The University of Rochester

Area Paper
for the Degree Master of Science

Binaural Audio as a User Interface

Author:
Androwis Abumoussa

Supervisor:
Dr. Henry Kautz

July 13, 2013

Contents

1 Acknowledgements **iv**

2 Abstract **v**

3 Introduction **1**
- 3.0.1 Motivation . . . 1
- 3.0.2 Project . . . 2

4 Background **5**
- 4.1 Sound Localization . . . 5
- 4.2 Binaural Audio . . . 7
  - 4.2.1 Binaural Solutions . . . 8
  - 4.2.2 Binaural Issues . . . 9
  - 4.2.3 Accessibility . . . 10

5 Related Work **11**
- 5.1 Audio Only Interfaces . . . 11
- 5.2 Assistive Audio Interfaces . . . 12
- 5.3 Binaural Interfaces . . . 13
- 5.4 Binaural Audio Interfaces with other Modals . . . 16
- 5.5 HCI : Interface Design . . . 16
- 5.6 Psychology of Interface Design . . . 17
  - 5.6.1 Mental Models and Cognitive Loads . . . 17
  - 5.6.2 Mental Model Elasticity . . . 18
- 5.7 Interface Elements . . . 19
  - 5.7.1 Visual Notifications . . . 19
- 5.8 Tactile Notifications . . . 20
- 5.9 Auditory Notifications . . . 21

6 Our Approach **22**
6.0.1 Analytics 23
6.0.2 Notifications 24
6.0.3 Current Results and Frameworks 24

7 Future Work **26**
7.1 Empirical Measure of Efficiency 26
7.2 Artificial Intelligence Domain 27
7.3 Crowd Sourced Domain 27
7.4 Analytics Domain 28
7.5 Cultural Domain 28

8 Conclusions **29**

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. Henry Kautz, whom 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 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.

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. The use of this technology has been explored in the context of user interface design but has not been implemented in any mainstream accessible user interfaces.

This paper explores previously researched conceptual models of binaural audio interfaces and intelligent agent designs. The goal is to integrate the theory behind intelligent notification systems and positional audio to assist users in multi-tasking environments. This paper demonstrates that the interplay of these types of systems offers the potential of increased productivity, smoother transitions, and a more fluid user experience.

Introduction

This paper focuses on the roles of audio and notifications generated by intelligent agents within user interface design. It begins by reviewing previous findings of HCI research where audio is used as *the* primary element of an interface as explored by Arons et al [2]. Attention is then shifted to the work by Goose et al who used binaural audio as an interface for HTML presentation [10]. Finally we explore a number of works which have looked at binaural audio's role within interfaces incorporating other modals [24, 41].

Our work continues with the exploration of the role notifications have within interface design. Experiments have shown that the impact of variable reliability in a notification system biases users to ignore future notifications [21]. This paper examines the differences between two types of notifications: informative [23] and interpreted [14]. We also explore the effective and judicial use of notification systems to inform future designs of our work [6].

The structure of this exploration is motivated by audible interfaces' ability to provide an alternate mode of access to computers whenever a user's visual focus is unavailable. Visual focus can be unavailable for a number of reasons, such as when users are driving, using hand-held devices while engaged in other physical activities that command your visual attention, or even as a result ofphysical disabilities [26]. There are a number of audible interfaces which only provide a text-to-speech component allowing systems to read the content of a visual interface to a user. Text-to-speech solutions often implement a monoaural speech pattern that provides a user with a single channel of communication. Techniques exist to provide spatially placed (3D / binaural) audio to listeners. Our interest is in the use of spatial sound as a main interface element for an audio interface.

3.0.1 Motivation

T.V. Raman enumerated the differences between speech output interfaces and screen reading solutions . He argued that the key difference was the system's ability to give an application a voice (like Emacspeak) versus simply reading a screen. Raman was introducing Emacspeak, an application intended to

give a voice to the terminal, not simply allowing a machine to read what it is displaying. The precise difference he was making was demonstrating that in creating Emacspeak as a subsystem of Emacs, Emacspeak had context available to it that many screen readers simply cannot duplicate [28].

To this effect, our project researches how binaural audio and notification theory can be combined to create new user interface designs. Our paper begins to lay the foundation for future work by studying the roles of four particular combinations of speech (ambient versus direct) and intelligent agent notifications (purely informative versus interpretive). With our work we are exploring the ways in which sound and notifications can be combined by different types of intelligent agents to create interfaces with differing goals. We will explore the benefits of context aware interfaces and the impact vocalization has on multi-tasking environments.

We motivate our work by using a common scenario like driving to work on a typical morning. Eric Horvitz [CITE] described how intelligent interfaces could be utilized these situations. We ask the reader to imagine a world where a user has a number of assistants each of which are smart and effective communicators. These assistants could be orchestrated either by the user or the interface. A typical interaction might be so that the computer powering the navigation system is also checking current road conditions relative to your location. A smartphone has resumed polling the work email account, your calendar has been updated by a colleague, and your family is messaging you with a reminder for a prior engagement. With each type of stimuli, there is a stochastic decision that must be maderegarding both the best manner and the proper time to alert the user. In his work, Horvitz quantified the risks and benefits of using different notification paradigms to express this information [CITE]. We explore how these trade offs present themselves in the auditory domain.

3.0.2 Project

We describe a system that supports a new conceptual model which maps interface elements into a 3D audio space. Using binaural audio as the mechanism, novel features are explored that provide information to the user in terms of spatial attentuation, audio structural survey of content on the web, accurate positional audio feedback and an audible progress indicator. These

new features may improve both the users' comprehension of content presented to them while providing users with cues to assist in the recall of information.

Human auditory localization has been studied extensively [5, 40]. Because humans are especially adept at localizing sounds in three dimensions we want to explore how predominant auditory cues for determining whether a sound is coming from the left or right directions can be used to portray other types of information.

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. This will be the mechanism powering the spatialization of this interface.

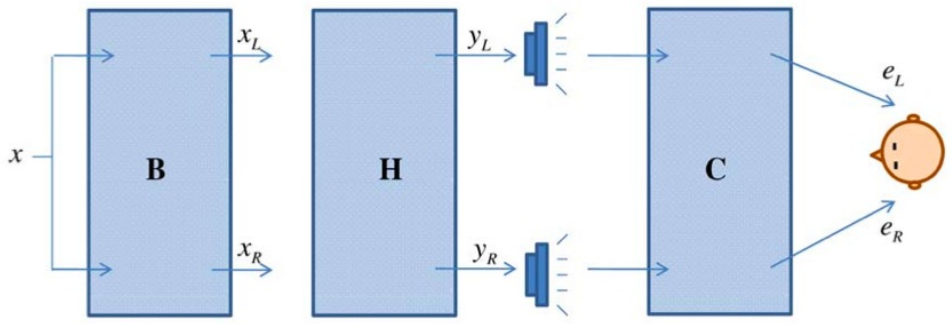

Figure 3.1: Schematic of a binaural audio system

This thesis begins explore questions of efficiency of different types of intelligent agents and how auditory interfaces can be used to achieve multi-tasking. While exploring the research in the utilization of audible interfaces, this paper works towards two major goals. First a review of the prior 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 are provided in the following section. The related work connects ideas and findings both in the fields of Human Computer Interaction and Artificial Intelligence to form the basis for 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 could 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.

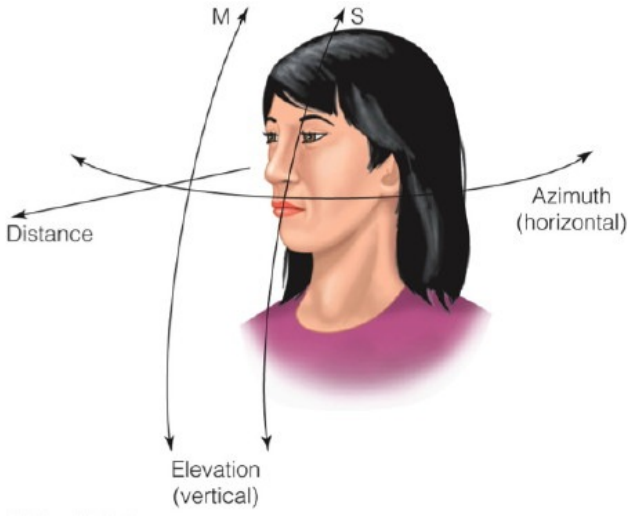

Figure 4.1: Measurements used for spatial sound processing analysis.

The brain is able to utilize subtle differences in a sound's intensities, spectral and timing cues to locate the origin of a given sound source. Often the measurements used to place a sound relative to a subject are the *azimuth* (the horizontal angle), the *elevation* (the vertical angle) and the *distance* (for static sounds) or *velocity* (for sounds that are moving) [33]. These measurements are illustrated in figure 4.1. Also portrayed in this illustration are two vertical axis: 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.

The primary mechanisms that an auditory system uses to determine a sound's location including time and level differences between the signals arrival at each of the subject's ears are illustrated below:

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. Interaural time differences is precisely the time difference that the ears perceive the same sound. ITD applies to low frequency localization for sounds that are less than approximately 1500Hz. The average distance between human ears is 20 cm resulting in a 600 microsecond delay between the incident sound in one ear and hearing in the other for a sound emitted from point A. As the azimuth changes, the delays increase, and provide a feature that the brain can use to place sound.

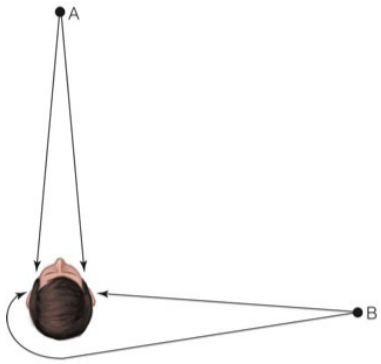

Figure 4.2: Diagramed interaural time differences.

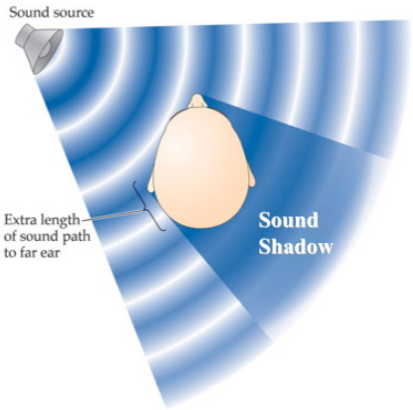

Figure 4.3: Diagram of sound level differences.

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. The difference in magnitude of the frequency perceived by both ears is what is measured by interaural intensity delays.

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 [36]. In the case of 3D audio systems that use 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 primary difference between this technique and binaural audio is that transaural filters must take into account room reverberations. Transaural audio filters a binaural signal such that the subject can process the subsequent stereo representation as a 3D signal given the reverberations of the listener's environment. This technique was first put into practice by Shroeder and Atal [31, 32]. 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 web cams have been demonstrated. Researchers have shown 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 real-time using only loudspeakers [35].

Ultimately, both binaural audio and transaural audio systems aim 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 [17]. 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 within the audiophile communities as headphones have become cheaper and capable of producing higher quality audio.

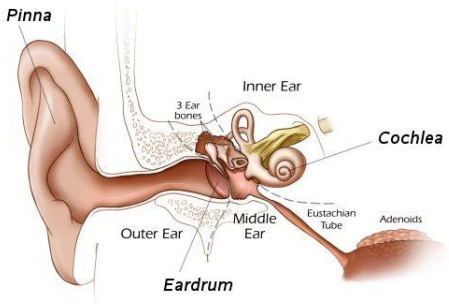

Figure 4.4: Anatomy of the ear.

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 using signal transformations 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].

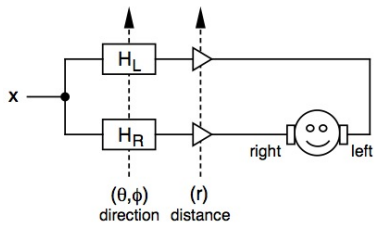

Figure 4.6: Single source bianaural spatializer

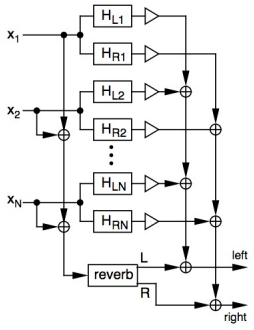

Figure 4.7: Multiple source binaural spatializer.

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 [38]. 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 minor issues include the variance between individual user's in their ability to perceive sound elevations.

4.2.3 Accessibility

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. Free alternative solutions do exist, such as NVDA, Orca and the Linux Screen Reader available to users on Linux and Windows [4]. 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.

Related Work

The following section aims to organize previous research enumerating aspects of both binaural audio and intelligent agents. By exploring prior work in this space, we begin to observe the benefits arising from the interplay of these components which will be explored in future work.

5.1 Audio Only Interfaces

MIT's Media Lab has explored the area of "hyperspeech", a corollary to hypertext where the paradigm is presenting speech as data. The goal of the work was to understand the nuances necessary to both extract information from audio while also accessing it through speech [2].

Though speech is an incredible tool for communication, its temporal existence prevents humans from browsing sounds the way eye can scan text. Speech and audio interfaces must be sequential, while visual interfaces can be simultaneous in nature. Therefore, designing an interface focused solely on sound and speech, rather than text, video, or graphics presents a number of nuanced challenges. Navigation with speech is much more difficult as instructions are inherently fleeting, and as Arons demonstrated "one cannot simply 'click here' in the audio domain, for as the selection is made, time has passed and here has disappeared" [2].

To introduce hyperspeech, or the use of sound as data, Arons conducted interviews with a number of leading interface design specialists. The interviews were brief in nature, completed in approximately 5 minutes. The speech recorded from the interview was then processed off line and put into a data base. Labels and connections were created between associated statements, creating a graph of the topics discussed. What was studied was the interface with which hyperspeech could be used. The goal of the interface was to answer questions such as, "How do persons A and B feel about C?" or "Who disagrees with person A's stance on B?". Navigation between different topics or through specific interfaces were studied.

By exploring how the design can influence a user's ability to navigate through the information of a database, it was demonstrated that it is important to be able to navigate in any direction arbitrarily from any node. Having all of the commands available while navigating was the most natural presentation of an interface. One benefit of using speech as an input modality is that it is goal directed and very succinct. Complexity and variability proved to be a difficulty when creating this interface.

5.2 Assistive Audio Interfaces

Work by researchers at Georgia Tech demonstrated the use of a single, radial axis in an audio interface to present information to blind individuals. They presented multiple user studies and work of assisting blind users navigate new or unfamiliar rooms. With smart phones and computer vision techniques the researchers in Bruce Walker's Sonification lab created an application helping blind users understand the orientation and placement of objects in new environments. Of note are the current practices that blind users employ to navigate these settings. The focus group Jeon et al. interviewed demonstrated that blind users employ various strategies to place objects in a room such as one subjects summarization that by "listening to the room, we can detect where running refrigerators, computers, or other machinery exists" [16]. The second most common response was that layout is also often inferred by touch, or assisted navigation either with the use of inanimate objects (such as walking canes) or sighted volunteers.

The smart phone application presented in this work allows visually handicapped individuals to scan a room by holding an external camera connected to the phone, or a phone's native camera left to right across the room. The application then provides the user with 3 modes of information retrieval after video processing has been performed to identify the objects and their locations in the environment.

1) Linear Searching : allows the interface to communicate which objects are present in a room and the order in which they were observed. The authors found that this mode allowed users to quickly "scan" a setting and identify whether an object of interest existed in the space.

2) Directional Retrieval: provides the users of the interface with information on objects by grouping related objects and returning their location with respect to room walls and room center.

3) Spatially, using 3D audio, the app vocalizes the position of an object through the use of binaural audio when the user is wearing a headset. This allows the user to understand where objects are in relation to their current orientation.

Additionally, the lab has explored optimizations for auditory interfaces. The lab demonstrated how the use of spearcons (injections of compressed speech [15] that allows the sound to act as a cue within a given context) can be used to enhance menu navigation on mobile devices [37]. This work extends previous findings that auditory menus that rely solely on TTS vocalizations of the text prove to be much less efficient than menus that employ spearcons which allow the user to navigate based on injections and rely on the full speech for context. This work demonstrates how new constructs for navigation can be learned by users, and then used to optimize the navigation of auditory menus. Using a Nokia phone with 50 contacts with randomized names, Walker et al. studied how 89 undergraduates navigated to a target user using the visual display only, visual display with TTS, visual display with spearcon, visuals off with TTS, and then TTS with spearcons. The results show that conditions with visual cues led to the fastest responses. Once visuals are disabled, spearcons helped drive performance to the lower limits, while TTS solution initially afforded faster initial performance. With the same amount of practice, spearcons outperformed TTS engines due to their compressed nature. This work demonstrates some of the possible optimizations that can be built in to auditory interfaces.

5.3 Binaural Interfaces

Researchers at Siemens Corporate Labs demonstrated the comprehensive benefits of delivering HTML content to users with a binaural interface. They created an interactive browser using audio as the primary interface which placed page elements around a user, focusing on aiding the user maintain orientation while browsing on the web [10].

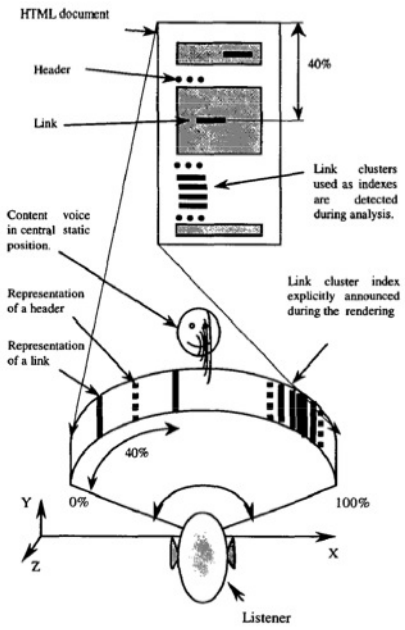

Figure 5.1: Mapping an HTML document horizontally

The work by Goose et al. was the fist work demonstrating that binaural techniques could be applied to non prerecorded media. The authors began by researching how well multiple users could spatialize sound presented to them from regular PC audio hardware to determine the best design choice for the interface. The observation that most impacted their design was the inability of most users to accurately identify the source of sound along the y-axis. This motivated Goose et al. to adopt a horizontal axis interface depicted in 5.1.

By showing the inefficiency of the y-axis to convey accurate information in the audio space, they demonstrated why an interface simulating the reading down of a page would be ineffective. Instead, by projecting a document onto the x-axis, Goose et al. effectively translated web pages into time lines. The interface could then compute the length of a given web page by calculating the time it would take to read.

The horizontal time domain allows the interface to communicate both inter-document and intra-document links effectively. To disambiguate the two forms of links, a "rhetoric of arrival and departure" was created where the interface

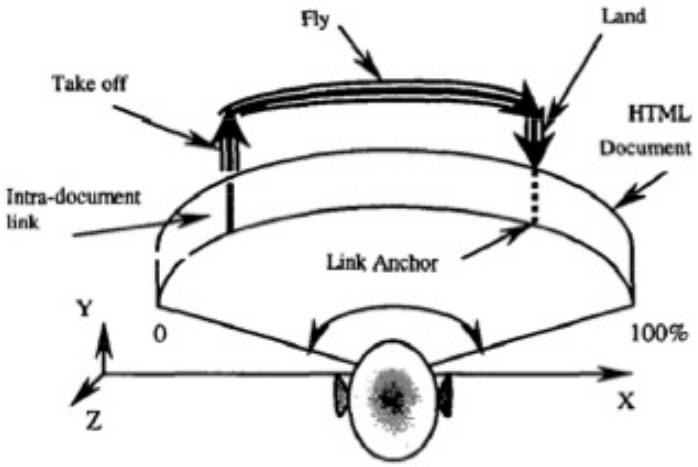

Figure 5.2: Binaural intra document traversal

would "take-off", "fly", and then "land" in the appropriate section of the horizontal display as depicted in figure 5.2. This rhetoric utilized unique and specific sounds that the user could associate with given transitions. With these unique transitions, later tests demonstrated that users could correctly identify the action performed by links, as well as maintain their orientation while navigating web pages every time. Secondly, the interface could exaggerate this rhetoric by playing the sound of a launch into space, and then a subsequent landing at the beginning of a new document to communicate inter-document transitions. This demonstrated the elasticity of the metaphor which was able to deliver appropriate locational feedback (similar to scroll bar's location indicating where you are on a page).

Goose et al. also demonstrated how a system can be engineered to prevent a user from listening to every single document to extract relevant information. The interface they built allowed for multiple rendering modes (one reads the document, one that only announced links within a document, one announcing section headers). By having multiple rendering engines, multiple voices could present different bits of information simultaneously to the user, allowing the user to focus on the elements of interest when arriving at a given page.

Goose et al provided early results which indicated that binaural audio improved the user's comprehension of the HTML document's structure and their

overall orientation within multiple pages. The authors argue that these factors can improve the effectiveness of the browsing experience for its users [10].

5.4 Binaural Audio Interfaces with other Modals

The viability of this work is exciting when considering how binaural audio can be used in conjunction with other interfaces. Research from the University of Glasgow noted the ergonomics and usability of different tactile interfaces when coupled with a 3D audio soundscape. Their work demonstrated how several interaction patterns (from gestural on mobile devices, to physical motions such as head nodding) might work with a given soundscape measuring the tactile interfaces accuracy and effectiveness. The researchers demonstrated some of the difficulties encountered when working with audio as an interface, citing the ambiguity of 'exact locations' in space when asking subjects to locate a sound. They observed that some of these ambiguities arose from blur introduced by non-calibrated HRTFs. Of note are the design choices Marentakis et al. made when integrating sound and visual displays that allowed for feedback to the user when guiding the user to interact with a specific sound source. They concluded that a 3D soundscape was effective with all of the tested interfaces and could be used as for flexible, eyes free interaction with a computer [24].

5.5 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 [3]. 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 [29].

5.6 Psychology of Interface Design

User interface design explores how systems can mediate effective operation. Before you can begin looking at the effectiveness of a particular operation or efficiency measure for a specific interface, metrics have to be established for the desired or given task. One goal for future work may be to increase the ability for a blind or visually disabled user to multi task in a given computing environment. In our work though, we are exploring how the use of 3D audio interfaces can be used to provide a new kind of interaction for users

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 [18].

5.6.1 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 more efficiently by reducing the high cognitive load needed for captioning [20].

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, Patel et al. demonstrated that users preferred summarized directions that incorporated familiar paths to more detailed and exact routing instructions. The users in the study valued traveling on familiar routes enough to sacrifice overall trip distance and travel time required [27].

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 [19].

5.6.2 Mental Model Elasticity

Since this paper 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 [9]. 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 commands [9]. This work suggests that deviations from common interaction patterns can be beneficial to users when training is provided.

5.7 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 perceived 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 our work, a binaural system with intelligent agent powered notifications will have to learn how to notify users of changes by evaluating the costs and benefits associated with interruptions at different times.

5.7.1 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 [21].

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 [21]. Empirical evidence show that users will ignore all notifications that are not highly valid when performing demanding visual tasks [23], 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.

5.8 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].

5.9 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 [30]. 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

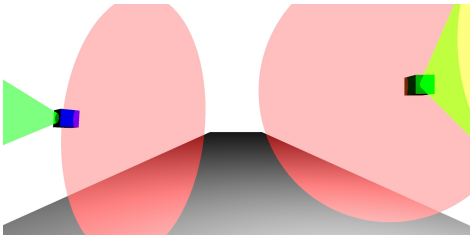

Figure 6.1: Sample rendition of sound images in 3D space

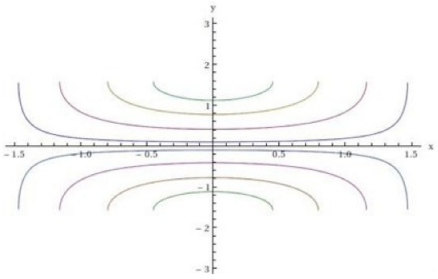

Figure 6.2: Sample sound paths tested

This project explores an audible interface designed for use primarily on the web with devices that can produce sound through headphones. The interface we are exploring places sound images around the user, providing them with a 3D sound environment allowing them to interact with technology when visual attention is either not available or not desired.

Conventional binaural or transaural audio systems work well if the listener is stationary at the position (usually along the perpendicular bisector of the two points of sound) as graphically represented by rendition 6.1. However, once the listener moves away from a sweet spot, system performance degrades rapidly (headphones alleviate these problems as the user's location does not change relative to a sweet spot).

If a system intends to keep the virtual sound source at the same location, when the head moves independent of the sound sources, the binaural synthesizer should update its HRTF matrix to reflect the movement. Figure 6.3 shows the components of an example binaural system. B is the binaural engine representing the HRTF matrix, C is an acoustic crosstalk canceller that can be used to account for room reverberations, which is what needs to be updated in realtime. The updates of B and H were referred as dynamic binaural synthesis and dynamic crosstalk canceler, respectively and are presented in previous

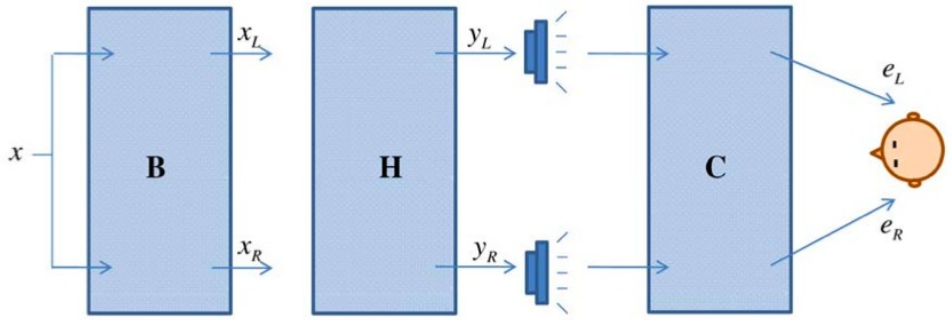

Figure 6.3: Schematic of a binaural audio system

work with the application that allowed the user to move independently of speakers [22].

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 transformations.

In this paper, we argue that the imlpementation of 3D audio algorithms to create a personal 3D audio interface which draws sound images around the user, as depicted the in virtual rendition found in figure 6.1, can be used to create a more efficient interface when vision is not available. By creating a virtual environment that allows sound images to be placed arbitrarily around the user, content that is depicted audibly can be maniupulated to drive a number of key metrics. In this section, we enumerate the direction of future work and the proposed aspects of exploration.

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. Previous 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, 25]. In this regard, any exploration of a binaural interface should empirically test the value of choosing different types of notifications and the presentation of the notifications to the user.

6.0.3 Current Results and Frameworks

Currently, a prototype of this system has been developed. The existing work has explored four physical interfaces : 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 in real time. It is the platform of choice as it is available across all devices and currently only relies on the availability of an HTML5 compliant browser and a device that can produce stereo sound.

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 users to interact with a given interface through multi-tasking techniques afforded by independent sound objects.

7.1 Empirical Measure of Efficiency

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.

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 [39]. Expanding on this research in the auditory domain with empirical studies on user satisfaction using these measures is the first step in understanding the potential for 3D interfaces.

7.2 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 if acute attenuations are mis-calibrated 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 [17].

7.3 Crowd Sourced Domain

Crowd sourced 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 orchestrated guidance or remote control of a user to a location (such as navigating a blind user to a bathroom in a mall). Another application could be providing more natural instruction to web workers through the use of sound images that guide workers to complete

tasks.

7.4 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 first step for this project.

7.5 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 optimal configuration. Surveys have shown that culture influences on communication preferences for patients seeking treatment at the end of life stage present different requirements in both the content and structure of the information relating to care [34]. This cultural sensitivity to content can be explored to tailor this interface to users with diverse backgrounds.

Conclusions

The lack of adoption of 3D audio as a primary interface is often attributed to a number of factors in the literature. Prior research in psychology has shown that humans base much of their communication on gestures, nuance, and inflection [36]. As a result of modern speech synthesizers inability to communicate using these components, 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) [36].

When using sound 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. 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.

We have explored these psychological effects and explored the background research needed to create four different patterns of interfaces. 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 [25].

We have presented an exploration of prior research and posed necessary

research questions that should be explored when designing an interface that uses 3D audio to place sound around a user on any device. Such an interface could leverage techniques in binaural audio, artificial intelligence, and human computer interaction to provide users with an immersive environment to interact with their technology. This interaction would be useful, both as a tool for enabling blind users, augmenting the capabilities of non-handicapped individuals, and 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] Barry Arons. Hyperspeech: Navigating in speech-only hypermedia. In *Proceedings of the third annual ACM conference on Hypertext*, pages 133–146. ACM, 1991.

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

[4] 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.

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

[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] Myounghoon Jeon, Benjamin K Davison, Michael A Nees, Jeff Wilson, and Bruce N Walker. Enhanced auditory menu cues improve dual task performance and are preferred with in-vehicle technologies. In *Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 91–98. ACM, 2009.

[16] Myounghoon Jeon, Nazneen Nazneen, Ozum Akanser, Abner Ayala-Acevedo, and Bruce Walker. Listen2droom: helping blind individuals understand room layouts. In *Proceedings of the 2012 ACM annual conference extended abstracts on Human Factors in Computing Systems Extended Abstracts*, pages 1577–1582. ACM, 2012.

[17] 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.

[18] 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.

[19] 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.

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

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

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

[23] 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.

[24] 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.

[25] 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¿.

[26] 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.

[27] 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.

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

[29] 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.

[30] 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.

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

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

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

[34] 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.

[35] 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.

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

[37] Bruce N Walker and Anya Kogan. Spearcon performance and preference for auditory menus on a mobile phone. In *Universal Access in Human-Computer Interaction. Intelligent and Ubiquitous Interaction Environments*, pages 445–454. Springer, 2009.

[38] 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.

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

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

[41] Wai Yu, Ravi Kuber, Emma Murphy, Philip Strain, and Graham McAllister. A novel multimodal interface for improving visually impaired peoples web accessibility. *Virtual Reality*, 9(2-3):133–148, 2006.