current position. Initially, with the user being positioned at the center of the semi-circle, the major and minor axes equal in length, so the semi-circle is conserved for the overview setting. When moving towards a source, the user's offset from the center of the semi-circle is added to the major axis.

The exact position of a source on the elliptical path is determined by projecting the position of the source on the semi-circle onto the ellipse, using the listener's current position as a reference point. Thus, the angular position of a source as heard by the listener stays the same, while the increased distance leads to sound attenuation.

This setup provides enhanced differentiation between overview and focus settings: The initial (overview) position of the listener being located in the center of the semi-circle does not differ in the elliptical setup, as an ellipse with equal axes yields a circle. However, when the listener moves into the semi-circle to focus on a source, most other sources, especially those being located around 90° to the left and right substantially increase their distance to the listener. From a psychoacoustical perspective, this is of particular importance, as sound sources positioned around $+/-90°$ from the looking direction add much to distraction due to angular proximity to the ears [3].

3.3. Analysis of Sound Levels

When using the projection algorithm described above, it is worth taking a look at how the sound level of spatially arranged sources is altered when the listener changes from the overview setting to a focused position. Changes in attenuation for a monaural setup were computed for seven sources using a linear fall off model. Initially, sources are arranged equidistantly on a semi-circle, their azimuthal values ranging from $-90°$ to $90°$ in steps of $30°$ (overview position). In this position all sources have the same reference level of 0dB. The computed values depicted in Figure 2 show a scenario where the listener is rotated by $+90°$, thus facing the leftmost source. For auditory perception this can be considered the worst-case scenario, as all sources are now located to the right of the listener, which makes them prone to front-back confusion. However, the full range of elliptical projections can be seen in this setting. For every source, two sound levels are depicted, with the listener having moved 1/3 (33%) and 2/3 (67%) of the circle's radius towards the target source.

With respect to the original position of the listener at the center of the semi-circle, a significant amplification of the target source located in front of the listener can be seen for both distance values. The source next to it (at $60°$) also gets amplified due to spatial proximity, although significantly less than the target source, especially at 67% offset. All other sources are attenuated below their initial level of 0dB, allowing the listener to focus on the target source. Although this monaural computation of attenuation values does not take binaural effects into account, it provides valuable clues about the level of distraction from background sources in different settings.

4. CONCLUSION

In this paper we explored a new way to interact with spatial audio applications. Especially when using low-dimensional input devices, splitting spatial navigation into primary and secondary parameter can improve interaction with the system: This method frees the designer from having to use an input device capable of delivering two degrees of freedom. The general strategy we followed was to simplify interaction so it distracts the user

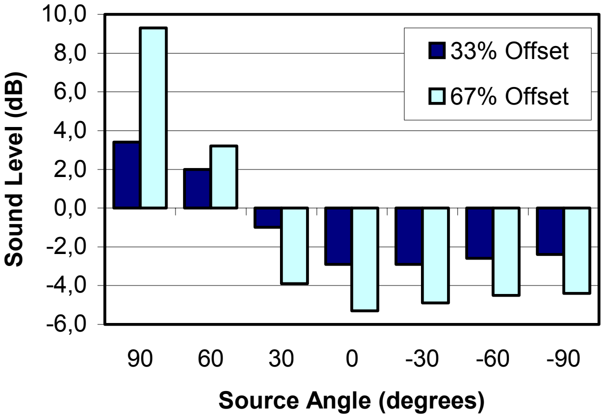

Fig. 2: Changes in sound level for seven audio sources as the listener is moved by 1/3 (33%) and 2/3 (67%) of the circle's radius towards the target source located at 90°.

as little as possible from other simultaneous tasks, such as walking. At the same time we developed a scheme for the arrangement of sound sources and the listener's position relative to them that makes it easy to reach "desirable" states in the design space with low-dimensional input. The setup was chosen in such a way that comprehensibility and source separation are achieved in order to provide an overview over the auditory scene. At the same time the interaction method allows the user or the application designer, respectively, to choose a focus level based on user preferences or application needs.

Future work will deal with evaluation of this interaction scheme: First, finding the optimal function for coupling the elliptical path to the user's position can be done best with auditory tests. Second, the interaction paradigm of using two weighted parameters will have to be evaluated, in order to find out about how intuitive this setting is when put into practice, especially when using aforementioned gesturebased input methods.

5. REFERENCES

[1] B. Arons. A review of the cocktail party effect. *Journal of the American Voice I/O Society*, 12:35–50, July 1992.

[2] M. Billinghurst, S. Deo, N. Adams, and J. Lehikoinen. Motion-tracking in spatial mobile audio-conferencing. In *Workshop on Spatial Audio for Mobile Devices (SAMD 2007) at Mobile HCI 2007*, September 2007.

[3] M.L. Hawley, R.Y. Litovsky, and J.F.Culling. The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *Journal of the Acoustical Society of America*, 115:833–843, 2004.

[4] J. Huopaniemi. Future of personal audio - smart applications and immersive communication. In *AES 30th Intl. Conf.*, September 2007.

[5] J. Lumsden and S. Brewster. A paradigm shift: alternative interaction techniques for use with mobile & wearable devices. In *CASCON '03: Proceedings of the 2003 conference of the Centre for Advanced Studies on Collaborative research*, pages 197–210. IBM Press, 2003.

[6] G. Marentakis and S. Brewster. A study on gestural interaction with a 3d audio display. In *Proceedings of MobileHCI 2004*, pages 180–191. IBM Press, 2003.

[7] The Khronos Group. http://www.khronos.org.