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

May 6, 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, while also providing guidance to recover when my steps faltered.

My co-advisor, Dr. Henry Kautz, has always been 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 and for the incredible guidance he has provided throughout my academic career.

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, compassion, 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 a more immersive experience for 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 be properly calibrated to provide the proper sound experience. Minor deviations from calibrated settings result in a degraded 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 and lose contextual information that can easily be provided by 3D audio, such as transitional cues using common Doppler effects.

This paper introduces Irys, a binaural audio interface used 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.

Introduction

Audible interfaces provide an alternative style of access to computers when users visual focus is unavailable. Visual focus can be unavailable for a number of reasons, such as when users are driving, using handheld devices while engaged in other physical activities and are unable to provide the interface full visual attention, to other causes such as physical disabilities. Audible interfaces benefit also have the combined benefit of being able to interact with screen based structures and the users sensual representations of his or her environment [24].

Most audible interfaces only provide a text-to-speech layer that allows systems to read the content of a visual interface to a user. Some problems with current solutions that modify the intended behavior of an interface (such as Apples VoiceOver, JAWS, etc) is that these interfaces only present monophonic audio. To an end user, monophonic audio is perceived as coming from one stationary speaker and source. Techniques exist to provide spatially placed (3D) audio to listeners. The use of binaural audio has been explored by researchers for the use of web accessibility, virtual reality, and gaming settings to allow interfaces to provide cues to the user about their surroundings in an environment.

The lack of adoption of 3D audio as a primary interface is often attributed to a few factors in the literature. Prior research in psychology has shown that humans base much of their communication on gestures, nuance, and inflection [35]. As a result of modern speech synthesizers poor performance on these measures existing systems are often perceived as ineffective or poor communicators. Audible interfaces that are not based purely on speech, but complement other sensation modals, focusing on other kinds of non-communication based sounds have experienced more promising results (as is often demonstrated with games and movies) [35].

When using sounds as a communication medium to interact with humans, certain factors need to be considered due to humans sensitivity to sound. An interface designed around audio must understand the psychological basis for tone, nuance, and inflections when portraying information to a subject. As Thackara mentioned [35], humans have no choice but to follow an auditory patterning as long as it does not consist of too much distortion in the sense

of noise pollution. Humans inability to ignore most sound plays an important role when creating an interface based primarily on sound to avoid user frustration.

Much of our society is driven by information. Armed with devices that are constantly 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 have explored how to best 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 disruptions have on a multitasking computing environment. The goal of much of this research has been to study how relevant and correct information can be efficiently delivered to a user in a manner that does not distract from their current tasks [23].

3.1 Motivation

As the push for the graphical representation of information continues within computer science, the most prevalent solution to providing visually impaired users with a usable computing system continue to rely heavily on screen reading solutions. Though these solutions provide 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.

Instead, imagine a world where you had an army of assistants. Furthermore, these assistants were smart and effective communicators that could succinctly and politely portray just the right amount of information to you. You could orchestrate these assistants any way you'd like, but with the ultimate goal being to carry conversations with your computational tasks, not have them all yell at you through disruptive notifications.

Take a common scenario like driving to work on a typical morning. Eric Horvitz described how intelligent interfaces have been realized, so that the computer powering the navigation system is also checking current road conditions relative to your location. 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. In his work, he quantified the risks and benefits of using different notification paradigms to express this information.

If a system existed that conformed to common patterns of communication, audio would be a major component. By exploring human psychology and surveying user preferences, can a system be created that can discern exactly how much or how little to say when presenting a user with updated information? Can a system be created that successfully evaluates an individual's preferences and needs, adjusting settings for both physical constraints such as fine tuning the user specific head related transfer functions, to the verbosity of given notifications? By understanding usage patterns for common tasks, such as navigation, communication the system aims to provide just the right amount of context for each task a user is performing, creating a framework that minimizes cognitive load and maximizes productivity.

3.2 Project

This paper describes a system that supports a new conceptual model that maps interface elements into a 3D audio 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 users comprehension of content presented to them while provided with cues to assist recall of information.

Human auditory localization has been studied extensively [4,38]. 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 auditory 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.

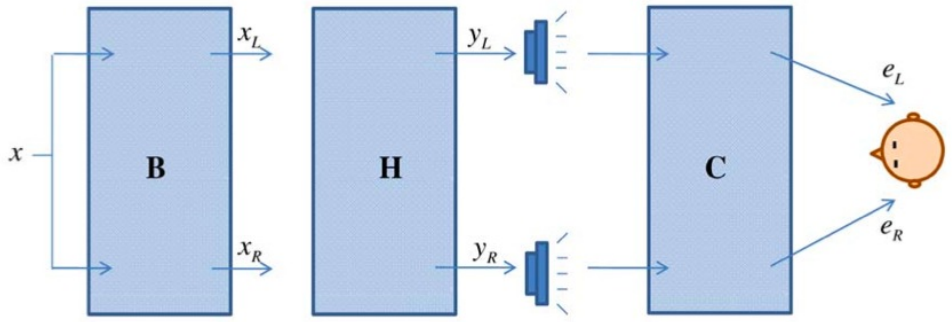

Figure 3.1: Schematic of a binaural audio system

The remainder of this thesis explores both the questions that remain to be explored as well as the steps to achieve these goals.

While exploring the research in the utilization 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 a binaural audio interface as well as goals research in this area would pursue.

Background

4.1 Sound Localization

Sound localization refers to a listener's ability to identify the origin of a detected sound in both direction and distance. Mammalian sound localization mechanisms have been extensively studied. The following section provides the reader with background information necessary for the remainder of the work.

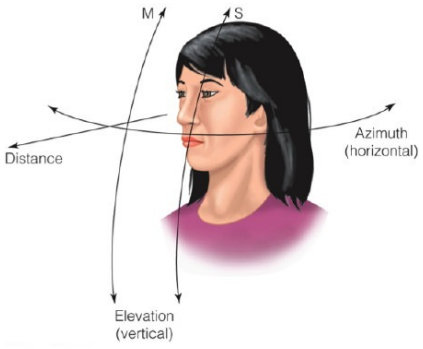

Figure 4.1: Measurements used for spatial sound processing analysis.

The brain is able to utilize subtle differences in intensities, spectral and timing cues. Often the terminology used to place a sound relative to a subject are *azimuth* (the horizontal angle), the *elevation* (the vertical angle) and the *distance* (static sounds) or *velocity* (sounds that are moving) [32]. These measurements are illustrated in figure 4.1. In this illustration two vertical axis are also shown: M refers to the axis aligned to the median of the subject and S refers to the vertical axis aligned to a side parallel with the subject's ear.

Auditory systems use several important cues to determine a sound's location including time and level differences between the signals arrival at each of the subject's ears.

Interaural Time Difference (ITD) is shown in figure 4.2. The figure illustrates how sound emitted from different sources reach both ears. The sound emitted from point (A) which is directly in front of the subject arrives at both ears at the same time. However when the tone is off to the side (B) it reaches the listener's right ear before it reaches the left. ITD applies to low frequency localization for sounds that are less than approximately 1500Hz. The average distance between human ears is 20cm resulting in a 600 microsecond delay between the incident sound in one ear and hearing in the other.

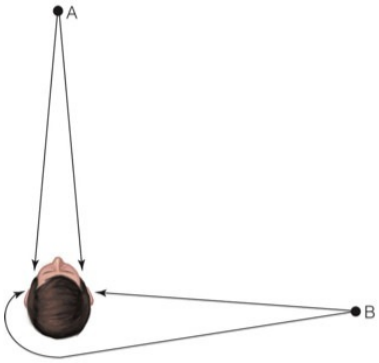

Figure 4.2: Diagramed interaural time differences.

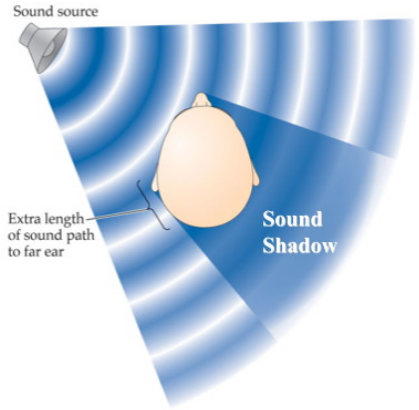

Figure 4.3: Diagram of sound level differences.

Interaural Intensity Delay (IID), like ITD is also dependent on the frequency of the sound emitted. If a sounds wavelength is equal to or greater than the listener's head then the sound will diffract around the head and be heard with the same intensity as the incident wave. What is being displayed in figure 4.3 is a high frequency sound where absorption of the sound energy occurs by a solid medium (in this case, the subject's head). This absorption of sound energy is called a sound shadow because there is effectively zero sound energy from the original source in that space.

4.2 Binaural Audio

Binaural audio is simply audio that is engineered with the intention of creating a 3-D stereo sound sensation for a listener. The goal is often to

simulate the listener's actual presence in a virtual environment where the recording was actually made. This acoustic virtualization can be accomplished with the use of specialized hardware or signal processing techniques.

The only requirement to produce three dimensional audio systems or to render sound images around a listener is stereo sound output from either headphones or loud speakers [35]. In the case of 3D audio systems that usie 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.

Transaural audio is the name for the technique of delivering 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 given the reverberations of the listener's environment. This technique was first put into practice by Shroeder and Atal [30, 31]. It is possible to produce the accuracy of binaural audio from headphones through the use of loudspeakers, as is often observed with high end sound systems. Systems capable of calibrating the audio to a moving user with stationary stereo speakers in an open environment with the aid of head tracking webcams have been developed. In system, researchers have demonstrated how both the audio transformations necessary to mimic the physics of the 3D sound waves and the placement of the virtual sound sources relative to the listener can be updated in realtime using only loudspeakers [34].

The ultimate goal of binaural audio systems are to perfectly calibrate sound placement to create an experience that provides the user with the perception that sound is placed in a 3D environment that exists around them.

4.2.1 Binaural Solutions

Binaural audio has a history dating back to 1881 where an array of microphones were installed on the front edge of the Opera Garnier allowing telephone subscribers to enjoy the music through their telephones with specialized headsets [15]. Since then, the novelty of the technology has waxed and waned with the introduction of the radio, television, and personal walkmans. Binaural audio is experiencing a resurgence in popularity, specifically with the audiophile communities as headphones have become cheaper.

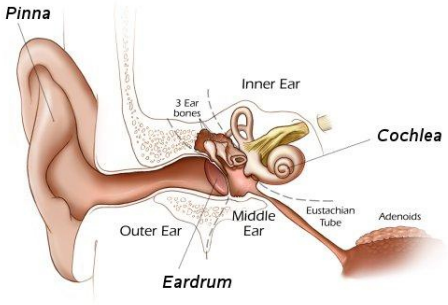

Figure 4.4: Anatomy of the ear.

Figure 4.5: Binaural microphone with the exact dimension and density as a human head. There are two input channels corresponding exactly to the location of human eardrums.

There are primarily two ways to produce binaural audio. The audio can either be generated or recorded. Figure 4.5 is a binaural microphone where two high-fidelity microphones are mounted inside a dummy head inset in ear shaped molds allowing the system to fully capture all of the audio frequency adjustments that happen naturally as sound is warped by the human head. There exist many variations of the microphone, each targeting different types of output (e.g. playback on headphones versus loud speakers).

The second method to produce binaural audio is through the actual manipulation of two sound channels using a **Head Related Transfer Function** (HRTF). The general idea is to reproduce the acoustic transformations that would normally occur at the listener's ears in a natural listening situation. Specifically, the HRTF describes exactly how a given sound wave input, originating from some location and having some frequency, is filtered by the diffraction and reflection properties of the head, pinna, and torso before the sound reaches the mechanical parts of the eardrum and inner ear (figure 4.4).

This process is accomplished by convolving each source signal with a pair of HRTFs corresponding to the sound sources intended location. The resulting signal is presented to the user through headphones. Figure 4.6 demonstrates the spatialization of a single sound source from an arbitrary distance and azimuth. The direction of the source (θ = azimuth, ϕ = elevation) determines which pair of HRTFs to use and the distance (r) determines the gain. Figure

4.7 demonstrates how to spatialize multiple sound sources with constant level reverberation to enhance the listener's perception of distance [8].

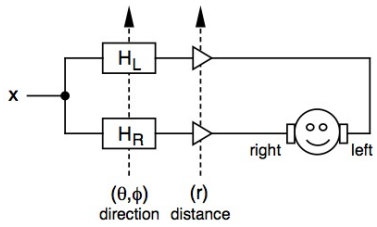

Figure 4.6: Single source bianaural spatializer

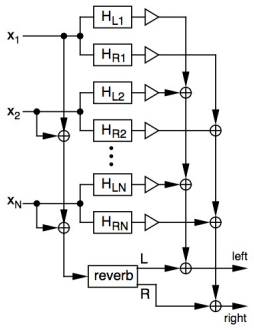

Figure 4.7: Multiple source binaural spatializer.

4.2.2 Binaural Issues

Given that humans are capable of processing sound signals to place the audio source in space, either prerecorded or simulated sounds used to produce the auditory experience of one or more sound sources arbitrarily located around a listener are subject to certain constraints. Once recorded or produced, the spatialization requires stereo playback. To accurately reproduce the effect of hearing a sound in person given the location of objects, user specific HRTFs should be calibrated to the individual user. In practice, average HRTFs work well enough for most young adults [36]. Secondly, the concept of externalization which is the placement of the sound sources in the 3D plane suffers when users use in-ear headphones. What can occur with in ear headphones is that the 3D plane is perceived to be inside of the head (a process called internalization).

Other issues include a user's ability to perceive sound elevations.

Related Work

Before constructing a binaural interface, it is important to decompose the proposed system into its key components. An interface such as this would be comprised of two main elements. The first and most tangible of which would be a pseudo intelligent agent as described by Eric Horwitz. Such an agent would be able to classify the importance of specific notifications, allowing the system to discern when and how to notify the user of an event. Psychologically, this agent would act as the mind for the conversational behavior of the system. The second component would be the binaural engine. To understand the interplay of these two components we turn to prior research exploring interface design, the psychology of certain notification modals, accessibility research and 3D audio. Again, the following section aims to organize previous research in this space as well as demonstrating the benefit arising from the interplay of these components which will be explored in the future work.

5.1 Audio Based Accessibility Solutions

There are many existing solutions that enable blind users to access computer interfaces. Screen readers such as JAWS and Windows-Eyes are proprietary special purpose software solutions that cost more than $1000 per installation. There are free alternative solutions such as NVDA, Orca and the Linux Screen Reader available to users on Linux and Windows [3]. The Macintosh operating system comes standard with Voice Over and is available on all Macintosh computers and iOS devices. There are other solutions available for web content and browsers as well as custom hardware solutions aimed at providing accessibility solutions. But these products that provide accessible solutions only produce monaural audio and provide no audio localization.

T.V. Raman demonstrated enumerated the differences between speech output interfaces and screen reading solutions, namely giving an application a voice (via Emacspeak) versus simply reading a screen. Emacspeak is intended to give a voice to the terminal, not simply allow a machine to read what it is displaying. The key difference being that as a subsystem, Emacspeak has

context specific information available to it that many screen readers simply cannot duplicate [27].

To this effect, the ultimate goal for this project is to create a context aware operating system that is able to vocalize each task in a multi-tasking environment in a natural manner.

5.2 HCI : Interface Design

Interface designers focus on three main questions:

Who is the target user?

What is the task the user is trying to accomplish?

How can you make an interface that is the most sensible for this task?

These design principles were established to make computers more usable and receptive to a user's immediate needs. The goal of designing the best interface with given constraints, simultaneously optimizing for a desired property (such as efficiency or learn-ability) is not a trivial undertaking.

The field of HCI as such, experiments with new design methodologies and hardware, prototyping new systems, and exploring new paradigms for interaction to minimize the barriers between the human's cognitive model of what they are trying to accomplish and the computer's representation of the task [2]. The following sections explore findings of each goal as presented in the current research. What is important to note is that this project, though geared towards users with visual disabilities, adheres to the mantra of maximizing human cognitive and physical abilities using the guidelines found in [28].

5.3 Psychology of Interface Design

The ultimate goal for user interface design is to mediate effective operation. Before you can begin looking at the effectiveness of a particular operation or efficiency for a specific interface, metrics have to be established for a

the desired or given task. The goal of our future work is to increase the ability for blind or visually disabled users to multi task on a given computing environment.

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 a conversational metaphor would be an appropriate interface design choice because the user could leverage existing biological mechanisms to understand the interface and that the existing conversational mental model directly supports direct and simple inference of what information this interface is presenting [16].

5.3.1 Mental Models and Cognitive Loads

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 [18].

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

Exploring a users collective preference for simpler representations of information. For example, traveling users preferred summarized directions that incorporated familiar paths to more detailed and exact routing instructions. The users in the study actually valued traveling on familiar routes enough to sacrifice overall trip distance [26].

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 [17].

5.3.2 Mental Model Elasticity

Since Irys is an exploration of a new interaction technique, it's important to consider user's ability to learn how the system works. Work in HCI explores how well users are at creating new mental models for interactions as evidenced by shifting paradigms of technology hardware (changing from numeric keypads, to stylus, to keyboard, to touchscreen). Work from the Universitite Paris Sud explored how well users could remap existing mental models by exploring how users adapt to interaction techniques that use the temporal dimension versus more advanced uses of rhythmic patterns to convey information. The researchers explored how rhythmic techniques, though a deviation from user expectation of temporal interactions (such as multiple clicks), can be particularly useful in situations where the visual channel is overloaded or even not available. They concluded that (i) rhythmic patterns can be efficiently reproduced by novice users and recognized by computer algorithms, and (ii) rhythmic patterns can be memorized as efficiently as traditional shortcuts when associating them with visual command [9]. This work suggests that deviations from common interaction patterns can be beneficial to users when training is provided.

5.4 Immersive Interfaces

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.

One of the guiding principles for this interface is providing the user with the abstraction that they are immersed in their interface by placing them in a 3D virtual audio environment. One of the motivations for such a design choice is our natural ability to discern the location of sound around us. Another, perhaps more compelling reason to provide such an interface are the added

benefits of immersive virtual environments. Kayur Patel et al. defined a fully immersive virtual reality setting to be an extension of virtual reality. Specifically, immersive virtual reality captures a user's full body motion while immersing the subject into the virtual environment. Their exploration of such an environment for educational purposes demonstrated that the increased interactions led to more effective education in a variety of tasks [25].

Other works confirm the beneficial effects of immersive audio through the use of binaural audio.MIT's Media Lab chose to incorporate binaural and transaural audio for the Artificial Life Interactive Video Environment to augment the interface [5]. Researchers at Siemen Corporate Labs demonstrated the comprehensive benefits of an HTML page's structure when provided with an interactive browser using binaural audio to place elements of a page around a user [10].

Research from the University of Glasgow researched the ergonomics of different tactile interfaces to a 3D audio soundscape. They demonstrated how several interaction patterns (gestural on mobile devices, to physical motions such as head nodding) might work with a given soundscape measuring the tactile interfaces accuracy and effectiveness. They concluded that a 3D soundscape was effective with all of the tested interfaces and could be used as a flexible, eyes free interaction with a computer [22].

5.5 Interface Elements

Eric Horvitz depicted the divide in research focusing on the promise of creating new metaphors and tools that enhance a user's ability to directly manipulate objects *versus* an applications ability to provide automation. In his system LookOut, Horvitz explored the role of a user interface in guiding the user to task completion. He explained the benefits of building an interface that tailored the computer to drive task completions through what he coined as "Mixed Initiative" actions. He demonstrated how different modalities of interactions could allow a system to overcome uncertainty while delighting the user with a percieved automated experience that performed repetitive tasks for the user [14]. This application demonstrated how user interfaces could adapt to noisy data, such as uncertainty about times referring to an event that needs to be scheduled in an email. The influence LookOut has

on modern interfaces can be seen with many common email clients such as Google's Gmail and Microsoft's Outlook performing similar mixed-initiative actions. In relation to Irys, the system will have to learn how to notify users of changes by evaluating the costs and benefits associated with interruptions at different times.

5.5.1 Visual Notifications

While performing a task, it is important that an interface assists the user to work efficiently. Certain assumptions are made between the user and the interface as they are in fact entering a working relationship. When using an interface, users inherently trust that the system will assist them in completing tasks. 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 notifications [19].

The study of information design and the options that are suitable for three often conflicting design objectives for notifications - interruption to primary tasks, reactions to specific notifications, and comprehension of information over time - are necessary to create the most efficient computing environment. Understanding the impact visual notifications have on a user directly inform the decisions made for an audible interface.

Given the stochastic nature of notifications, probabilistic reliability can be used to create intelligent messaging systems that gain a user's trust, rather than lose it [19]. Empirical evidence show that users will ignore all notifications that are not highly valid when performing demanding visual tasks [21], but there are also findings that demonstrate that a notifications timing determines the amount of disruption it causes a user [6]. In a study, Microsoft Researchers found that notifications occurring at the beginning of a task are more likely to completely derail focus while those occurring near the completion of tasks have mitigated effects.

5.6 Tactile Notifications

When exploring non-visual interfaces, it is important to consider how the human mind operates when interfaces utilize other sensation modalities. Consider interfaces that utilize haptic feedback. It has been shown that for many dynamic tasks (such as remote control of robots), haptic interfaces are able to provide the user with important types of feedback. For example, a user is able to detect the same forces that might impede the operation of the target object. In other situations, users may prefer haptic feedback to audible interactions when in social situations where audible interruptions are discouraged. Other uses of haptic interfaces are seen in high precision tasks (such as surgery) where it is important to apply the precise torques and forces on both ends. Common to all of these interfaces is the purpose of the device is human actuation, often at a physically smaller interface [7]. There has been work looking at the psychophysical use of power amplification as seen in prosthetic engineering, but those are not considered for this project.

What is of note is that interfaces that utilize haptic feedback are often set in a master/slave relationship between the user and the object on the other end of the interface. Complexity is often measured in degrees of freedom that the operator has over the slave device. What is important with haptic feedback is that notifications have to be performed in real-time or else stale information propagates into large control errors [7].

5.7 Auditory Notifications

SOCIAL ISSUES: In their paper Hanson et al. discussed the interplay of social situations and auditory cues [12] Current auditory notifications cues can be attention demanding, distinct, and can be perceived 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 [12]. Nitin Sawhney performed an empirical study demonstrating how existing devices with audible notifications (read: pagers and phones) can detract from an owner's concentration throughout a day. Their work demonstrated that there was around 10 minutes per hour spent on interruption handling by the average user in their study. To address this issue,

researchers at MIT created Nomadic Radio, a physical device that would infer when the ideal time to present a notification to the user by listening to the environment [29]. In this paper, the researchers provide a mechanism for scaling a cue to the user providing the user with the ability to prevent an interruption. Other techniques are presented to allow the user to quickly identify senders based on items such as auditory signatures. Of interest is the fact that the ambient auditory introduction was the most requested feature as user's cited the least cognitive effort required to anticipate the subject or content of the new notification.

Our Approach

TOWARDS AN UBIQUITOUS BINAURAL AUDIO INTERFACE

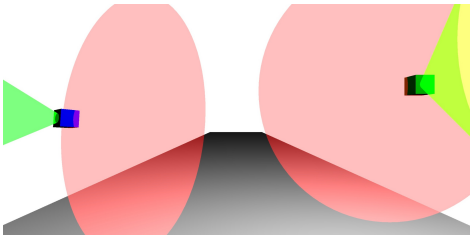

Figure 6.1: Sample rendition of sound images in 3D space

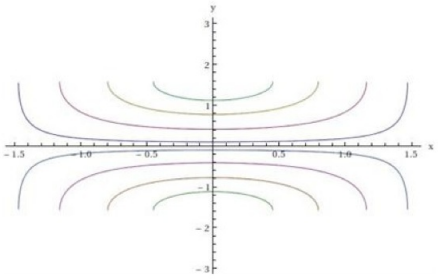

Figure 6.2: Sample sound paths tested

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 allowing users to interact with technology when visual attention is either not available or not desired.

Conventional binaural or transaural audio systems works well if the listener is stationary 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 a sweet spot, system performance degrades rapidly (though headphones alleviate these problems).

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 canceler matrix H. The updates of B and H were referred as dynamic binaural synthesis and dynamic crosstalk canceler, respectively and were presented in [20].

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, but rather this is left for future work.

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 6.1. The goal of this project is to create a virtual environment that allows sound images to be placed arbitrarily around the user, depicting content audibly. The following sections enumerate the methods and project goals for continued research.

6.0.1 Analytics

Before implementing many of the features discussed in this area paper, developing an understanding of the ways blind users interact with the web and other interfaces would be invaluable to this work. In that regard, I will carry on work started by Jeffrey Bigham in reverse engineering Text-to-Speech so as to develop a probabilistic model of where a user listening to a TTS engine is located. The ground work for this project has been completed by training machines to recognize speech synthesized from the Macintosh Text-to-Speech engines. There is much work left implement a particle filter so as to allow a system to develop a speech tracking mechanism for a given interface. Using particle filters as simulation-based probabilistic approximations for tracking has been proven effective in the robotics domain [13] and we hope this technique informs much of the work that will be performed in the future interface design.

6.0.2 Notifications

The essence of any event or update to user data is condensed 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 perspectives that overlap with our ability to handle interruptions and distractions [6, 23].

6.0.3 Current Results and Frameworks

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.

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, I found that OpenAL provided exciting abstractions, but distance was only provided by volume amplitude attenuation and not properly calculated with a delay.

Creating a native desktop, iOS and android application using OpenAL was 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 traveling 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 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 specific tasks.

The targeted user base for this interface are members of the blind or low vision community. We 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 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 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, were very excited to explore the ability of 3D audio in helping users remember information by providing a tangible dimension to their information processing.

7.1 Artificial Intelligence Domain

Before exploring the contributions of such an interface to the field of human and computer interaction, there are a number of technical implementations that would contribute to artificial intelligence. Prior literature has demonstrated that acute attenuations are miscalibrated in a user's head related transfer function, then the auditory experience degrades sharply [1]. One contribution to this domain would be a systemic learning procedure that is able to model the parameters needed for an individual's head related transfer function. The CIPIC HRTF database provides high resolution anthropometric measurements and will act as a strong source of data for the parameter tuning for such a classifier. Manual Techniques for calibrating HRTF were presented by researchers from Creative Audio and would provide a great stepping stone for this [15].

7.2 Crowd Sourced Domain

Crowdsourced work has proven to be an effective medium for solving hard AI problems, computer vision and other computationally ambiguous problems. With a binaural interface, a new control mechanism could be created that would allow other users to provide navigational cues where a full audible stream may not be beneficial. Examples that demonstrate how to utilize audio as a control have been the dropping of digital soundtracks in the augmented reality space. Other uses for this could be

7.3 Analytics Domain

Researchers have shown that the use of analytics such as eye-tracking for visually enabled users looking at search data provided rich insights into the user's intent, the page's efficiencies, and the user's goals when browsing search results. By looking at the time the eye focuses on a given web page, they were able to extract rich features impacting areas such as usability, interface design, and other valuable aspects [11]. To this we ask whether the same insights can be gained by observing how a non-sighted user interacts with content. The

implementation of the particle filtering algorithm to infer location is currently the set as the first step for this project.

7.4 Cultural Domain

How does culture, age, or language affect an individual's choice or mode of communication? As this interface is primarily targeted to providing auditory speech-like interactions between a number of running processes and applications to a user, communication preferences begin to play an important role in determining the surveys have shown that culture influences communication preferences for patients seeking treatment at the end of life stage [33].

7.5 Human Computer Interaction

The study of Human Computer Interaction has very well established guidelines and methods for evaluating graphic interfaces. There are over 13 commonly accepted design principles encompassing many facets of visual perception and information processing [37]. Expanding on this research in the auditory domain with empirical studies on user satisfaction with a

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, artificial intelligence, and human computer interaction to provide users 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 when interacting with machines.

Bibliography

[1] V Ralph Algazi, Richard O Duda, Dennis M Thompson, and Carlos Avendano. The cipic hrtf database. In *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, pages 99–102. IEEE, 2001.

[2] Ronald M Baecker and William A.S. Buxton. Human-computer interaction. *A multidisciplinary up*, 1987.

[3] Jeffrey P. Bigham, Craig M. Prince, and Richard E. Ladner. Webanywhere: a screen reader on-the-go. In *Proceedings of the 2008 international cross-disciplinary conference on Web accessibility (W4A)*, W4A '08, pages 73–82, New York, NY, USA, 2008. ACM.

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

[5] Michael A. Casey, William G. Gardner, and Sumit Basu. Vision steered beam-forming and transaural rendering for the artificial life interactive video environment (alive). In *Audio Engineering Society Convention 99*, 10 1995.

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

[7] RE Ellis, OM Ismaeil, and MG Lipsett. Design and evaluation of a high-performance haptic interface. *Robotica*, 14(3):321–328, 1996.

[8] William G. Gardner. Transaural 3-d audio. *M.I.T Media Laboratory Perceptual Computing Section Technical Report*, (342), 1995.

[9] Emilien Ghomi, Guillaume Faure, Stéphane Huot, Olivier Chapuis, and Michel Beaudouin-Lafon. Using rhythmic patterns as an input method. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, pages 1253–1262. ACM, 2012.

[10] Stuart Goose and Carsten Möller. A 3d audio only interactive web browser: using spatialization to convey hypermedia document structure. In *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, pages 363–371. ACM, 1999.

[11] Laura A Granka, Thorsten Joachims, and Geri Gay. Eye-tracking analysis of user behavior in www search. 2004.

[12] 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.

[13] Jeffrey Hightower and Gaetano Borriello. Particle filters for location estimation in ubiquitous computing: A case study. In *UbiComp 2004: Ubiquitous Computing*, pages 88–106. Springer, 2004.

[14] Eric Horvitz. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems: the CHI is the limit*, pages 159–166. ACM, 1999.

[15] Adrian Jost and Jean-Marc Jot. Transaural 3d audio with user-controlled calibration. In *Proc. Of the COST G-6 Conference on Digital Audio Effects (DAFx-00), Verona, Italy*, 2000.

[16] David E Kieras and Susan Bovair. The role of a mental model in learning to operate a device. *Cognitive science*, 8(3):255–273, 1984.

[17] 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.

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

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

[20] Tobias Lentz. Dynamic crosstalk cancellation for binaural synthesis in virtual reality environments. *Journal of the Audio Engineering Society*, 54(4):283–294, 2006.

[21] 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.

[22] Georgios Marentakis and Stephen A Brewster. A study on gestural interaction with a 3d audio display. In *Mobile Human-Computer Interaction-MobileHCI 2004*, pages 180–191. Springer, 2004.

[23] 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¿.

[24] 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.

[25] 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.

[26] 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.

[27] TV Raman. Emacspeakdirect speech access. In *Proceedings of the second annual ACM conference on Assistive technologies*, pages 32–36. ACM, 1996.

[28] Leah M Reeves, Jennifer Lai, James A Larson, Sharon Oviatt, TS Balaji, Stéphanie Buisine, Penny Collings, Phil Cohen, Ben Kraal, Jean-Claude Martin, et al. Guidelines for multimodal user interface design. *Communications of the ACM*, 47(1):57–59, 2004.

[29] Nitin Sawhney and Chris Schmandt. Nomadic radio: scaleable and contextual notification for wearable audio messaging. In *Proceedings of*

the SIGCHI conference on Human factors in computing systems, pages 96–103. ACM, 1999.

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

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

[32] Brian Shannan. Audiology update, November 2010.

[33] William H Shrank, Jean S Kutner, Terri Richardson, Richard A Mularski, Stacy Fischer, and Marjorie Kagawa-Singer. Focus group findings about the influence of culture on communication preferences in end-of-life care. *Journal of general internal medicine*, 20(8):703–709, 2005.

[34] 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.

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

[36] Elizabeth M Wenzel, Marianne Arruda, Doris J Kistler, and Frederic L Wightman. Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America*, 94:111, 1993.

[37] Christopher D. Wickens, John D. Lee, Yili Liu, and Sallie E. Gordon Becker. *An Introduction to Human Factors Engineering.* Pearson Prentice Hall, 2004.

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