The University of Rochester

Area Paper
for the Degree Master of Science

Binaural Audio as an Accessible User Interface

Author:
Androwis Abumoussa

Supervisor:
Dr. Jeffrey Bigham

April 24, 2013

Contents

Acknowledgements

Though only my name appears on the cover of this thesis, there are a number of people who played an integral part in its production. I am forever grateful for those who have made this thesis possible and because of whom, my graduate experience thus far has been one that I will cherish forever.

My gratitude belongs to my advisor, Dr. Jeffrey Bigham. I have been amazingly fortunate to have an advisor who gave me the freedom to explore on my own, and at the same time the guidance to recover when my steps faltered.

My co-advisor, Dr. Henry Kautz, has been always there to listen and give advice. I am deeply grateful to him for the long discussions that helped me sort out many technical details of my work.

I am also indebted to Walter S. Lasecki whose interactions and discourse have proven to be the some of the most valuable ones. As a peer, he is second to none in compassion and guidance.

Most importantly, none of this would have been happened without the love, patience and support of my family. My mother, to whom this dissertation is dedicated to, has been a constant source of love, concern, support and strength.

Abstract

Traditional 3D audio systems are often used to enhance the experience of movies and games. Their popularity has risen as more entertainment centers 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 re-quiring 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 inter- faces. 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.

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 [9].

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 Apples 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. 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 experienced more promising results [15].

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 [15], 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.

Our society is driven by information. Armed with devices that are contstantly connected, the current generation of technology has the potential to communicate massive amounts of information, everything from weather forecasts and traffic conditions, to neighboring attractions, restaurant schedules, store specials, even to the location and discoveries of our friends.

Fields of research explore how to communicate constantly changing information to interested parties at the appropriate time. With the influx of mobile

devices that provide an always on channel, research has explored the effect these distruptions have on a multitasking computing environment. The goal of much of this research has beento study how relevant and correct information can be efficiently delivered to a user in a manner that does not distract from their current tasks [8].

While exploring the vast world of Audible Interfaces, this paper works towards two major goals. First a review of the methodologies, processes, and terms necessary to study audible interfaces is presented. In conjunction with this, an argument for the benefit of a 3D Audio Interface will be presented as an accessible solution for users with visual disabilities. A general background of Binaural Audio and Audible Interfaces is provided in the following section. The related work connects ideas and findings both in the fields of Human Computer Interaction and Signal Processing to form the basis of Our Approach in section 6. Finally, this work concludes with an overview and survey of future work necessary for the creation of an Binaural Audio Interface.

3.1 Motivation

As the push for the graphical representation of information continues within computer science, the most prevalent solution to providing the visually impaired with a usable computing system continues to rely heavily on screen reading solutions. Though this solution has provided tremendous benefits, many concepts are lost in the cross sensory translation and mapping. For example, users of screen readers lose the concurrency afforded by having multiple graphical windows open simultaneously, and have a hard time discerning unprompted changes of focus. As interface designs continue to leverage graphics, the attempt to translate graphic interfaces into a serial auditory stream becomes polluted with inefficiencies and downfalls.

Imagine driving to work on a typical morning. Eric Horvitz dream of an intelligent interface has been realized, so your navigation system is checking current road conditions relative to the location of the GPS inside of the car. Your smartphone has resumed polling your work email address, your calendar has been updated by a colleague, and your family is messaging you reminding you of a prior engagement.

system that knew exactly how much or how little to say. The system is able to provide just the right amount of context for each task you're performing.

Background

4.1 Binaural Audio

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 listeners 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 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.

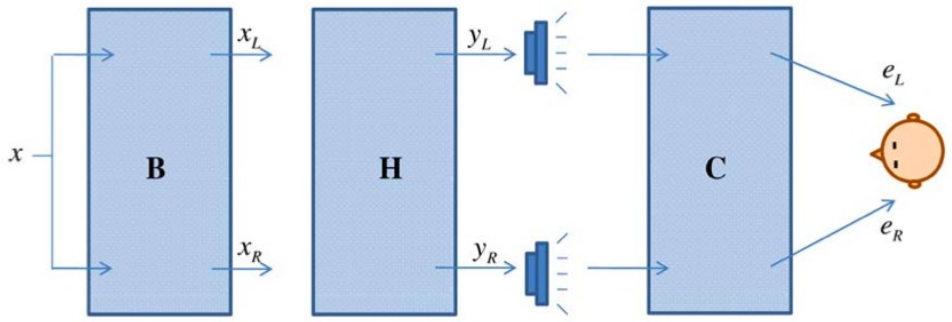

Figure 4.1: Schematic of a binaural audio system

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 at-tenuation, audio structural survey of content on the web, accurate positional audio feedback, and an audible progress indicator. These new features can improve both the users comprehension of content presented to them while provid-ed with cues to assist recall of information.

Human auditory localization has been studied extensively [1, 16]. Humans are especially adept at localizing sounds in three dimensions. Consider a sound source to the left of a listener. Sounds from the source arrive at the left ear first, and a short time after reach the right ear. The amplitude of the left ear sound will be attenuated due to head shadowing.

The predominant audiotory cues for determining whether a sound is coming from the left or right directions are the interaural intensity differences and the interaural time differences. Humans are also adept at identifying sound position that are in front of or behind them, along with estimating the sound source's elevation. This is possible because the incident sound waves interact with the torso, head, external ear (pinna) prior to arriving at the inner ear.

The directional dependent filtering to each of a subject's ears can be expressed as a frequency response, called a head related transfer function (HRTF), and thus a pair of HRTFs describe how sound from one location reaches the two ears. HRTFs are usually measured using human subjects or dummy-head microphones which consist of response pairs for the left an right ears corresponding to a large number of source positions surrounding the head.

There are two regions of interest when considering source locations of sound. When the sound is close to the head, the spherical curvature of the incident sound waves cause the HRTFs to change qualitatively as a function of distance, but at moderate distances, the incident waves can be considered planar. At extreme distances, humans are only capable to process auditory cues that only depend on the sound sources volume.

Given that humans are capable of processing sound signals to place the audio source in space, a binaural spatializer can then be implemented to simulate the auditory experience of one or more sound sources arbitrarily located around a listener. The general idea is reproduce the acoustical signals at the two ears that would occur normally in a natural listening situation.

This process is accomplished by convolving each source signal with a pair of HRTFs corresponding to the sound sources intended location. The resulting

Figure 4.2: Binaural microphone with exact dimension and density as human head

signal is presented to the user through headphones.

4.2 Current Solutions

4.2.1 Transaural Audio

Transaural audio is a method used to deliver signals to the ears of a listener using stereo loudspeakers. The idea is to filter the binaural signal such that the subject can process the subsequent stereo representation as a binaural signal. This technique was first put into practice by Shroeder and Atal [12,13]. Song et al. demonstrated how computer vision techniques could be applied to the field of Transaural Audio. By tracking a subject's head, the computer can recalculate the HRTFs necessary for the sound, as well as reposition the sound sources relative to the subject to provide an adaptive 3D audio system

[14].

4.3 Research Domain

4.3.1 Audio Used for Analytics

4.3.2 Audio Used for Interfaces

4.4 Our Approach

4.5 Notifications

The essence of any event or update to user data is condensened into a notification. Current research has explored how notification systems attempt to deliver current, important information to computer screens. Many works have also explored the costs, benefits, and optimal displays of these notifications from psychological prespectives that overlap with our ability to handle interruptions and distractions [2, 8].

Related Work

A binaural interface would consist of two main elements: a pseudo intelligent agent able to classify the importance of specific notifications, and the binaural engine. The following section explores research that have explored both components as well as demontstrating the interplay between the different elements of this interface.

When designing an interface, there should be three main considerations. How can you make it usable ? Who is the user ? What is the task?

5.1 Psychology of Interface Design

5.1.1 Mental Models and Cognitive Loads

The ultimate goal User Interface Design is to mediate effective operation and control of a machine from the operator's perspective. The interplay of ergonomics and psychology are known facets of the field. As such, the design of novel interfaces are encouraged to either increase efficiency or user satisfaction (preferably both).

To more concretely enumerate the goals, the literature provides a number of guiding examples. Exploring a number of different goal oriented tasks, general patterns of interactions emerge that suggest the style that Irys should present.

A limitation of most interfaces is simply the size and volatility of a user's short term memory. To make real time captioning easier for untrained works Lasecki et al. demonstrated how time could be warped for any given users task. Warping the playback speed of previously recorded audio enabled users to perform the captioning task by reducing the high cognitive load needed for captioning [5].

Simplifying information is not only beneficial for goal oriented and task driven processes, but they are also emperically preffered by users of a number of systems.

Exploring a users collective preference for simpler representations of information. For example, travelling users preferred summarized directions that incorporated familiar paths to more detailed and exact routing instructions. The users in the study actually valued travelling on familiar routes enough to sacrifice overall trip distance [11].

When an interface is controlled by a semi-autonomous intelligent agent, it becomes increasingly important to be aware of the interactions that users will have with the system as decisions are being made. In this system, the decisions that would be made relate to the verbosity for a given notification, the position to place the notification relative to the user, and what actions are available to the user when acting on a notification. Prior work has explored how an intelligent agent can explain itself to the user, the amount varying degrees by which an intelligent agent can reconfigure itself, and the effects of a user's mental model soundness [4].

5.2 Immersive Interfaces

Kayur Patel defined a fully immersive virtual reality setting to be an extension of virtual reality that also captures a user's full body motion while immersing the subject into a virtual environment. The exploration of such an environment for educational purposes allowed researchers at UC Berkeley to conclude that the increased interactions led to more effective education [10].

5.3 Interface Elements

5.3.1 Visual Notifications

Given the stochastic nature of notifications, probabilistic reliability can be used to create intelligent messaging sytems that gain a user's trust, rather than lose it [6]. Emperical evidence show that users will ignore all notifications that are not highly valid when performing demanding visual tasks [7].

The question of a users trust in a system is therefore important when designing notification interfaces, with the goal being mitigating all negative first

impressions. Empirical evidence shows that users carry a historical bias when dealing with actionable notification [6].

The study of information design and options that are suitable for three often conflicting design objectives of notifications - interruption to primary tasks, reactions to specific notificaitons, and comprehension of information over time - are necessary.

5.4 Tactile Notifications

5.5 Auditory Notifications

SOCIAL ISSUES: In their paper Hanson et al. discussed the interplay of social situations and auditory cues [3] Current auditory notifications cues can be attention demanding, distinct, and can be percieved as intrusive in social situations. "The beeping and ringing is by nature an intrusive sound not unlike the sound of an alarm clock" referring auditory cues often heard arising from cell phones [3].

5.6 Current Audio Based Accessibility Solutions

5.7 Interface Design

5.8 Interfaces

5.9 Engines

5.10 Analytics

5.10.1 Design Principles

5.10.2 Design Patterns

5.10.3 Design Methodologies

5.11 Artificial Intelligence

Our Approach

UBIQUITOUS BINAURAL AUDIO INTERFACE

Irys is an audible interface designed for use primarily on the web on devices that can produce sound through head-phones. 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 perpendicu-lar 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 perfor-mance 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 sys-tem to draw sound images around the user. The virtual en-vironment 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 ap-plications, Android handhelds, iOS handhelds, and the web. Each medium provides different drawbacks and benefits for the interface being built.

OpenAL is a cross-platform open sourced library that pro-vides 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 ampli-tude attenuation and not properly calculated with a delay.

Creating a native desktop, iOS and android application us-ing 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).

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 pag-es 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 capabili-ties in human computer interaction. With an initially com-pleted 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 fo-cus 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 infor-mation audibly. Systems, such as Apples 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 quantify-ing 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, were very excited to explore the ability of 3D audio in helping users remember information by providing a tangible dimension to their information pro-cessing.

7.1 Human Computer Interaction

The study of Human Computer Interaction has very well guidelines and methods for evaluating graphic interfaces.

7.2 Analytics Domain

7.3 Human Computer Interaction

Conclusions

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.

Bibliography

[1] Jens Blauert. *Spatial hearing: the psychophysics of human sound localization*. The MIT press, 1997.

[2] Edward Cutrell, Mary Czerwinski, and Eric Horvitz. Notification, disruption, and memory: Effects of messaging interruptions on memory and performance. 2001.

[3] Rebecca Hansson, Peter Ljungstrand, and Johan Redström. Subtle and public notification cues for mobile devices. In *Ubicomp 2001: Ubiquitous Computing*, pages 240–246. Springer, 2001.

[4] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. Tell me more?: the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, pages 1–10. ACM, 2012.

[5] Walter S Lasecki, Christopher D Miller, and Jeffrey P Bigham. Warping time for more effective real-time crowdsourcing. 2013.

[6] Scott LeeTiernan, Edward Cutrell, Mary Czerwinski, and Hunter Hoffman. Effective notification systems depend on user trust. In *Proceedings of Human-Computer Interaction–Interact*, 2001.

[7] Masha Maltz and Joachim Meyer. Cue utilization in a visually demanding task. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 44, pages 283–283. SAGE Publications, 2000.

[8] D.Scott McCrickard, Mary Czerwinski, and Lyn Bartram. Introduction: design and evaluation of notification user interfaces. *International Journal of Human-Computer Studies*, 58(5):509 – 514, 2003. ¡ce:title¿Notification User Interfaces¡/ce:title¿.

[9] Daniel Michelis, Florian Resatsch, Thomas Nicolai, and Thomas Schildhauer. The disappearing screen: scenarios for audible interfaces. *Personal and Ubiquitous Computing*, 12(1):27–33, 2008.

[10] Kayur Patel, Jeremy N Bailenson, Sang Hack-Jung, Rosen Diankov, and Ruzena Bajcsy. The effects of fully immersive virtual reality on the learning of physical tasks. In *Proceedings of the 9th Annual International Workshop on Presence, Ohio, USA*, pages 87–94, 2006.

[11] Kayur Patel, Mike Y Chen, Ian Smith, and James A Landay. Personalizing routes. In *Proceedings of the 19th annual ACM symposium on User interface software and technology*, pages 187–190. ACM, 2006.

[12] Manfred R Schroeder. Digital simulation of sound transmission in reverberant spaces. *The Journal of the Acoustical Society of America*, 47:424, 1970.

[13] MR Schroeder and BS Atal. Computer simulation of sound transmission in rooms. *Proceedings of the IEEE*, 51(3):536–537, 1963.

[14] Myung-Suk Song, Cha Zhang, Dinei Florencio, and Hong-Goo Kang. Personal 3d audio system with loudspeakers. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 1600–1605. IEEE, 2010.

[15] John Thackara. *In the bubble: designing in a complex world*. The MIT Press, 2005.

[16] William A Yost and George Gourevitch. *Directional hearing*. springer-Verlag New York, 1987.