Towards a Universal Binaural Audio User Interface

Androwis Abumoussa
University of Rochester
Computer Science, ROC HCI
Rochester, NY 14623 USA
androwis@cs.rochester.edu

ABSTRACT

Traditional 3D audio systems are often used to enhance the experience of movies and games. Their popularity has risen as more homes and theaters are outfitted with multiple speakers to provide the immersive experience of the content.

3D sound can actually be provided by stereo audio sources but implementations are often not realized or are limited because of the sensitivity of human audible perception requiring that the systems are properly calibrated to provide the proper sound experience.

3D sound is often used to provide context and supplement content in both video and games. The use of the technology has been explored in the context of user interface design but has not been implemented in any accessible user interfaces. Current solutions remain linear in scope, focusing on single focus interfaces. Since current audible interfaces are limited to only one dimension, users lose the ability to place sound in a 3D environment.

Irys uses binaural audio to spatially place sound in a 3D space relative to the user. The system offers the potential of increased productivity, smoother transitions, and a more fluid user experience. Context switches can provide users with known audible transitions to inform said user when focus is shifting, preparing the user for a different application.

ACM Classification Keywords

H5.2 [Information interfaces and presentation]: User Interfaces. - Audible user interfaces.

General terms

Design, Human Factors, Performance, Design, Experimentation

Author Keywords

Real-time, Binaural, Head-Related Transfer Function, Mobile

INTRODUCTION

Audible interfaces provide both sighted and visually impaired users with access to interfaces and content. These interfaces are often used when users are driving, using personal handheld computing devices, or unable to provide the interface full visual attention. These interfaces are able to interact with screen-based structures as well as sensual representations of our environment [1].

Most audible interfaces provide a text-to-speech layer that allows systems to read the content of the interface to the reader. There exist solutions that modify the behavior of an interface (such as Apple's VoiceOver). The audio that most of these interfaces rely on is monophonic, meaning that the audio is perceived as coming solely from one speaker. No effort is made to spatially place the source of audio to provide cues to the user.

The lack of adoption of audio as an interface is often attributed to a few factors. Prior research concludes that humans base their acceptance of sound synthesized by machines on three features: Gestures, Nuance, and Inflection [2]. Most modern speech synthesizers often perform poorly on these measures. Audible interfaces that are not based purely on speech, but focus on other kinds of sounds have more promising results [2].

When using sounds as a communication medium to interact with humans, these factors need to be considered due to humans sensitivity to sound. As Thackara mentioned [2], humans mostly have no choice but to follow an auditory pattering as long as it does not consist out of too much sound in the sense of noise pollution. It is important to keep these points in mind when creating an interface based primarily on sound.

Related Work

Three-dimensional audio systems render sound images around a listener by using either headphones or loud speakers [2]. In the case of 3D audio systems based on headphones, the 3D audio cues to localize a virtual source can be perfectly reproduced at the listener's eardrums because the headphones isolate the listener from external sounds and room reverberations. There exist systems that are capable of producing binaural audio to a user using stereo speakers in an open environment with the aid of head track

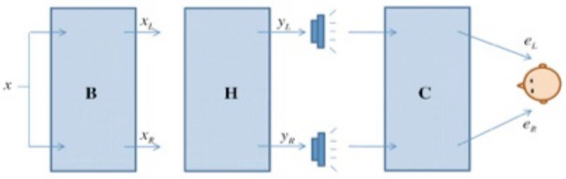

Figure 1: Schematic of binaural audio system

ing webcams [3]. Either of these systems are able to perfectly calibrate sound placement and create an experience that provide the user with the perception that sound is travelling around them.

Previous work have explored binaural audio as positional cues in navigation applications while the gaming industry rely on these interfaces to enhance the user experience.

This paper describes a system that supports a new conceptual model of interface mapping in 3D space. Using binaural audio as the mechanism, novel features are discussed that provide information to the user in terms of spatial attenuation, audio structural survey of content on the web, accurate positional audio feedback, and an audible progress indicator. These new features can improve both the user's comprehension of content presented to them while provided with cues to assist recall of information.

BINAURAL AUDIO SYSTEMS

The block diagram of a typical binaural audio playback system is depicted in Figure 1 [4]. The binaural audio engine itself consists primarily of the binaural synthesizer **B**. The goal of the binaural synthesizer is to produce sounds that should be heard by the listener's eardrum. More succinctly, we want e_L and e_R to be equal to x_L and x_R.

The crosstalk canceller **H** is the component that knows about the sounds pathway from the source of sound to the listener's eardrums. The sole role of the crosstalk canceller is to convolve the audio to equalize any effects of component **C**, the physical transmission pathway from the source of sound to the listener's ear.

Components **H** and **C** were the topics of research discussed by Song et al [4] when considering loud speakers as a source of input. If the user has headphones, then these components can be disregarded.

Binaural Synthesis

The binaural synthesizer B is responsible for creating one or more virtual sound images at different locations around the listener using 3D audio cues. Among many binaural cues for the human auditory system to localize sounds in 3D such as the interaural time difference (ITD) and the interaural intensity difference (IID), I explore the use of HRTF, which is the Fourier transform of the head-related impulse response (HRIR). Since HRTF captures most of the physical cues that human relies on for source localization. Once the HRTFs of the ears are known, it is possible to synthesize accurate binaural signals from a monaural source [5].

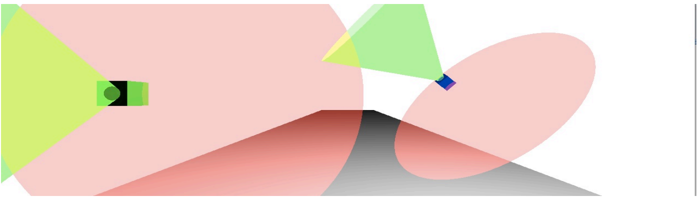

Figure 2: Sample rendition of sound sources in 3D space

UBIQUITOUS BINAURAL AUDIO INTERFACE

Irys is an audible interface designed for use primarily on the web on devices that can produce sound through headphones. The interface places sound images around the user, providing them with a 3D sound environment to interact with their technology.

The conventional binaural audio system works well if the listener stays at the position (usually along the perpendicular bisector of the two points of sound) corresponding to the presumed binaural synthesizer **B**. However, once the listener moves away from the sweet spot, system performance degrades rapidly.

If the system intends to keep the virtual sound source at the same location, when the head moves independent of the sound sources, the binaural synthesizer shall update its HRTF matrix to reflect the movement. In addition, the acoustic transfer matrix C needs to be updated too, which leads to a varying crosstalk canceller matrix H. The updates of B and H were referred as "dynamic binaural synthesis" and "dynamic crosstalk canceller", respectively [6].

For this project, we will not concern ourselves with the case that the user is moving independent of the audio source and assume that the user is either in a stationary environment or has headphones to remove the need for external monitoring and real time updating of the audio convolutions.

In this paper, we propose to build a personal 3D audio system to draw sound images around the user. The virtual environment is depicted in Figure 2. The goal of this project is to create a development environment that allows sound images to be placed arbitrarily around the user to depict content audibly.

Four physical interfaces were explored: native desktop applications, Android handhelds, iOS handhelds, and the web. Each medium provides different drawbacks and benefits for the interface being built.

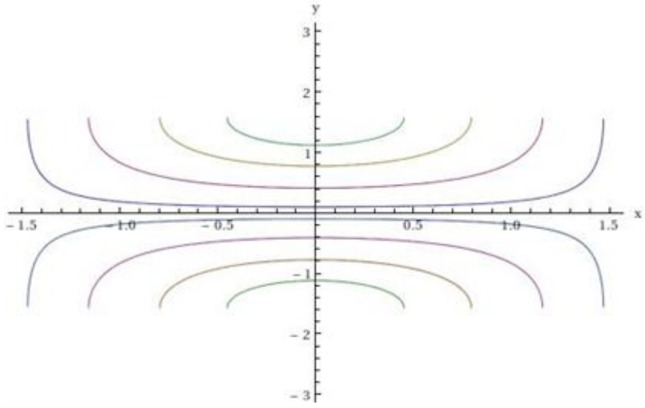

Figure 3: Sample sound pathways tested with OpenAL

Methods - OpenAL

OpenAL is a cross-platform open sourced library that provides efficient rendering of multichannel three-dimensional positional audio. It has implementations on most native application frameworks, and at the onset of the project, seemed to provide a silver bullet for much of the interface across multiple devices. During the implementation cycle of this project, we found that OpenAL provided exciting abstractions, but distance was provided by volume amplitude attenuation and not properly calculated with a delay.

Creating a native desktop, iOS and android application using OpenAL was relatively quick, but upon evaluation by human subjects, it became apparent that the framework was too limiting. By using volume to place the sound, the user was left with jarring edge conditions as the sound image crossed planes of reference. Figure 3 represents pathways tested on users, where each line represents a sound traversal pattern relative to the user centered at the origin. Because OpenAL uses volume based attenuation and not delays in sound queuing, items were perceived to be travelling along a single flattened left-to-right (x-axis) plane.

Despite the flattening perception of the library, it was a great tool to test some of the concepts on both mobile and desktop environments to initially understand if such a framework was feasible and useful. With OpenAL, a framework is provided that allows for audio streams to be created and played in real-time (this is contrary to what can currently be done on the web, as HTML requires that the audio to be played already exist).

Methods – HTML5

For the web, we were able to utilize a new audio tag introduced by the web standard community. The new HTML5 audio tag allowed us to modify the JavaScript on web pages to generate the necessary transforms and delays to place the sound in a 3D environment. Using the library Three.js we were also able to create the necessary callback scripts to perform the transformations as well.

The framework we present here, allows an individual to occupy a space and interact with the surroundings. The browser is able to perform the necessary transformations and displace the audio around the user.

FUTURE WORK

3D audio interfaces present a number of exciting capabilities in human computer interaction. With an initially completed framework and literature review, the immediate short-term goals for this project are to perform user studies and performance measures on the efficacy of this type of interface as it relates to different tasks.

The target individuals for this interface are blind and low vision users. I plan on performing more in-depth studies of how this interface can be used to best enable blind people to interact with a given interface through multi-tasking techniques afforded by independent sound objects.

Having multiple sources of audio may be distracting for a user, so evaluation on the number of voices a user can focus on, what types of information are best presented to the user, and time locality are all other metrics of interest for this interface.

Search and navigation within this type of interface becomes an interesting research topic. How can a user query information audibly. Systems, such as Apple's Siri and Google Voice attempt to provide an interface for general query and answer interactions, but how can a system be built to allow for in-depth querying of content in a spatial manner? Should context be provided to search or should the system only search the locality around the user?

Most importantly, the next major focus will be on quantifying the benefits of this type of interface. Metrics on goal completion on tasks in a 3D space as compared to regular interfaces as well as throughput as measured by multi-task capacity would assist in understanding the efficacy of this system. Finally, we're very excited to explore the ability of 3D audio in helping users remember information by providing a tangible dimension to their information processing.

CONCLUSION

We have presented Irys, an interface that uses 3D audio to place sound around a user on any device. Irys leverages techniques in binaural audio to provide the user with an immersive environment to interact with their technology. Irys is useful, both as a tool for enabling blind users, but as an approach to test spatial layout of information for humans.

REFERENCES

1. Michelis, Daniel, et al. "The disappearing screen: scenarios for audible interfaces." *Personal and Ubiquitous Computing* 12.1 (2008): 27-33.

2. Thackara J (2005) In the bubble—designing in a complex world. MIT Press, Cambridge.

3. C.Kyriakakis, “Fundamental and technological limitations of immersive audio systems,” *Proc. IEEE*, vol. 86, pp.941?951, 1998.

4. Song, Myung-Suk, et al. "Personal 3D audio system with loudspeakers." *Multimedia and Expo (ICME), 2010 IEEE International Conference on IEEE, 2010*.

5. W. Gardner, “3-D audio using loudspeakers,” Ph.D. thesis, Massachusetts Institute of Technology, 1997.

8.HTML5 Rocks. Heikennan, Ilmari. Feb 16, 2012. Google. Oct 15, 2012.

6. T.Lentz, G.Behler, “Dynamic Crosstalk Cancellation for Binaural Synthesis in Virtual Reality Environments,” *J. Audio Eng. Soc.*, Vol. 54, Issue 4, pp. 283-294, 2006.

7.HTML5 Rocks. Smus, Boris. Oct 14, 2011. Google. Oct 15, 2012. http://www.html5rocks.com/en/tutorials/webaudio/intro/.

http://www.html5rocks.com/en/tutorials/webaudio/positional_audio/.