AuralEyes: Investigating the Fusion of Eye-Tracking and Spatial Audio in Electronic Sensory Aids for the Blind

by

Frank Jones

B.S., University of Idaho, 2007

M.S., University of Idaho, 2010

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Computer Science

2013

This thesis entitled:
AuralEyes: Investigating the Fusion of Eye-Tracking and Spatial Audio in Electronic Sensory Aids for the Blind
written by Frank Jones
has been approved for the Department of Computer Science

_____
Dirk Grunwald

_____
Qin Lv

_____
Nicolaus Correll

_____
Lewis Harvey

_____
Michael Lightner

Date _____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Jones, Frank (Ph.D., Computer Science)

AuralEyes: Investigating the Fusion of Eye-Tracking and Spatial Audio in Electronic Sensory Aids for the Blind

Thesis directed by Prof. Dirk Grunwald

At least as early as 1945 researchers have sought to utilize electronic devices to communicate spatial and environmental information to the blind [9]. Despite significant research and development efforts since then, the number of electronic sensory aids (ESAs) actively utilized by the blind community remains small.

A major challenge to the adoption of ESAs is the steep and protracted learning curve generally associated with such devices[20]. Impractical and/or non-intuitive man-machine interfaces contribute to this problem. Existing ESAs lack a natural and hands-free method of providing direct user manipulation of the input stream, such as is available to sighted persons by the moving of their eyes. In the case of audio devices, a secondary effect of this shortcoming can be the implementation of obscure audio codes in an attempt to disambiguate positional elements of the data-stream. Such deficiencies limit the usability of an ESA.

In this work I propose the fusion of eye-tracking with spatialized audio feedback as a means of increasing ESA usability - by enabling direct user control over synthetic sensory feedback. To this end, I submit AuralEyes, a novel man-machine interface designed for use in ESAs for the visually impaired. Experimental results show that users of AuralEyes are able to perform simple range disparity tasks on simulated input with only a few minutes of training. There is evidence that user preference favors an AuralEyes implementation that employs spatialized audio feedback over a similar implementation with non-spatialized feedback. Finally, I present a fully functional implementation of the AuralEyes framework.

Dedication

To my wife Erin. More than anything else, your sacrifice, patience and support has made the completion of this effort possible. And to my children: Hannah, William, Oliver and Clara - four of my best reasons for doing anything good. A bit of all of you is in here.

Acknowledgements

I would like to thank all of my committee members - most especially my major professor Dr. Dirk Grunwald - for your correction, your ideas, your flexibility, and your encouragement. I know your time is precious.

I am also grateful to Marlin De May, who has assisted with several aspects of the investigations in this work, and who continues to support and participate in this research.

I also thank Daniel Houck of Harvey Mudd, who dedicated a summer to assist me in understanding and developing the algorithms and approaches to spatial audio employed in this work.

Finally, I would like to express my appreciation to Greg Johnson and Dharmjeet Rattan at The College of Idaho; exceptional undergraduates whose enthusiasm to further this research has been a breath of fresh air.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Introduction

An estimated 284 million people worldwide have vision impairment, 39 million of these are totally blind[65].

For the visually impaired (VI), computing technology can facilitate access to information that would be otherwise difficult or impossible to obtain. For decades researchers have studied technologies and techniques for enhancing the range and depth of information accessible to the VI, through alternative senses such as touch and hearing. Refreshable braille displays and screen readers represent well known and relatively early examples of such efforts. Extending this concept into the mobile domain presents numerous additional challenges and opportunities for both the designer and user of computing technology.

Electronic sensory aids (ESAs) utilize computing technology to transform information typically perceived via one sensory modality (such as sight) into another (such as hearing) for utilization by the user. Virtual audio displays (VADs) which utilize spatialized digital audio have proven particularly effective in this domain. Moving beyond text bound information, modern ESAs seek to transmute contextual and environmental information into a consumable format for the user. Increasingly affordable and powerful mobile computing devices have rendered these technologies highly portable, even wearable, making possible their incorporation into everyday living[9, 58, 81, 95] .

Despite a significant body of research and development efforts, few electronic sensory aids have escaped the realms of academia. None have succeeded in displacing the white cane as the defacto sensory aid of choice within the blind community. Numerous factors have likely contributed to

the low adoption rates of ESAs including cost, maintenance, usability, reliability and accessibility. The principle focus of this work is on advancing the state of the art in audio-based mobile sensory aids for the blind. Specifically I focus on the usability and accessibility of such systems.

A major challenge to the adoption of ESAs is the steep and protracted learning curve generally associated with such devices[20]. Impractical and/or non-intuitive man-machine interfaces contribute to this problem. Historically much emphasis has been placed on the acquisition, processing and transmittal of environmental information with insufficient consideration given to user interaction. Consequently the operational model employed by numerous ESA's borrows little from the sense that they are seeking to augment or replace, namely sight. The level of "user control" that sighted person's exert over their optical sense is significant; intentionally and independently selecting where and when to acquire information via manipulation of their eyes. This ability not only reduces computational complexity, when compared to a 360 degree field of view, but also augments the incoming sensory data with orientation information. Existing ESAs lack a natural and hands-free method of providing such directed user manipulation of the input stream. In the case of audio devices, a secondary effect of this shortcoming can be the implementation of obscure audio codes in an attempt to disambiguate positional elements of the data-stream. Such deficiencies limit the usability of otherwise promising technologies.

When the user interface of an ESA is sufficiently non-intuitive that it requires dozens of hours of orientation before use, the accessibility of such a device is impacted. Setting aside the necessary time and resources to afford and/or attend needed training can become not only inconvenient - but impossible. Thus the usability of a newly designed ESA must be considered, alongside purchasing and maintenance costs, as an accessibility concern.

I argue that inadequate user control over synthesized sensory input has a negative impact on the usability and accessibility of existing ESAs.

I propose the fusion of eye-tracking with spatial audio feedback as a means of providing an intuitive and powerful method of user control in assistive devices for the blind. Specifically I offer the following theses:

(1) A user interface that fuses eye/gaze tracking with audio feedback can enable a user to extract meaningful spatial information about their environment through directed exploration.

(2) Spatialized, gaze-directed audio feedback provides a reinforced sense of orientation, resulting in a more intuitive user experience relative to gaze-directed audio feedback alone.

The outline of this dissertation is as follows. Chapter 2 provides technical background information, beginning with an introduction to some of the technologies employed in ESAs for the blind. The chapter culminates with an introduction to spatial audio, its foundational principles and the importance and challenges of HRTFs and HRTF individualization. Chapter 3 presents foundational and current research efforts related to my work, focussing especially on assistive technologies for the blind that employ advanced audio displays. In chapter 4 I introduce AuralEyes, presenting the model and assumptions around which the interface is built. Chapter 5 outlines the materials, methods and procedures of a comparative "zero-day" usability study designed to investigate the effectiveness of AuralEyes. Chapter 6 is dedicated to experimental results and discussion. Finally, in chapter 7 I present my conclusions and plans for future work.

Chapter 2

Background

In this chapter I present some of the technological and historical foundations of modern mobile electronic sensory aids. The basic architectural elements and enabling technologies of ESA's for the blind are discussed, followed by an introduction to spatial audio. The chapter concludes with a brief discussion of open problems related to spatial audio and their implications to ESA design.

2.1 Architectural Considerations of A Mobile Sensory Aid

Mobile Electronic sensory aids must address at least four challenges:

(1) The acquisition of environmental data.

(2) Processing and transformation of environmental data.

(3) Presentation of relevant data.

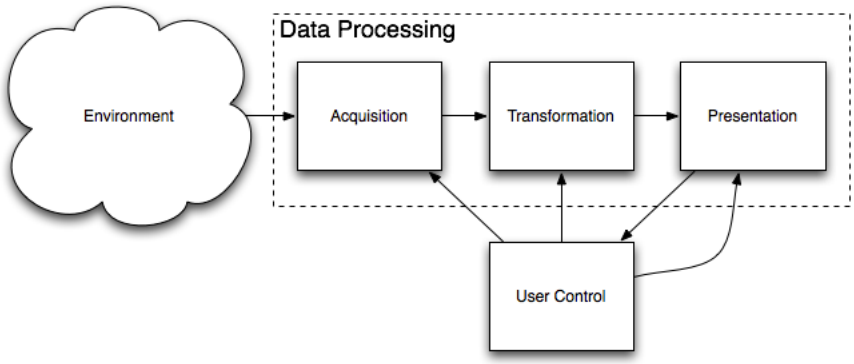
(4) User control of the device.

Figure 2.1 shows the architecture of an ESA partitioned into components that address each of these tasks.

2.1.1 Acquisition / Sensors

An obvious first step in delivering useful information about a person's environment is the acquisition of such information. Recent decades have provided the designer of modern ESAs with

Figure 2.1: Architecture of A Mobile Sensory Aid



an impressive array of powerful and inexpensive sensors, cameras, radios etc. capable of collecting information about a user's surroundings, location, and orientation. The following subsections provide a brief survey of the characteristics, benefits and drawbacks of some of these technologies. This section is not intended to be exhaustive but to provide the reader with a sense of the capabilities and trade-offs that ESA designers must negotiate.

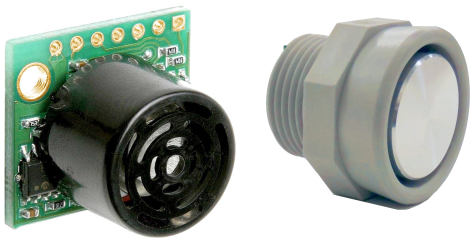
Some of the most useful information that a person can have about their environment is the proximity of objects in the immediate vicinity. Fortunately there are numerous and inexpensive technologies available for gathering such information.

2.1.1.1 Range-finding

Ultrasonic Sensors

Ultrasonic sensors utilize the relatively consistent speed of sound through air (aprox. 1,130 ft/sec

Figure 2.2: Ultrasonic Sensors [41]



[26])to determine the distance to one or more objects in the immediate environment. Such sensors operate by emitting sounds at frequencies well above the range of human hearing[1] and then measuring the amount of time that elapses before an echo is sensed by an ultrasonic detector[2] . Multiplying this time by the speed of sound through air and dividing by two produces the range to the object that reflected the sound waves.

$$Range = \frac{((1,130 ft/sec) * ElapsedTime)}{2} \tag{2.1}$$

Utilizing ultrasonic frequencies has multiple benefits. Because the frequencies are outside the range of human hearing the device does not create an audible disturbance. Additionally, the shorter wavelength of the sound waves makes them less susceptible to the effects of diffraction[3] resulting in measurable reflections from much smaller objects than can be detected with audible sound waves. Commercial ultrasonic sensors operating at 42KHz are capable of detecting reflections from nearby objects on the order of 1 cm in diameter[41].

ESA designers may purchase pre-assembled commercial range-finding sensors similar to those shown in figure 2.2 on the left for around thirty to forty dollars per unit. Weather resistant sensors like the one on the right are typically two to three times more expensive. Common communication interfaces for such sensors include pulse width modulation, serial bus, analog voltage differential, and etc. Typically such sensors are activated by initiating a simple control signal and a range measurement is made available on the communication bus after a predetermined amount of time.

If tighter integration or more direct measurement control is desired, the discrete components utilized in ultrasonic sensors can be purchased individually for tightly integrated custom designs. Though versatile and inexpensive, ultrasonic rangefinders do have certain limitations. Sound wave attenuation in air generally increases with frequency[26], meaning that the useful range of ultrasonic echo detection is much less than for audible echoes. Reliable sensing ranges begin at 3 - 4cm (minimum reportable range) and extend to distances on the order of 10 - 20ft maximum.

[1] Humans can hear up to around 20KHz, ultrasonic sensors typically operate upwards of 40KHz.

[2] Because emitting and sensing capabilities can be built into the same device these sensors can be made extremely compact.

[3] Diffraction allows waves to "bend" around objects whose diameter is small relative to their own wavelength.

Ultrasonic sensors suffer from limited radial resolution due to the diffusive nature of sound. Consequently, though the range to an object can be accurately determined to centimeter resolution - size, shape and bearing are ambiguous. Utilizing multiple sensors with varying orientations can help but interference between sensors causes this approach to suffer from diminishing returns.

Infrared Sensors

Infrared (IR) range-finding sensors offer an interesting alternative/compliment to the capabilities

Figure 2.3: Infrared Sensors [24]

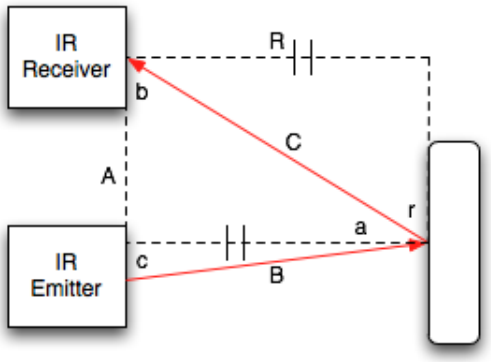
of their ultrasonic cousins. As the name suggests these devices utilize a beam of infrared light to perform their function. Rather than propagation delay - which would be extremely small for the range within which these devices can be effectively used - IR sensors rely on triangulation.

Range-finding is accomplished by projecting a narrow beam of IR light forward at a predetermined angle and detecting the reflection of that beam using an IR sensor.

By measuring the angle of the reflection at the receiver, the law of sines can be used to calculate the distance from the sensor to the object (see figure 2.4 and formula 2.2).

$$
\begin{aligned}
&\frac{90^\circ}{\sin(C)} = \frac{r}{\sin(R)} \rightarrow R = \sin^{-1}(\frac{r\sin(C)}{90^\circ}) \\
&\frac{a}{\sin(A)} = \frac{c}{\sin(C)} \rightarrow \sin(C) = \frac{c\sin(A)}{a} \\
&a = 180^\circ - (b + c) \\
&r = b \\
&\therefore R = \sin^{-1}(\tfrac{bc\sin(A)}{90^\circ a})
\end{aligned}
\tag{2.2}
$$

Figure 2.4: IR Rangefinding



The narrowness of the IR beam is advantageous because it offers a high degree of spatial resolution - but disadvantageous because the sensor will only be able to detect an object if it happens to be within the relatively small path of the beam. A partial solution to this problem is to mount such sensors on pivoting servos and continuously "sweep" side to side. This approach increases the angular coverage of the sensor on a horizontal plane in space but at the expense of temporal resolution since the range information of a particular location in front of the sensor can only be updated each time the sensor sweeps past it. From a wearable computing standpoint the practicality of a solution involving constantly oscillating sensors is highly suspect; however one might conceive of an approach that takes advantage of the naturally occurring movements of the body and head as an alternative.

The sensing range of IR range-finders is affected by design decisions (the angle the IR beam is emitted at and the field of view of the detecting sensor) as well as operating conditions (i.e. indoors vs. outdoors where ample IR light increases the level of background noise for the IR sensor). Typical sensing ranges fall between a few inches up to several feet[24].

Commercial IR sensors can be purchased for as little as around ten dollars per unit [24].

2.1.1.2 Digital Cameras

Advances in digital imaging sensors have ushered in an age of inexpensive digital cameras - including the seemingly ubiquitous "web-cams" that are increasingly included in mobile computing devices. Imagery from a single digital camera may contain numerous types of information that are useful to the user of an ESA. Examples include: text that can be processed by OCR software, features and objects detectable by image recognition algorithms, color, boundaries/transitions etc.

Visible Light

Generally speaking digital cameras come in two types, visible and IR. Not surprisingly the majority of digital cameras that we **intentionally** purchase operate exclusively in the visible spectrum. The majority of these devices utilize CCD or CMOS imaging sensors that range in resolution from thousands to tens of millions of pixels[83]. Visible light cameras are an effective method of gathering data that is generally accessible to sighted individuals including virtually any kind of "human-generated" information such as text, signs, or other man-made imagery.

Infrared

Infrared cameras are useful for at least two reasons. First, there are situations in which specific information is more easily extracted from an infrared image than one generated with visible light. For instance the features of the human eye that are useful for eye-tracking are more easily detected in infrared than in visible light as these features become more pronounced as less relevant "noise" is reduced in the image (see image 2.5).

Figure 2.5: Visible and Infrared Images of a Human Eye

Secondly, IR cameras can be used to detect system generated information and features that are intentionally hidden from the casual sighted observer. Consider the IR "dot" projected by the IR rangefinders mentioned in section 2.1.1.1 - such capabilities can enable the extraction of additional information from a user's surroundings without cluttering the visual environment for others in the area. This idea is discussed in greater detail shortly.

An interesting irony is that the technology used to generate the sensors for cameras that operate in the visible spectrum results in image sensors that are sensitive to a range of the electromagnetic spectrum that includes near-visible infrared light. By inserting an IR filter between the optical lens and sensor array a camera that is only sensitive to visible light is produced. With care one can remove this filter, replace it with a visible light filter (such as overexposed color negative) and effectively turn a visible light camera into a near visible IR camera. This technique can been used to construct an eye-tracking IR cameras 'on the cheap'.

Stereo Vision

As useful as a single camera can be - two or more provide the added benefits of parallax. Parallax is the difference between objects in images taken of the same scene but from different vantage points in space and/or time. By analyzing these differences information about the range, relative position, shape, and dimensions of objects can be extracted from two or more images. This is one of the cues that the human brain uses to construct a three-dimensional perception of our surroundings [16].

Computer vision is an active research field in its own right and an in-depth discussion of the capabilities and techniques employed in stereo vision and object detection is well beyond the scope of this work. I provide here only one example of a commercially available stereo vision product as a sample of the resources available to ESA designers.

The Bumblebee XB3 (shown in figure 2.6) is a commercially available stereo vision product. This system utilizes three cameras and is capable of capturing and processing images at resolutions ranging from 648X488 pixels at 48 frames per second (FPS), up to 1280X960 at 16 FPS. The sensor can operate in black and white or color modes and is interfaced via an IEEE-1394b FireWire bus. The system comes with software development kits (SDKs) for image acquisition and camera control

Figure 2.6: Stereo Vision [35]

as well as image rectification and stereo processing.

The benefits of such systems is that they provide powerful off the shelf access to effective object detection and depth mapping capabilities. Some of the drawbacks include a hefty price tag (tens of thousands of dollars) and significant computational requirements.

Figure 2.7: A Structured Light Based Infrared Depth Sensor: The Kinect Sensor [62]

Structured Light

An inexpensive alternative to industrial grade stereo vision systems are the structured light approachs exemplified by the Microsoft Kinect and ASUS Xtion Pro Live sensors[40, 62, 93]. The technology utilized in these devices [4] employs an IR projector that "paints" an invisible complex

[4] Developed and licensed by Primesense

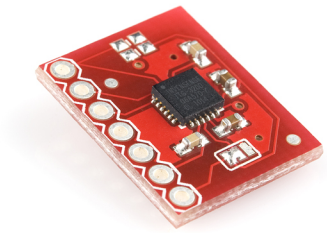
pattern on the surface of objects. An IR camera captures an image of the resulting scene. A depth map for the scene is generated by analyzing the projected pattern(see figure 2.8). A third camera operating in the visible spectrum can be used to tie visual information to the depth map generated by the sensor[93]. Microsoft has released a Windows 7 SDK for the Kinect[61] and open-source alternatives such as the freenect library being developed by the OpenKinect project are currently under developement[64]. The OpenNI project provides open source libraries and binaries for use with both the Kinect and Xtion Pro Live. Such resources are allowing hobbyists and researchers to utilize the sensing capabilities of the Kinect for more than simple gaming interaction[40, 62, 93].

Figure 2.8: Infrared Pattern Projected by the Kinect Sensor [93]

2.1.1.3 Orientation Sensing: Electronic Accelerometer, Gyroscope and Compass

Using physical phenomena such as inductance, capacitance, voltage etc. it is possible to construct small, lightweight sensors capable of measuring acceleration, orientation and magnetic fields. Commercially available electronic accelerometers, gyroscopes and compasses offer ESA designers lightweight and inexpensive access to positioning and orientation information that may be useful in the acquisition and processing of information in a user's immediate environment. Such sensors often employ simple analog voltage outputs or simple serial bus interfaces such as I^2C and may be purchased for thirty to fifty dollars - though more sophisticated and expensive options are also

Figure 2.9: A Digital Accelerometer From Sparkfun [25]



available.

2.1.1.4 Radios

Recent decades have seen an explosion in the use of portable radio technology. Radio transmitters and receivers offer multiple enhanced capabilities to ESA designers. **Cellular Networks** The ability to remotely access computational resources and databases via the internet offers access to greater computational resources than can be physically included in a mobile device - making possible more advanced data processing in a mobile solution. Additionally, the presence of multiple cellular phone towers can be used to triangulate the position of cellular radio. Currently available commercial implementations of this technique claim in/outdoor accuracy from 1- 30 meters [33]. The financial obligations associate with cellular networks are an obvious limitation to this technology.

Global Positioning System

The well known Global Positioning System (GPS) utilizes radio transmissions from satellites in geosynchronous orbit to determine geographic location. Multiple researchers have suggested/utilized the use of GPS in ESAs for the blind[52, 48]. Financially use of the GPS system does not incur the long term costs associated with cellular technologies, an obvious advantage. However, the inability to utilize the GPS network reliably indoors or underground, due to a relatively weak transmission signal, may nullify this benefit in some instances.

Local Area / Ad - Hoc Network

Similar to cellular networks, local area or WiFi networks can allow access to remote stationary computing capabilities, as well as provide positioning capabilities - assuming the presence of multiple static transmitters within the network. The limited range of these networks poses an obvious problem for this technology as a standalone solution in ESA design.

2.1.2 Data Transformation

Once the desired environmental/situational data have been collected, an ESA must transform this raw data into useful information for the user. Multiple levels of technological sophistication are available depending upon the nature of the data and the presentation method to be employed.

2.1.2.1 Embedded Circuitry

In simple cases, where the amount of input is limited and conversion between sensory readings and the signal to be presented to the user is trivial, minimal embedded circuitry is sufficient. Examples of commercial devices that utilize this approach include the Ultracane and K-Sonar which transmute ultrasonic information into haptic and audio feedback respectively. The straightforward transformation from input to output signal leverages the ability of the user's mind to adapt to and understand the feedback from the device, removing the need for significant computational resources[81, 79].

2.1.2.2 Computer Based

When large amounts or more complex data are utilized (i.e. vision based systems), or sophisticated methods of presenting the information to the user are employed (i.e. spatial audio, synthetic voice etc.) more substantial computing capabilities are required.

Portable/mobile computing devices such as smartphones/PDAs and laptops/netbooks can provide impressive computational capabilities while allowing an ESA to remain portable and self contained[17]. So called "web-enabled" devices, such as smartphones or properly equipped laptops,

can access additional computational resources via the internet, allowing for cloud-computing and "crowd-sourcing" based capabilities [48, 95].

2.1.3 Presentation

With the necessary data acquired and processed, an ESA must present the relevant information to the user through one or more available senses. Both the mode of transmission and format of the data must be considered.

2.1.3.1 Audio

Audio is an oft-leveraged avenue for conveying information to the blind. Next to sight, hearing offers the highest data throughput of the human senses. The human auditory system is highly sensitive to pitch, capable of detecting as little as a 0.3 percent change in frequency[16]. The perception of other factors such as volume, timing, and sound source orientation, offer multiple dimensions through which information may be encoded.

At a high-level there are multiple formats through which audio-based information may be presented.

Verbal

Depending on the function of the device, verbal feedback may be an effective method of presentation. The personal guidance system proposed by Loomis et al. experimented with verbal feedback to provide navigational guidance to the user. Spoken directions such as "left eighty" directs users toward a desired destination[52]. One of the advantages of verbal feedback is that it can be immediately recognized and understood by a user - no significant training is required. A drawback to this approach is the amount of computation and time necessary to generate, present, and consume verbal cues.

Nonverbal

As an alternative to speech ESAs may employ a mapping between spatial/contextual information

and nonverbal signals. This approach has the advantage of being more general[5] , requiring less computational overhead, and requiring less time to present (i.e. the duration of a special sound may be much shorter than a typical spoken word). As mentioned previously, frequency, volume, and timing offer avenues for delivering low-level non-verbal cues to a user. Such attributes of sound may be used to convey low level information such as brightness, color, range, and relative position/direction[12, 17, 28, 58].

Auditory Icons/Earcons Gaver and Blattner et al. have proposed auditory icons and earcons respectively as nonverbal techniques of presenting high-level information through sound[11, 30, 31]. Gaver's work proposed simple sounds that are related to the task or information they represent such as a crackling sound to indicate a test file[30]. Blattner's approach built upon this concept constructing "families of sounds" from "compound audio elements". In their work with the System for Wearable Audio Navigation (SWAN) developed at Georgia Tech, Walker and Lindsay make a strong case for the advantages of nonverbal audio signals including auditory icons and earcons[88]. They point out that such signals take less time to transmit (and thus create less audio clutter) and are more quickly comprehended than their lengthy verbal counterparts[88]. Additional advantages include the fact that such sounds are not bound to a specific language and can therefore improve the accessibility of a system without the need for multilingual capabilities. Walker and Lindsay successfully demonstrate the use of multiple nonverbal "beacons" to perform navigation with the SWAN system in [88].

Virtual Audio Displays

Both verbal and nonverbal feedback may be spatialized (see section 2.2) to add positional information and relevance. The use of spatial audio to communicate an immersive and convincing sense of space in an audio display constitutes what is known as a virtual audio display (VAD). Numerous studies have demonstrated the effectiveness of incorporating VADs in ESAs for the blind [6, 52, 48]

.

[5] Low level information is often not effectively transmuted to speech.

2.1.3.2 Haptic

Haptic feedback is another method for communicating information to blind users of ESAs. Haptic interfaces have the advantage of not cluttering the existing audio environment but at the expense of reduced throughput. Vibrotactile displays have been experimented with for the back, abdomen, fingers and forehead [5]. The "Ultracane" is a modern ESA that communicates range information from two ultrasonic sensors to the user via two vibrotactile "buttons" located on the handle of the device[79].

Figure 2.10: The Tongue Display Unit [47]

Electrotactile stimulation is an interesting alternative to vibrotactile feedback. The technique utilizes electrical impulses instead of vibration to stimulate a region of the body. This technique is utilized in the tongue display unit (TDU) developed at the University of Wisconsin Madison[4]. The TDU operates by transmitting information to the user in the form of mild electrical impulses delivered to the user's tongue via a grid of up to 144 electrodes. Various types of information such as basic shapes or imagery from a region on a computer screen may be encoded and transmitted via the device. The TDU has been successfully utilized in numerous studies exploring sensory

substitution and neural plasticity[4, 5, 47].

2.1.4 User Control

Considerable research effort has been brought to bear on the tasks of data acquisition, transformation and presentation in ESAs. Unfortunately the remaining element of an ESA's architecture - the user interface, has receive disproportionately less attention.

The PGS demonstrated by Loomis allows the user to select the operational mode of the device using a traditional keyboard interface[52]. The SWAN system utilizes an audio menu interface that is navigated using a handheld device derived from the traditional desktop mouse. Used in conjunction with speech recognition users can interact with the system and record custom annotations linked to geographic locations (see section 3.2.1.5 for more information). Such interfaces are demonstrative of the kind of interactions that user's are allowed with typical mobile ESAs. In general the user does not have direct control over the synthetic sensory feedback being presented. Aside from physically changing the orientation and location of their body, thereby repositioning the ESA itself, users must allow the ESA to determine which information to present.

A notable exception is demonstrated in the function of less complex ESAs, typically pointing devices, such as the KSonar and Ultracane[81, 79]. Because the user has direct physical control over the orientation of the device, they can perform deliberate actions to control the nature of information being retrieved from their environment. Put simply, they can point it towards regions they want to get information from and point it away from areas they do not care to know about.

I posit that the lack of a natural and effective hands-free method of directing the behavior of sophisticated ESAs has had a negative impact on the overall usability of these devices. One of the theses of this work is that eye-tracking techniques can provide such an interface. This idea is discussed in greater detail in chapters 4 and 5.

2.2 Spatial Audio

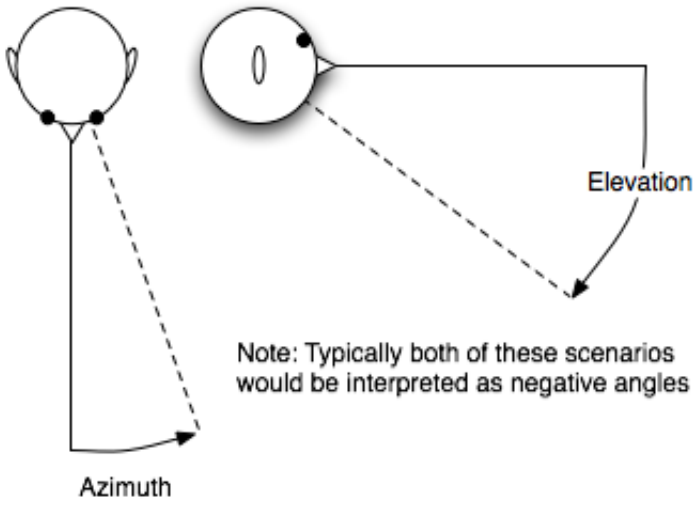
At the cost of an increase in complexity, sophisticated digital audio technology can provide significant increases in the amount of information that can be conveyed to a user through sound. In particular, the spatial nature of sound can be a powerful medium for conveying critical information about the user's environment.

Exploiting the spatial qualities of sound requires at least a cursory understanding of the mechanics involved in human sound source localization. The following section is intended to provide a high level introduction to these concepts. Readers familiar with the theory of spatial audio can safely skip section 2.2.1.

2.2.1 Human Sound Source Localization

The human ability to localize an active sound source in terms of azimuth and elevation (see figure 2.11) has been a topic of interest to researchers for over one hundred years[63]. Studies have shown that under proper conditions listeners are able to perceive as little as one degree of change in a sound source's position[63]. In fact, in terms of determining the azimuth and elevation of an active signal, the human auditory system appears to be superior to that of a bat![91]

Figure 2.11: Azimuth and Elevation

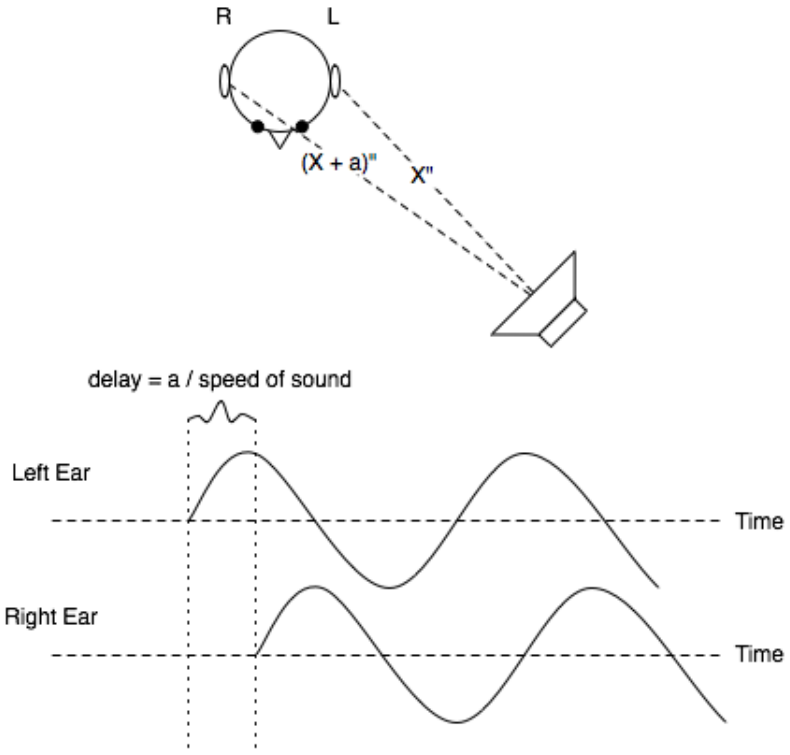


The mechanisms by which humans localize sounds rely upon the positional geometry of a sound source and the physical features of human anatomy. The simple fact that our ears are positioned on opposite sides of the head results in differences in timing and volume between the two ears, providing two valuable cues for localization.

2.2.1.1 Interaural Time Difference

Interaural time difference or ITD results from the fact that in most cases the distance from a sound source to the left and right ears is different. Notable exceptions occur when a sound originates on the vertical plane that separates the left and right hemispheres of the head (an azimuth angle of 0 or 180 degrees), these and other special cases will be addressed shortly.

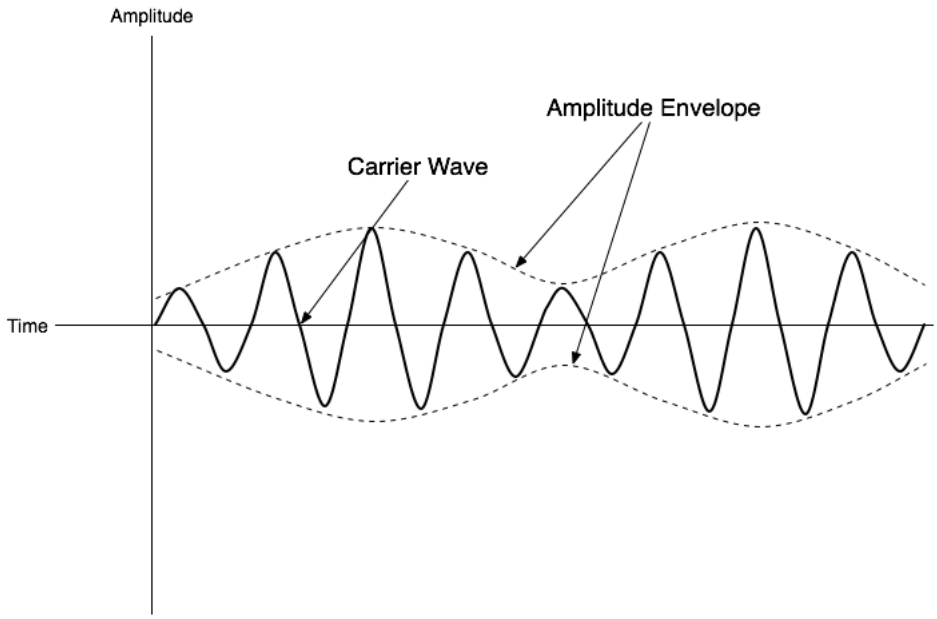
Figure 2.12: Interaural Time Difference



The difference in path length to each ear results in a difference for the onset of a sound as well as a phase shift between the two ears (see figure 2.12). These timing differences can be

detected and utilized by the brain for pure tones below approximately 1.5KHz. Above 1.5KHz the wavelength of sound in air approaches and becomes smaller than the size of the human head and ambiguity makes it impossible to determine which waveform is leading. Amazingly, if amplitude modulation is present in the sound then the brain is able to extract ITD cues from the amplitude envelope, provided that it has a frequency below 1.5KHz, even if the carrier frequency is well above the 1.5KHz threshold (see figure 2.13)[8, 44].

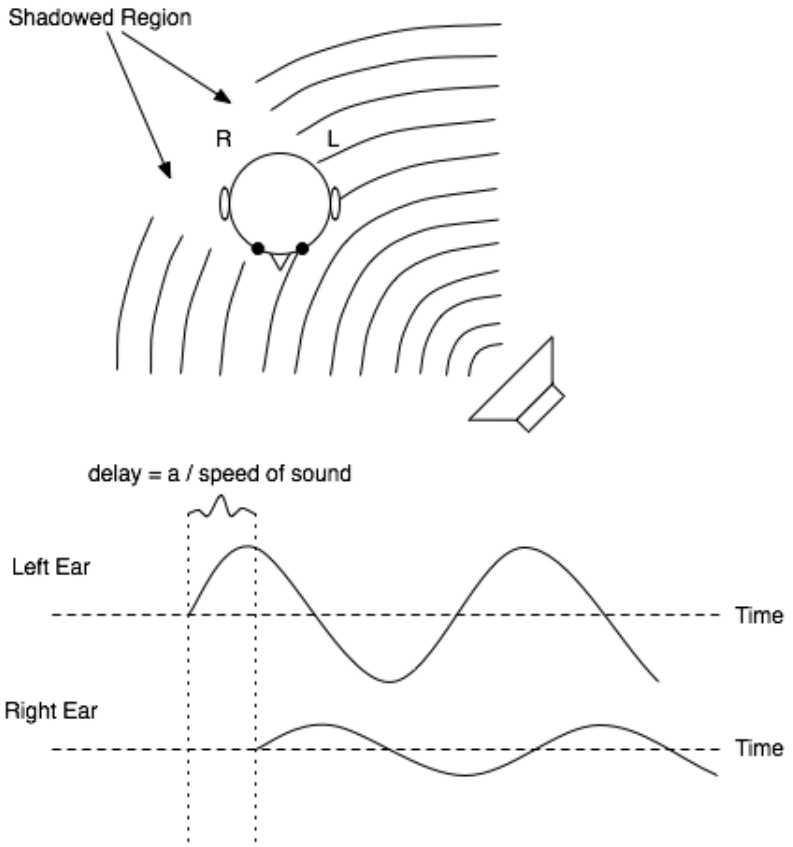
Figure 2.13: Amplitude Envelope



2.2.1.2 Interaural Level Difference

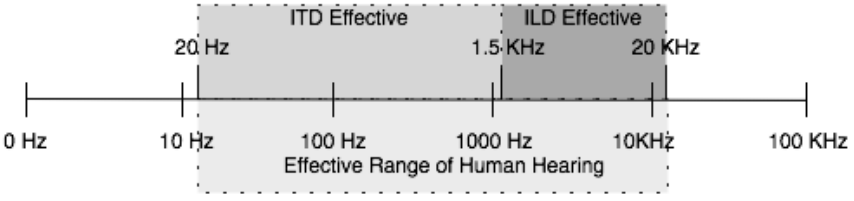
The second localization cue resulting from the apposing positioning of the ears is a difference in volume perceived at the two ears. This is referred to as as interaural level difference (ILD). For frequencies above 1.5KHz the head partially occludes sounds originating from the opposite side of a given ear - thus we hear sounds "louder" in the ear that faces a sound source. Lower frequency sound waves are able to bend around the head (i.e. diffraction) mitigating this effect. Above 1.5KHz the size of the head is large relative to the sound's wavelength allowing ILDs to become detectable (see figure 2.14)[44].

Figure 2.14: Interaural Level Difference



Together ITD and ILD provide localization cues for both "low" and "high" frequency sounds within the range of human hearing (see figure 2.15) . These cues alone, however, are not sufficient to unambiguously localize a sound in space due to a phenomenon known as the "cone of confusion".

Figure 2.15: Effective Ranges of ITD and ILD

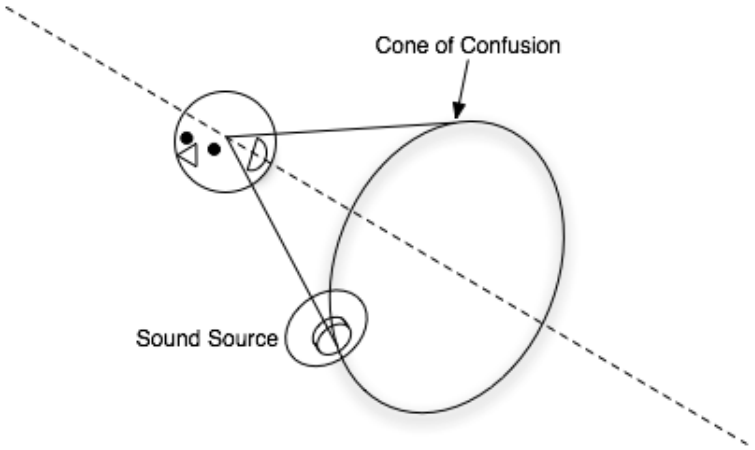


Note: Frequencies are plotted on a log scale

2.2.1.3 The Cone of Confusion

The approximately spherical shape of the human head results in ambiguity when attempting to localize sounds based upon ITD and ILD alone. The ITD and ILD differences resulting from a sound originating at a certain elevation and azimuth are identical anywhere on the surface of a cone generated by rotating the line passing through the source and the center of the listener's head around a horizontal axis passing through the ears of the listener (see figure 2.16)[8, 44]. More information is needed to determine precisely the elevation and front/back positioning of a sound upon the surface of this cone[8, 44].

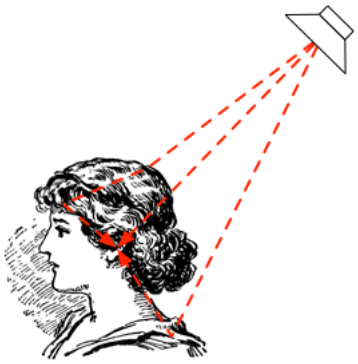
Figure 2.16: The Cone of Confusion



2.2.1.4 Spectral Cues

Additional information about the exact location of a sound source is embedded within the content of the sound itself. Some of the sound waves reaching a listeners inner ear may travel in a straight line from the source to the eardrum without interference. Many, however, will take a less direct route, passing through portions of the listener's head and outer ear (or pinnae), and/or reflect off of one or more surfaces (such as the shoulders or the inside of the skull) on their way to the inner ear.

Figure 2.17: The Many Paths to the Eardrum



As sound waves pass through or bounce off of objects they are modified by the characteristics of the materials they encounter. Various frequencies are more readily absorbed, reflected or even amplified, resulting in alterations to the overall spectrum of the original sound. The effects of these alterations is known as filtering. Thus, in a very real sense, the anatomical features of a person's head, pinnae, shoulders etc. act as individualized filters for incoming sound. Because these features are not front/back or top/bottom symmetric the filtering effects vary for sounds emanating from different elevations and azimuths. These additional cues (known as spectral cues) provide the necessary distinctions whereby the cone of confusion can be disambiguated.

ITD, ILD, and spectral cues are understood to be the primary physical mechanisms by which humans perform sound source localization in three dimensional space[6] . Using these cues humans are capable of impressive degrees of accuracy when determining the azimuth and elevation of an incoming sound[8, 63, 91].

2.2.1.5 Range to Source

It should be observed that lacking from ITD, ILD and spectral cues are strong, consistent indicators of range. Intensity (loudness) serves as a cue for range, but is dependent upon familiarity with a given sound (i.e. the typical loudness of a human voice), and suffers from the fact that intensity at a given range can vary with environmental conditions (i.e. a reverberant environment

[6] Other psychological clues such as familiarity with a sound may also contribute to localization [8]

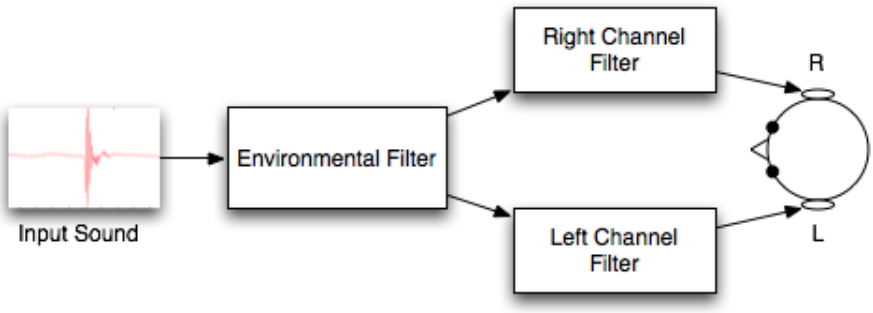
such as an enclosed room vs an open field)[8].

Reverberation provides another indicator of range, but at a cost. The ratio of reflected (reverberant) to direct sound (the R/D ration) provides cues about the environment and the distance to a sound source. Studies have shown that the presence of reverberation improves listener estimation of range; however, these improvements are accompanied by degraded azimuth and elevation estimation[8].

2.2.2 Spatialized Digital Audio

Spatialized digital audio employs computing technology to simulate the effects of ITD, ILD, spectral cues, and reverberation discussed in the preceding paragraphs[8, 14, 15, 18, 26]. Such systems typically implement a series of digital filters that simulate the filtering effects of the target environment and the listener's body (see figure 2.18). Generally headphones or a pair of loudspeakers transmit two channels of filtered audio which approximate the sound waves that would reach the left and right ears as a result of the configuration of sound-source, listener and environment that is being simulated[8, 44].

Figure 2.18: Audio Spatialization

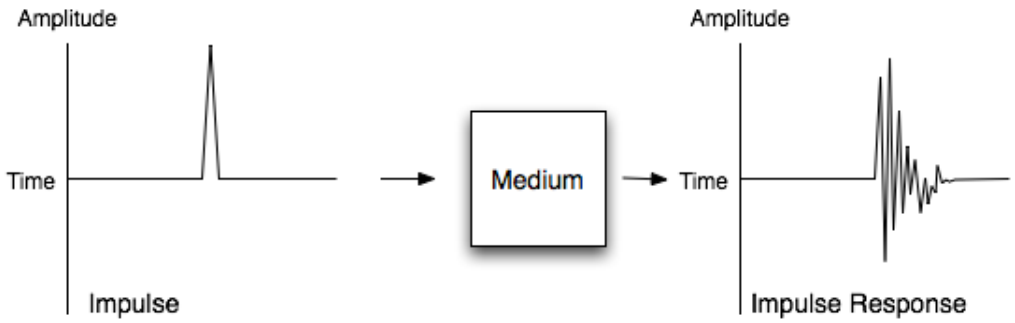


The filters diagramed in figure 2.18 are most often implemented as impulse responses which are convolved with the input waveform using hardware or software techniques[14, 15, 18].

2.2.2.1 Impulse Responses

An impulse response is a waveform that captures the response of a system to a perfect impulse (see figure 2.19). This response, which elucidates a system's reaction to all possible input frequencies, contains the information necessary to filter a sound such that it will "sound" as though it had occurred within the context that the impulse response was recorded.

Figure 2.19: An Impulse Response



Measuring An Impulse Response

Numerous methods of measuring impulse responses have been proposed. A commonly used method is to generate a special test signal containing all frequencies in the audible spectrum. Recordings of white/pink noise or a swept sine wave from 20Hz - 20KHz are common examples of such signals. Generation of an impulse response is accomplished by recording the signal with the loud speaker and microphone(s) positioned in the locations of sound source and listener respectively. Deconvolving the original signal from the recorded test signal results in the impulse response of the system. The resulting finite impulse response (FIR) will include room reflections and the effects of the microphone, speaker etc. If desired the effects of the microphone and speaker can be removed by deconvolving with a recording of the test signal taken using the same microphone and speaker in an echo-free environment instead of the pure test signal. In this way the filtering effects of these devices are removed by the deconvolution.

Impulse responses may be used to capture the filtering effects of a variety of contexts. Examples include the reverberant characteristics of a singer on stage as heard from a certain position

in a lecture hall or stadium, or the filtering effects of a listener's upper body, head, and outer ear (known as "head related impulse responses" (HRIRs). The concept of HRIRs, their significance, and methods of obtaining or approximating them are discussed in greater detail below. An in depth discussion concerning impulse responses in general and the numerous means by which they may be collected is beyond the scope of this work. For more information on this topic the reader is directed to [8] and [26].

2.2.2.2 Head Related Impulse Responses (HRIRs)

By placing probe microphones at the opening of a person's (or dummy head's) ears in an echo free environment, and playing specially designed sounds (such as those discussed in section 2.2.2.1) a pair of FIRs can be recorded for a particular azimuth, elevation and range with respect to the listener. Any sound filtered with these FIRs and played through headphones will sound to the listener as though it originated from the direction of the loud speaker at the moment the impulse was recorded[7] [1, 8].

Beyond one meter the spectral changes resulting from the listener's body are approximately constant as a function of distance. Therefore, by taking such measurements in a spherical pattern around a listener (at a range greater than one meter) a set of impulse responses can be recorded which can be used to virtualize a sound coming from any direction in three-dimensional space around the user[8] [8, 44].

The collection of all measured impulse responses for a given person are referred to as their "head related impulse responses" (HRIRs). When measured at both ears simultaneously (binaural HRIRs) HRIRs capture ITD, ILD, and the spectral filtering effects of the listener's body. For the remainder of this paper references to HRIRs imply binaural HRIRs.

By utilizing an individual's HRIRs it is possible to simulate truly three dimensional sound, giving the listener the impression that sound sources are actually located in space around him/her.

[7] Unless care is taken (i.e. an echo free environment) the FIR will contain the effects of the listener's immediate environment as well

[8] ...with a range greater than one meter of course.

Audio systems that utilize this capability to immerse a user in a virtual three dimensional audio environment are often called "virtual audio displays" (VADs).

2.2.2.3 Head Related Transfer Functions

Due to the computational advantages offered, it is common for software based audio spatialization systems to perform convolution in the frequency domain where the process of convolution can be accomplished via simple multiplication of the two signals to be convolved. To accomplish this the desired filters and input waveform(s) are first transformed into the frequency domain via the Fast Fourier Transform (FFT) or one of its derivatives. Convolution is effected by multiplying the real and imaginary coefficients of the transformed waveforms and then computing the inverse FFT of the result. The end product is the convolution of the filter and the waveform.

When represented in the frequency domain HRIRs are commonly referred to as "Head Related Transfer Functions" or HRTFs. Aside from the computational differences pointed out above there is no significant difference between audio systems that utilize HRIRs and HRTFs to perform spatialization. The information represented by HRIRs and HRTFs is equivalent - merely represented in different domains (time and frequency respectively) - and the terms are at times applied interchangeably. As the term HRTF appears to be used more often than HRIR in the literature, I will utilize the former throughout the remainder of this document.

2.2.2.4 Individualized HRTFs

It is important to note that because every person's body is unique no two people have identical HRTFs. This means that for highly accurate spatialization a listener's individualized HRTFs are optimal. This presents a difficult problem since collecting high quality HRTFs is time consuming and requires expensive facilities and equipment[8, 44, 43, 49].

Numerous studies have investigated techniques for generating generic HRTFs through methods such as averaging of measured sets from multiple individuals or utilizing dummy head measurements [8, 44]. In general localization error increases when such non-individualized HRTFs are

utilized; front/back reversals[9] , for example, increase significantly[97].

Zotkin et al. proposed generating individualized HRTFs based on anthropomorphic measurements[99]. Haraszy et al. built upon this approach by incorporating an artificial neural network (ANN) that used anthropomorphic measurements as inputs[37]. In experiments the ANN generated HRTFs with an error margin of 3 to 5 percent relative to measured HRTFs. It remains unclear how this level of numerical inaccuracy would manifest itself in actual perceptual performance[99]. While anthropomorphic techniques result in a significant reduction in time and resource requirements over direct measurement, a nontrivial investment of time and effort on the part of the user as well as the required participation of additional individuals remains (i.e. take pictures, measurements, etc.).

Subjectively Fitted HRTFs

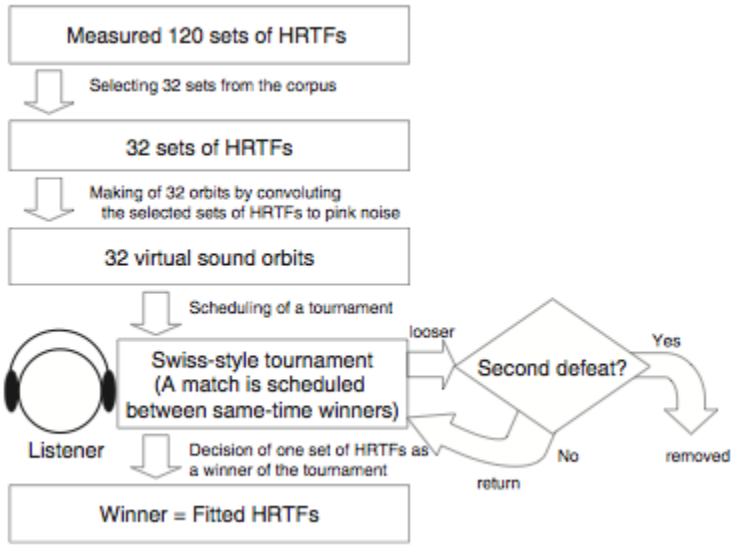
Seeber and Fastl experimented with subjective selection of an optimal HRTF set from a small database of measured HRTFs. Their technique utilizes a two step selection process in an attempt to mitigate the difficulties associated with subjectively evaluating numerous criteria on a large sample set. The results of this work demonstrate promise but reveal problems associated with a small HRTF database and the need for carefully designed selection criteria[76].

Iwaya built upon the ideas in [76] experimenting with a "tournament-stye listening test" called "Determination method of OptimuM Impulse-response by Sound Orientation" (DOMISO). Using this method participants subjectively select a "fitted" set of HRTFs via a modified Swiss-style tournament (see figure 2.20)[43]. The procedure utilized in the experiment calls for 32 HRTF sets to be selected at random from a 120 set corpus to participate in the tournament. These HRTFs are used to generate 32 separate audio "orbits". These orbits simulate a sound source emitting pink noise moving clockwise around the listener on a horizontal plane. Prior to the listening test each participant is shown a diagram of the simulated orbit against which they should evaluate each HRTF set's relative performance.

Iwaya reports that users of DOMISO are able to select an HRTF set that performs nearly as

[9] a sound source is perceived on the wrong side of the head

Figure 2.20: The DOMISO Procedure [43]



well as their own measured HRTFs in sound localization tests. Reported incidence of front-back confusion with fitted HRTFs is lower than in the random case and the difference is shown to be statistically significant. Front-back confusion with fitted HRTFs is higher when compared with an individuals own HRTFs but the difference is not statistically significant. These results, coupled with the fact that the DOMISO procedure can be carried out in **minutes** as apposed to the hours typically required to measure an individual's HRTFs support Iwaya's suggestion that "DOMISO might be an effective method for the individualization of HRTFs"[43].

Though far more accessible to the general populace than direct measurement, the above mentioned techniques require participants to render subjective evaluations - something that they may be uncomfortable or have difficulty with. Seeber alludes to this problem and suggests the concept of "directional anchors for comparison with perceived auditory direction" as a possible solution but cites "complicated hardware" as a precluding factor to this approach[76].

Iwaya's evaluation of HRTFs fitted using DOMISO is limited to azimuth estimation on a horizontal plane, which is not as difficult as azimuth and elevation estimation in 3-dimensional space. Furthermore, Iwaya reports on a simulated "worst-case" or "away" condition[10] , but does

[10] utilizing a set of HRTFs that is never chosen in tournament

not provide a comparison with a more meaningful baseline, such as performance with an averaged or generic set of HRTFs[11] .

2.2.2.5 Implications to ESA Design

To date the rapid and effective acquisition of individualized HRTFs remains an open question. Where spatialization **is** performed, often a generic set or sets of HRTFs is employed[8]. The result is an audio display that generates a sense of three dimensional space but is not highly accurate at positioning virtual sound sources around the listener, particularly when perceived elevation is considered. This is of primary concern in an ESA design that seeks to employ spatial audio to indicate obstacle position and orientation.

Fusing Eye-tracking with Spatial Audio

It is possible that the audio errors introduced by generic HRTFs can be mitigated if coupled with some other method of perceptual positioning. The AuralEyes interface I propose In this work investigates the spatialization of gaze selected audio feedback. Under this model a user's pupil position is used to select a subset of input data from the direction of the user's gaze. The resulting audio signal is spatialized, such that the user perceives the origin of the sound to be located in the direction of their gaze. The resulting cross-modal interaction between the user's eye-position and their auditory perception of orientation may result in reduced ambiguity and increased confidence when interpreting audio feedback. This idea is loosely related to the concept of visual capture[46, 75] often cited in the phenomenon known as the "ventriloquism effect". I develop this concept further in chapter 4.

[11] as recorded from a KEMAR manikin for instance

Chapter 3

Historical Perspectives and Related Works

Having discussed many of the enabling technologies and principles behind mobile electronic sensory aids, I now present representative examples of related work in electronic sensory aids for the blind. I begin by presenting some of the foundational efforts in the field and then proceed, relatively progressively, through more recent developments in ESA research.

3.1 Audio Based Mobile Electronic Sensory Aids

3.1.1 Pointing Devices

At least as early as 1945 electronic pointing devices have been designed with the intent to augment or replace the traditional "white cane" [9]. These devices emit one or more signals in the form of electromagnetic or ultrasonic waves and utilize sensors to detect and decode the reflections of nearby objects. These reflected signals are processed and presented to the user in the form of audible sounds indicating contextual information such as range and/or height of the obstacle.

The 'K'Sonar is a modern example of such a device. 'K'Sonar is an ultrasonic "wand" that clips onto a traditional cane. The device emits a broadband ultrasonic sweep and converts the reflected sound into audible frequencies transmitted to the user via a headphone. The user learns to "decode" the spatial content of the transmuted echo which is reinforced by pitch; as reflections get closer they are decreased in frequency and increased in volume.[81]

Such "audio canes" have advantages over the traditional white cane, including increased perceptual range and the ability to detect obstacles without physically touching them, but carry

the added burden of power and other maintenance requirements. Obstruction and/or obfuscation of the existing audio environment is also a concern. Finally there is always the risk of causing/receiving interference at critical times or other forms of malfunction.

Haptic implementations of ultrasonic canes, such as the UltraCane[79], utilize tactile vibrations in lieu of sound thus leaving the user's natural audio environment unaltered. This preservation comes at the cost of reduced informational bandwidth. These devices retain the other aforementioned drawbacks associated with audio canes.

3.1.2 Navigational Beacons

Another approach to assisting the VI in navigation and self orientation is to modify the environment with verbal beacons and/or cues. Talking Signs®, an evolution of the "talking lights" concept proposed by Loughborough in 1979[53], implements this idea in a commercial product[19, 55].

Designed especially with public transit in mind, Talking Signs® employs infrared (IR) transmitters located at strategic locations in public areas such as transit centers, intersections, museums etc. The transmitters continuously emit pre-recorded messages providing contextual information ranging in complexity from simple "verbal signs" to detailed exhibit explanations. The IR messages are detected, decoded and presented to travelers as audio messages via handheld directional receivers. By "scanning" their surroundings with the receiver VI travelers can obtain detailed information about their environment. Research has found the effects of such systems to be highly beneficial to the VI community[19, 55].

Some of the most obvious, and perhaps most serious, limitations of this approach are the implementation and maintenance costs associated with wide-scale deployment of such technologies. A partial solution to this problem is to replace or augment the transmitters with "virtual beacons" simulated through the use of a wearable computing platform with an active positioning system.

3.1.3 A Personal Guidance System

Loomis et al. investigated a wayfinding system that utilizes a portable computer, electronic compass, the global positioning system (GPS), a spatial database and route planning software. The "Personal Guidance System" (PGS) receives keyboard input and is capable of providing the user with a conventional or virtual audio display[52].

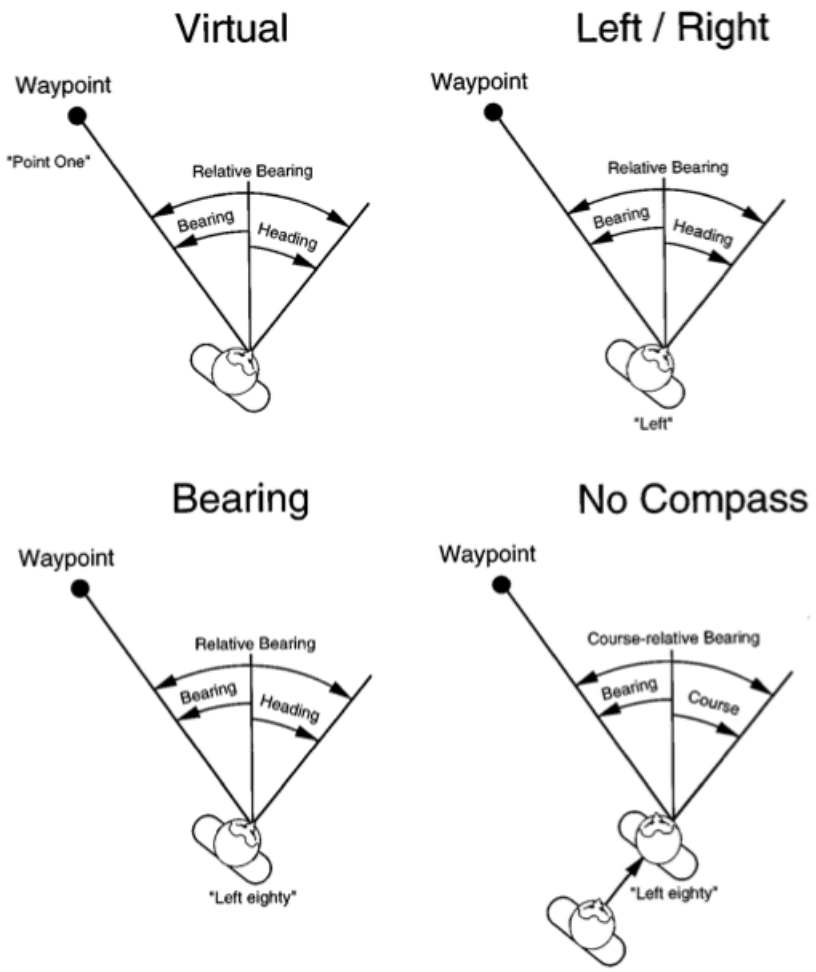
Using this system Loomis investigated four audio display configurations or "modes" referred to as, "virtual", "left/right", "bearing", and "no compass". In the first three modes the electronic compass is used to collect bearing information to be provided verbally to the traveler. The virtual mode employs spatialized audio (see section 2.2.2) to direct a traveler with a message such as "point one" which is heard emanating from the direction of the waypoint - thus providing directional information through auditory cues. The left/right mode provides simple ternary directions "left", "right", or "straight". The bearing mode adds information concerning the bearing to the target, i.e. "left eighty"; and the no compass mode provides the same information based upon the user's trajectory as extrapolated via the GPS system. See figure 3.1.[52]

Using this system users were able to navigate successfully using all four modes. The best performance and user preference ratings resulted from the virtual and bearing modes respectively[52].

Loomis' work illustrates the plausibility of a self-contained positioning and navigation system, as well as the potential value of virtual audio displays in assistive technologies. Virtual audio displays are discussed in greater detail in the following chapters.

Because GPS is typically ineffective inside of buildings, systems such as the PGS require a supplemental positioning system to operate indoors. The implementation details of such a system may re-introduce the cost and maintenance issues associated with physical beacons such as those utilized by Talking Signs®.

Figure 3.1: The Four Display Modes of the PGS [52]



3.2 Advanced Audio Displays in Assistive Devices

With the exception of the PGS' virtual mode, the assistive technologies discussed thus far can be implemented using very basic audio technology. Often a simple speaker or headphone is sufficient, and virtually no user information is required for implementation. Such simplicity has obvious benefits. However, in light of potential benefits such as added awareness and information throughput, researchers have long sought to utilize spatial audio in the context of assistive technologies.

Many systems and techniques have been devised to encode visual information through the use of VADs. Typically such systems spatialize synthetic sounds to communicate information about

both the relative location and type of objects within the user's environment. The sounds to be spatialized may be verbal cues such as the beacons referred to in section 3.1.2 or nonverbal audio signals which encode information according to a specially designed audio code.

Numerous approaches to this idea have been experimented with, including the work reported by Loomis et al. introduced in section 3.1.2 [52]. In the next section I discuss additional examples of VAD based assistive devices..

3.2.1 A Brief Chronological Survey of VAD Based Assistive Devices

Interest in the spatial qualities of audio can be traced back at least as far as the early 20th century. In 1930 experiments performed by C.C. Pratt revealed that listeners typically associate higher pitch with higher elevation in the vertical plane regardless of the actual elevation of the sound source[69]. Pratt's results were challenged in 1934 by Dimmick and Gaylord[21] who were unable to reproduce supporting results but subsequent works such as [66, 74] have reinforced his original findings.

By the latter half of the twentieth century researchers were leveraging technology to unlock the spatial characteristics of sound for application in ESAs.

3.2.1.1 An Early Audio Display

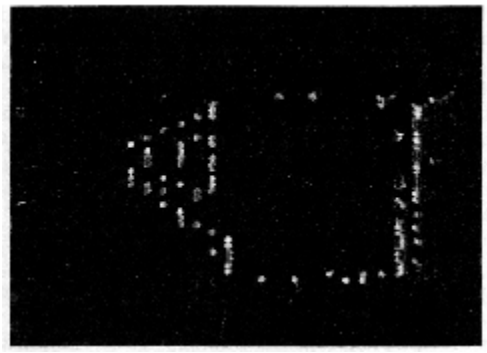
In 1975 Raymond Fish utilized the association between pitch and elevation in his work developing one of the first audio displays for the blind[28]. Perhaps the most important contribution of Fish's work was his "auditory code" for transmuting two dimensional patterns or images into audio representations that can be interpreted by human listeners.

Fish demonstrated two variants of his code. The first approach maps the bright regions of a scene to sound pulses. Scanning horizontally and then vertically (row-wise), bright regions are indicated by deliberately selected tones. The frequency of a tone is determined by the current position in the scan, with higher elevations being mapped to higher frequencies. Stereo ILD is applied to the sound to encode horizontal position within the scene. Reported frame rates for this

approach are on the order of seconds per frame.

The second variant of Fish's code is similar to the first but seeks to communicate edges rather than bright regions by producing a tone when a light-to-dark or dark-to-light transition occurs. Edges are detected both vertically and horizontally. Figure 3.2 demonstrates the kind of data resulting from this process when applied to an image of a coffee cup. Tone pulses would be generated corresponding to the bright "dots" in this image. This approach was found to be effective in conveying scene information while often allowing the system to transmit frames more quickly than the first variant.

Figure 3.2: Edge Detection of a Cup [28]



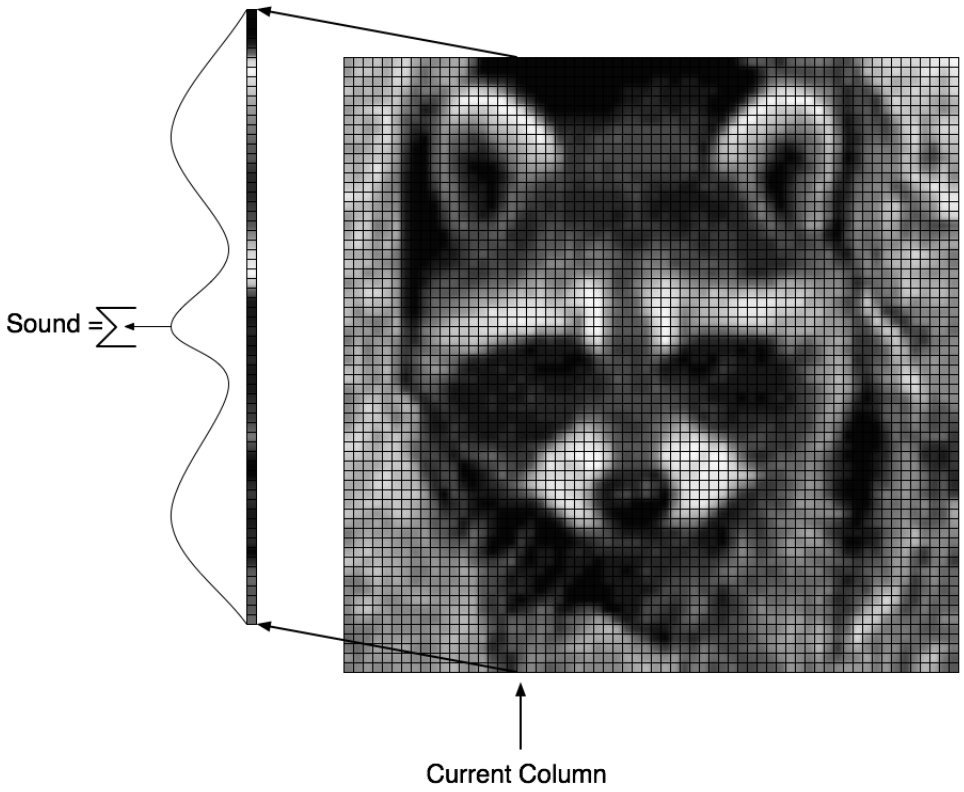
Using the technology of the time (namely photoelectric cells, oscilloscopes, tone generators, TV cameras, etc.) and a lot of ingenuity, Fish constructed four different prototype systems. These systems (termed systems I - IV) successfully demonstrated the potential of his audio code. User's of these systems were able to perform tasks such as identify simple and complex shapes, describe shapes they had not been exposed to previously, and (using one of the camera based systems) successfully navigate high-contrast obstacles in a room.

Though the bulkiness of the technology involved, and the operating frame rates of the systems were not suitable for everyday use; Fish's work established important foundational principles for transmitting visual information through an audio display.

3.2.1.2 The vOICe System

Building upon the work of Fish and others, in 1992 Peter B. L. Meijer proposed a low-cost portable system built around the idea of conveying images through an audio code with similarities to that proposed by Fish[58]. Like Fish, Meijer took advantage of the psychoacoustic nature of pitch to represent elevation. However, instead of performing a 2-dimensional pixel by pixel raster scan of an effectively black and white image Meijer's system combines the intensity values of an entire column of a grayscale image into a single sound comprised of multiple frequencies - somewhat like a musical chord comprised of multiple notes. The volume of each of the constituent frequencies is determined by the intensity value of the pixel it is associated with. Thus no sound is emitted for a completely black pixel and the maximum volume is employed for an all white pixel. An image is auralized by playing the combined frequencies for each column in sequence repetitively scanning from left to right(see figure 3.3).

Figure 3.3: The vOICe Audio Code



To emphasis the boundary between successive scans of a scene a "synchronization click" is produced between frames.

In general Meijer's approach allows a higher frame rate (on the order of 1 Hz) than those reported in Fish's work, especially in the case of complex images. Meijer's original system did not utilize ILD cues to enforce the current horizontal position of the scanline, instead relying solely upon the synchronization click and constant scan rate to indicate horizontal scanning position.[58]

The basic ideas of Meijer's original paper have resulted in several editions of the "vOICe"[1] . This system is described as an affordable, portable augmented reality system for the blind. The current system adds ILD as a horizontal cue, creating a "panning" effect as the scanline moves from left to right. By associating depth with brightness, depth maps can be auralized, allowing vOICe to operate in three-dimensions when coupled with a depth sensor. Though still considered under development by Meijer, users of the vOICe system are already reporting encouraging results. Beyond simply learning to identify objects and an increased sense of their surroundings, some users are reporting a limited restoration of the **perception** of sight[85]. This phenomenon is discussed further in section 3.3.1

Though the audio display utilized by the vOICe doesn't rely upon true spatialization as part of its audio code it represents an evolution of audio displays in assistive devices and demonstrates the viability of audio as a sensory substitute for sight.

3.2.1.3 The NavBelt

In 1990 J. Borenstein proposed the NavBelt, an electronic travel aid based upon the Obstacle Avoidance System developed for mobile robots at the University of Michigan[12]. The system utilizes an array of 16 ultrasonic sensors attached radially to a belt worn by the user. The sensors are arranged in two rows of eight with one row slightly angled upward and the other downward. The proposed NavBelt also incorporates an electronic compass and a doppler-effect distance sensor. [12, 77]

[1] OIC = "Oh I See"

The NavBelt can operate in two modes: Acoustic Guidance Mode - in which the user is actively guided toward a destination, or Acoustic Image Mode - in which an acoustic "image" is generated from the sensor data and presented to the user. Both modes utilize a binaural audio display that implements ITD cues to add a sense of direction to the auditory feedback.[12, 77]

In Acoustic Guidance Mode an audio beacon is played in the direction that the user is intended to move. Pitch and amplitude convey the recommended speed of travel via an inverse relationship. The theory behind this relationship is that higher volume and pitch attract more attention and cause a user to naturally slow down and pay attention to the sound[77]. Users of the NavBelt prototype were able to travel at an average rate of 0.76 m/s with an average directional deviation of 7.6 degrees.[77]

The Imaging mode of the NavBelt attempts to convey generalized information about the user's environment. Using the data from the sensors the system continuously presents an auditory "sweep" that is perceived from right to left. The amplitude of the signal indicates the range to the nearest obstacle detected by the device. The beginning of each sweep is indicated with a special "anchor signal" to assist the user in detecting the start of a new sweep. After several hours of training users were able to travel at an average speed of 0.4 m/s using the imaging mode of the NavBelt prototype.[77]

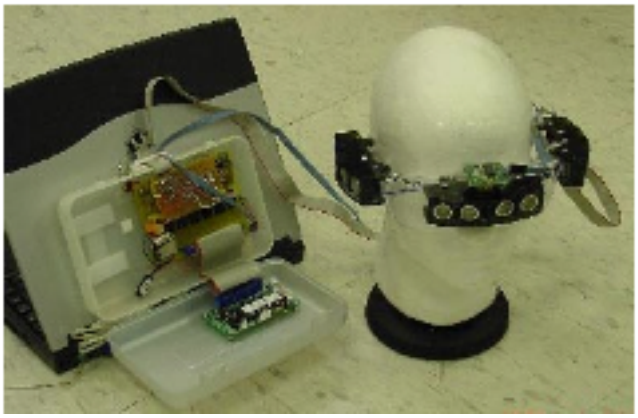
3.2.1.4 A Pocket-PC Based Navigational Aid for Blind Individuals

An issue of obvious concern when navigating any environment is the distance from a traveler to the obstacles in their immediate surroundings. The range-finding problem is one that seems particularly well suited for ultrasonic sensors. Not surprising numerous travel aids for the blind have been constructed around this technology [12, 13, 17, 79, 81]. I introduce here the work of Choudhury et al. as an example of such a system.

Choudhury et al. have proposed "A Pocket-PC Based Navigational Aid for Blind Individuals"[17]. Examples of the hardware utilized in this device can be seen in figure 3.4 which actually shows the notebook based precursor to their Pocket-PC based system. For range acquisition an array of three

head-mounted ultrasonic rangefinders are directed at 30, 60, and 120 degrees azimuth and mirrored by another three on the opposite hemisphere of the head for a total of six sensors. In addition to the sensors the headgear is equipped with a magnetic compass module used for determining magnetic North. Management of the sensors and compass is accomplished by a micro-controller based Sensor Control Unit (SCU). The SCU communicates with the host computer (a Pocket-PC in the final implementation) through an RS232 interface operating at 9600 bps. The SCU delivers range and bearing information in response to command sequences provided by the host.[17]

Figure 3.4: A Head Mounted Ultrasonic Navigational Aid [6]



Note: This image shows the laptop based precursor of the system discussed in this paper.

Users receive spatial or navigational information in the form of a VAD generated by the host computer. This VAD operates in two modes as selected by the user: "Obstacle Map Mode" (OMM) and "North Beacon Mode" (NBM). In OMM the host computer combines and plays six individual pre-convolved sound samples at varied intensities relative to the range information generated by the corresponding six rangefinders. A separate filter is used for the left and right channels for each of the six sounds to accomplish full binauralization. The filters used in the convolution are reported to be measured HRTFs though it is unclear if a different set of HRTFs was measured for each test subject participating in the experiment. In NBM a single sound source is active (i.e. one sound sample played through the two channels) whose perceived location indicates the direction of magnetic North relative to the user's heading.[17]

In experiments blindfolded users of this system were able to navigate through a building in a Northerly direction, negotiating corners as they traveled, at an average rate of 0.57 ft/sec. Though this rate is dismal when compared to normal walking speeds (on the order of 4 to 5 feet per second[84]) it is unclear how much practice time (if any) participants were allowed with the device. Since the volunteers were sighted it is unlikely they were accustomed to navigating in the absence of sight in general.

It is significant that the average efficiency ratio of the routes followed by the volunteers was approximately 0.93. This would seem to indicate that, though they were moving slowly, the participants were making good decisions about the path they chose[17]. This demonstrates that the information conveyed by the VAD was meaningful if not familiar and easily processed/trusted. With this in mind it seems plausible that with additional practice travel rates would improve.

3.2.1.5 SWAN

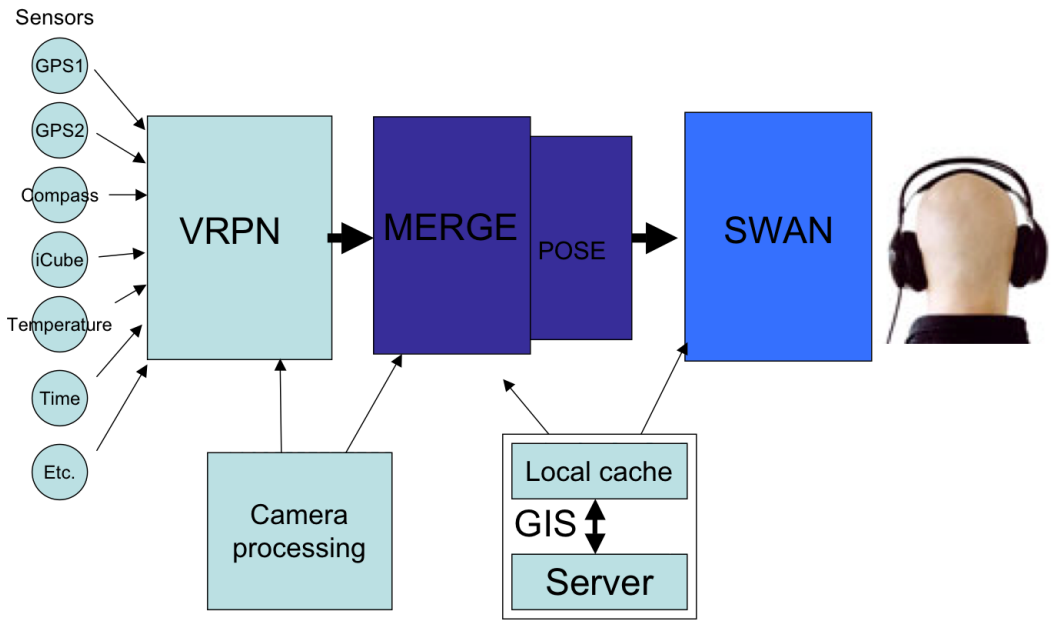
Wilson et al. at the Georgia Institute of Technology have developed the "System for Wearable Audio Navigation" (SWAN). This impressive system incorporates the principles and technologies of numerous previous research efforts (including many of those discussed in this paper) into a single configurable solution designed to be customizable to the specific needs of a given application.

The goal of SWAN is to aid users in "safe pedestrian navigation" by supporting "wayfinding, obstacle avoidance, and situational awareness"[48]. To accomplish these goals the system relies upon the fusion of sensing, localization and orientation data from numerous sensors and cameras as well as a remote Geographic Information System (GIS) database that users can use to share information. Assistive information is provided to the user through a 3D VAD utilizing non-individualized HRTFs[2]. The audio display is presented through traditional or bone conducting headphones. The system is controlled primarily through an audio menu that is navigated via a handheld PC-mouse derived controller[48, 95].

Of particular interest to this work is the audio display utilized by SWAN. Both verbal and

[2] Users are warned of potential front/back reversals etc.[95]

Figure 3.5: The SWAN Platform [48]



non-verbal cues are employed by the VAD, these include:

Navigational Beacons

Navigation assistance is available to guide a user along a specified path using spatialized non-speech "beacon sounds" that continuously indicate the direction the user needs to go - similar to the "virtual mode" implemented by Loomis in [52].

Environmental Features

Environmental features that may be useful to the user such as surface transitions or identifiable objects such as park benches, restrooms, bus stops etc. These are also represented as spatialized non-speech sounds. Such features may be extracted from camera or sensor input or stored in the GIS database.

Audio Annotations

Using the menu system users may add audio annotations to the GIS database to inform other users of relevant information about a specific location. This information might be advisory in nature (i.e.

"slippery here when wet") or informative (i.e. historical information etc.). This capability bears a striking resemblance to the purposes served by Talking Signs®discussed in section 3.1.2.[95]

In wayfinding experiments simulating travel in a virtual reality simulator SWAN users have demonstrated high levels of proficiency at speeds comparable to typical walking rates of sighted persons. "Other than the method of locomotion, participants who have used both the virtual prototype and the physical SWAN system do not report any major differences between the experiences."[88]

SWAN has successfully demonstrated itself as a capable and flexible sensory aid, solidly demonstrating the value of VADs in mobile ESAs for the blind. Furthermore the system serves as a powerful research tool for researchers seeking greater insight into the implications of new VAD based technologies and techniques.[88]

3.3 Related Developments

3.3.1 Neuroplasticity

When a sensory aid performs the role of replacing a non-functional sense (i.e. total blindness) it may be termed a sensory substitution device. Numerous neurological studies have investigated the neurological effects of long term use of sensory substitution devices (SSDs) like the vOICe system. There is mounting evidence that with sufficient training and exposure the human mind can "rewire" itself to take advantage of the information provided by remaining senses when one or more are lost. The ability of the brain to adapt in this manner is called crossmodal neuroplasticity. Studies have shown that users of Meijer's vOICe system are able to recruit portions of the mind typically utilized for the processing of sight to "decode" the spatial information contained within the audio code.[60, 85]

Experiments by Kupers et al. involving a tongue display unit (TDU) have produced similar results[47, 70]. The TDU is a two dimensional conductive grid that presents software processed image data to a user in the form of electrical stimuli to the tongue. Positron emission tomography (PET) scans pre and post-training with this device show an increase in activity in the visual

cortex of blind users of the device, demonstrating neuroplastic adaption to the augmented sensory input.[47, 70]

Neuroplasticity has also been observed in more "natural" forms of sensory substitution. Functional MRI scans of blind echo-locators who use tongue clicks or pops to "probe" their environment have revealed increased activity in the visual cortex when these individuals are exposed to recordings of their echolocation clicks and echoes. [82]

These results demonstrate the human brain's capacity to identify and utilize spatial information embedded within a variety of alternate stimuli. This is highly encouraging and suggest that more sophisticated systems and techniques might result in significant improvements in quality of life for VI persons who choose to utilize sensory substitution devices.

Chapter 4

AuralEyes: Increasing the Value of Synthetic Sensory Feedback Through a Reduction in Quantity and an Increase in Relevance

In section 2.1.4 I suggested that the usability of many ESAs for the blind would be improved if users had an effective hands-free method of exerting control over the synthetic sensory feedback generated by such devices. This is especially true of audio based devices which run the risk of overloading a user's audio sense - an unacceptable scenario for the blind. Allowing the user to identify a subset of the available information as relevant can result in a reduction of the **quantity**, accompanied by an increase in the **quality** of information presented. Furthermore, this mode of operation is more closely aligned with our natural use of sight and may produce more natural interactions between user and device.

4.1 Attention Driven Senses

For humans, hearing differs significantly from the other senses in that we do not have the same level of control over its application. Sight is easily directed, or limited, by moving the eyes and/or head, or closing of the eyelids.

Touch can be utilized to explore our environment through intentional and controlled movements. Taste and even smell can be physically directed or controlled[1] .

Attention driven manipulation of audio input for humans is significantly more limited. We can rotate our head or cover our ears - but neither action produces the level of affect we are able

[1] Though there is an obvious limit to the length of time a person can hold their breath!

to exert over our other senses. Our sense of hearing operates largely outside of our control both day and night. The audio displays of most ESAs operate in a manner more closely modeled after the way humans naturally interact with hearing than sight. This operational difference between the sense being utilized and the sense being replaced may be presenting an unrecognized obstacle to the adoption of audio based ESAs for the blind.

4.1.1 Lessons From Nature

A limited ability to physically direct hearing is not universal in the animal kingdom. Several mammals such as dogs, cats and horses have the ability to swivel concave ears, allowing them to focus their auditory sense toward a specific region of interest. Incorporating this ability into the audio display of an ESA for the blind would have numerous advantages.

Figure 4.1: Attention Directed Hearing in the Animal Kingdom

4.2 Fusing Eye Tracking with Spatial Audio

Eye tracking has been shown to be an effective method of enhancing computer interaction for sighted individuals [27, 50, 89]. To date I am unaware of this technology being utilized by the blind community. I propose user interfaces based upon the fusion of eye-tracking and audio in ESAs for the blind. I also propose the spatialization of the audio stream as a means of reinforcing

orientation in such interfaces.

This approach provide users a virtual "joystick", with which they can manipulate the behavior of a mobile ESA. For example, a device might limit the information presented to that which is collected from a region determined by the current direction of the user's gaze. Alternatively, all of the other information acquired could be presented, but at a greatly attenuated volume or level of detail. In this way a user is able to "listen" where they are looking.

A simple use-case illustrates some of the benefits of this approach. A person attempting to navigate a sidewalk is primarily concerned with obstacles that may be present in their intended path. Sighted individuals steer their gaze in the direction they intend on traveling, they are less concerned with what is going on in the periphery, and pay virtually no attention to obstacles behind or above them. Indeed the central, or foveal, vision model employed by humans further illustrates the focussed and directed nature of sight. By restricting the feedback from an ESA to a region indicated by the user's gaze, more detailed information about **that region** can be presented to the user without overloading the surrogate sense. The quantity of information can be reduced while its relevance (i.e. quality) enhanced.

Coupling eye-tracking with spatialized audio has the potential for all of the demonstrated benefits of spatial audio in ESAs, with the added benefit that the spatial qualities of the sound signals reinforce the intended actions of the user. Users are able to **perceive** that the information they are receiving is relevant to the intended area. Such consistency between the action of the user and the feedback of the device may increase confidence that the ESA is working properly and that the information is relevant.[2]

4.3 AuralEyes

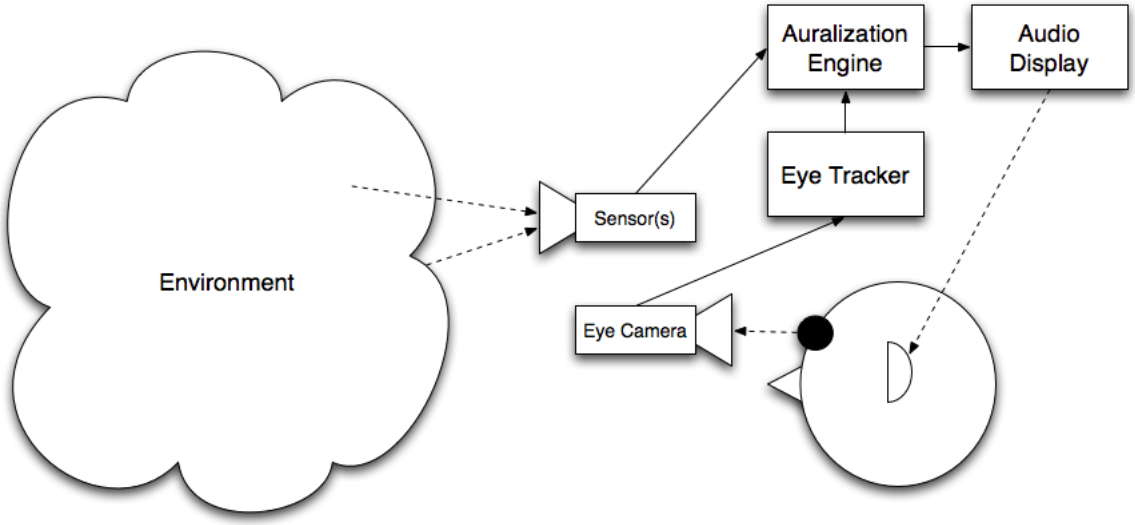
To investigate the theory that eye tracking coupled with spatial audio improves the usability of ESAs, I have developed a user interface called "AuralEyes". This system incorporates the fusion

[2] Conversely, if a user directs their gaze in one direction (i.e. left) and the feedback from the system appears to come from another (i.e. right) the user is immediately aware that something is amiss and can act accordingly.

of eye-tracking with audio feedback. The audio may or may not be spatialized as is appropriate.

Figure 4.2 presents a the high-level view of the proposed architecture.

Figure 4.2: AuralEyes System Overview

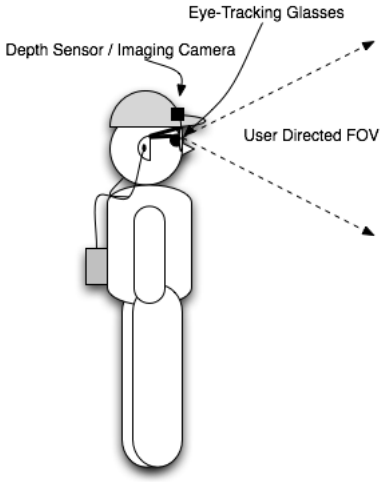


The system operates by first capturing information from the environment via one or more sensors. This data is fed into an Auralization Engine which performs a mapping from environmental information to audio signals based on a predetermined audio code or codes.

The Eye Tracker module captures an image from the eye camera and determines the elevation and azimuth of the user's attention based upon pupil position. The Auralization Engine selects/generates the appropriate audio feedback from the scene data based on the orientation of the user's eye. This audio signal is then presented to the user via an audio display. The display may be comprised of headphones, bone-conducting speakers etc. If left and right channels are available then the audio display may spatialize the audio via HRTF filtering. The result of this feedback loop is an audio signal reflective of the environment in the region the user is looking at. When spatialization is used, this audio signal **seems** to emanate from the region of focus.

Figure 4.3 depicts a hypothetical ESA that utilizes the AuralEyes interface. Eye-tracking glasses and a head mounted depth/video sensor gather gaze direction, head orientation information, and scene data. A mobile PC carried via a waist or back-pack performs processing of the input

Figure 4.3: Hypothetical AuralEyes Based ESA



data and provides the necessary audio signals via stereo headphones.

4.3.1 Considering Late and Early-Onset Blindness

Many people with blindness were once sighted and have already developed the motor control necessary to direct their eyes towards a region of interest. Individuals with early-onset blindness will likely not have this capability. This does not necessarily preclude such individuals from utilizing the kind of interfaces I propose in this work. Rather, this condition creates an additional opportunity to investigate the psychological issues surrounding the development of ocular motor skills through the use of the Aural Eyes system. Indeed it is possible that early use (i.e. childhood) of such systems may be advantageous to developing the necessary motor control. These topics are among the areas of future work discussed briefly in chapter 7.

Chapter 5

Materials and Methods: A "Zero-Day" Usability Study

The following research was conducted with the approval, and under the oversight, of the Institutional Review Board at The College of Idaho. The IRB approval number for this study is 1097.1.

In order to evaluate the hypotheses put forth in this work I designed and administered a comparative usability study. In contrast to numerous previous works, this study was not focussed primarily on the performance of well-trained subjects, but rather on a user's performance and experience while attempting to complete meaningful tasks with an ESA for the first time. A principal impetus for this work was, and is, the presupposition that increased usability in the short-term has a positive effect on an ESA's adoption rate in the long term. Therefore, evaluating the usability of AuralEyes in the short term was both a means of evaluating the hypotheses of this work as well as a first step towards understanding how the ideas embodied within AuralEyes might be effectively implemented in ESA designs that can achieve meaningful adoption rates within the blind community.

I begin by presenting details of the final protocol approved by the IRB at The College of Idaho. The official protocol, incorporating amendments, can be found in Appendix A. I close this chapter with a discussion of difficulties that arose during the execution of the experiment that have affected the data collected from the experiment and its subsequent analysis.

5.1 Study Design

5.1.1 Performance Evaluation

The first part of the experiment evaluated a user's ability to perform localization and identification tasks on simulated depth maps. Three ESA configurations were evaluated.

A - vOICe Learning Edition

B - AuralEyes Without Spatialization (AE Mono)[1]

C - AuralEyes With Spatialization (AE Spatial)[2]

In the experiment, and for the remainder of this chapter, these three systems are referred to as vOICe, AE Mono, and AE Spatial respectively.

5.1.1.1 Experimental Apparatus

Figure 5.1 presents the high-level architecture of the experimental apparatus used to simulate the three ESA interfaces and audio displays. The simulation software can be divided into modules that perform three major tasks - scene generation, audio auralization and data collection.

Scene Generation

The Scene Generation module provides simulated "scene data" in lieu of live sensor feedback. A precompiled image bank stores 81 synthetic depth maps, one of which is shown in figure 5.2. In these images depth is indicated by brightness or intensity, with brightness indicating "nearness". In other words the brighter a region the closer it is to the user. Nine regions of uniform depth, organized into three rows and three columns, are represented. The Scene Generation module randomly selects depth maps from the image bank, as appropriate, based upon user task. This is discussed in greater detail below.

Active scene data is not displayed on the computer monitor, though mouse-cursor position is used to determine the user's region of interest within the scene as if the image were being projected

[1] Referred to in the original protocol as AuralEyes Mode 1
[2] Referred to in the original protocol as AuralEyes Mode 2

Figure 5.1: Architecture of Test Apparatus

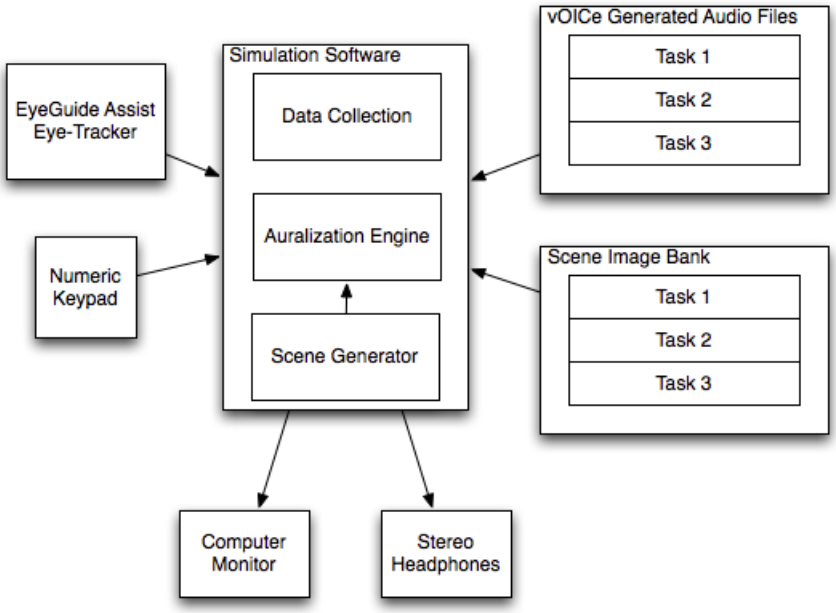


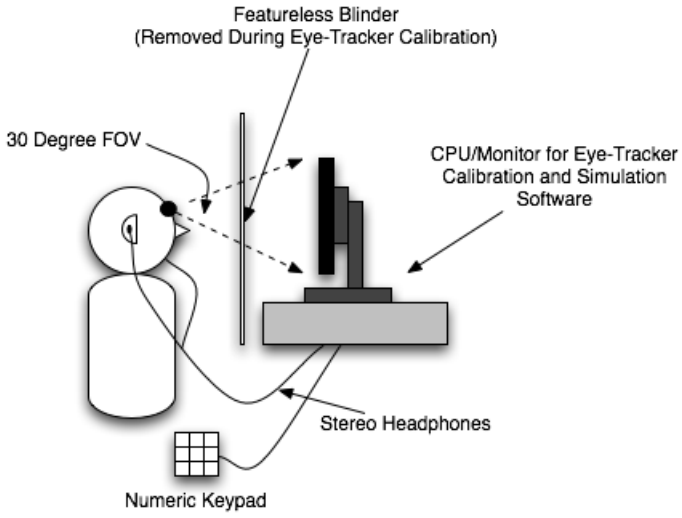
Figure 5.2: Sample Scene Data

onto a 30 by 30 degree viewport located at the center of the display (see figure5.3). The cursor's position is controlled via a head-mounted EyeGuide Assist eye tracking system. In this way eye tracking within the virtual environment is implemented.

Auralization Engine

In this experiment the audio signals of the three ESA systems are generated by the Auralization Engine based upon the depth information provided by the Scene Generator. This audio feedback

Figure 5.3: Physical Test Apparatus



is presented to the user via a pair of stereo "ear-bud" headphones.

In the case of vOICe, pre-recorded audio was generated for each depth-map via the vOICe learning edition software[59]. When a user is utilizing vOICe the Auralization Engine selects the appropriate audio file, based on the scene image selected by the Scene Generator, and plays it back to the user.

Audio feedback for AE Mono and AE Spatial is generated in a 3 step process.

(1) The orientation of the user's gaze within the virtualized depth scene is determined via the EyeGuid Assist Eye-Tracker.

(2) The depth at this orientation is determined from the depth-map and used to generate an appropriate audio signal.

(3) i) In the case of AE Mono, this audio is played back to the user as a mono-channel signal played in both ears.

ii) In the case of AE Spatial, this audio is spatialized using a generic set of HRTFs from the CIPIC database (subject 21 - KEMAR manikin) to simulate a sound source positioned in the direction of the user's gaze.

The audio signal generated for AE Mono and AE Spatial is a repeating audio pulse (a "tick") whose frequency is determined by a linear scaling with the intensity (proximity) of the region indicated by the eye-tracker. The closer the region the higher the frequency. For this experiment legal proximity values ranged from 0 - 255 (255 being as close as possible) and frequencies calculated on a range from 0 to 18 pulses per second[3] . The resulting pulse frequency is calculated as in equation 5.1.

$$PulseFrequency = Proximity * \frac{18}{255} \tag{5.1}$$

A floor of 0.25 Hz was enforced during the experiment to provide an active cue to participants that the system was still functioning even if they were investigating a region at the most distant range (i.e. Proximity = 0);

Data Collection The Data Collection module is responsible for recording participant performance during the experiment. User input is gathered by way of a numeric keypad. The details of each task being performed are captured, including user success/failure, and the amount of time taken to make a selection. These data are written to a file labeled with a participant identification number provided to the experimental software when the program is first invoked.

5.1.1.2 Experimental Procedure

During the experiment, participants were seated at a table with their head steadied by an adjustable chin rest. A flat-panel monitor was positioned a fixed distance in front of the subject's face for use in calibrating the eye-tracker. The distance to the screen (18.75 inches) was selected based upon the physical dimensions of the largest equilateral viewport that could be displayed on the screen (10.06 inches square) - such that this viewport occupies a 30 degree square region. An adjustable chair was utilized to ensure that each participant's eyes were positioned horizontally parallel with the center of the screen, thus centering the 30 degree viewport directly in front of the user's eyes.

[3] I selected this range subjectively in preliminary trials leading up to the experiment

Once the eye-tracker was calibrated a large, white, featureless poster board was placed directly in front of the user - serving as a virtual blindfold while enabling the eye-tracker to function.

Throughout the experiment the head mounted eye-tracking camera and ear bud headphones were appropriately affixed to the subjects head and ears. For consistency subjects wore the headphones and eye-tracker while performing tasks and during orientation periods, regardless of the system currently being evaluated. As the experiment progressed the performance of the eye-tracker was monitored and recalibrated as necessary to ensure consistent accuracy.

Each participant performed three iterations of tasks with variants of the AuralEyes and vOICe systems. The initial protocol specified that half of the subjects (group ALPHA) would utilize the vOICe system first, followed by AE Mono and then AE Spatial, the other half of the participants (group BETA) were to reverse this order. Subjects were thus partitioned into two sub-pools that performed three iterations of tasks as follows:

Table 5.1: Division of Participants into Two Groups By System Order

Iteration	ALPHA	BETA
1	vOICe	AE Spatial
2	AE Mono	AE Mono
3	AE Spatial	vOICe

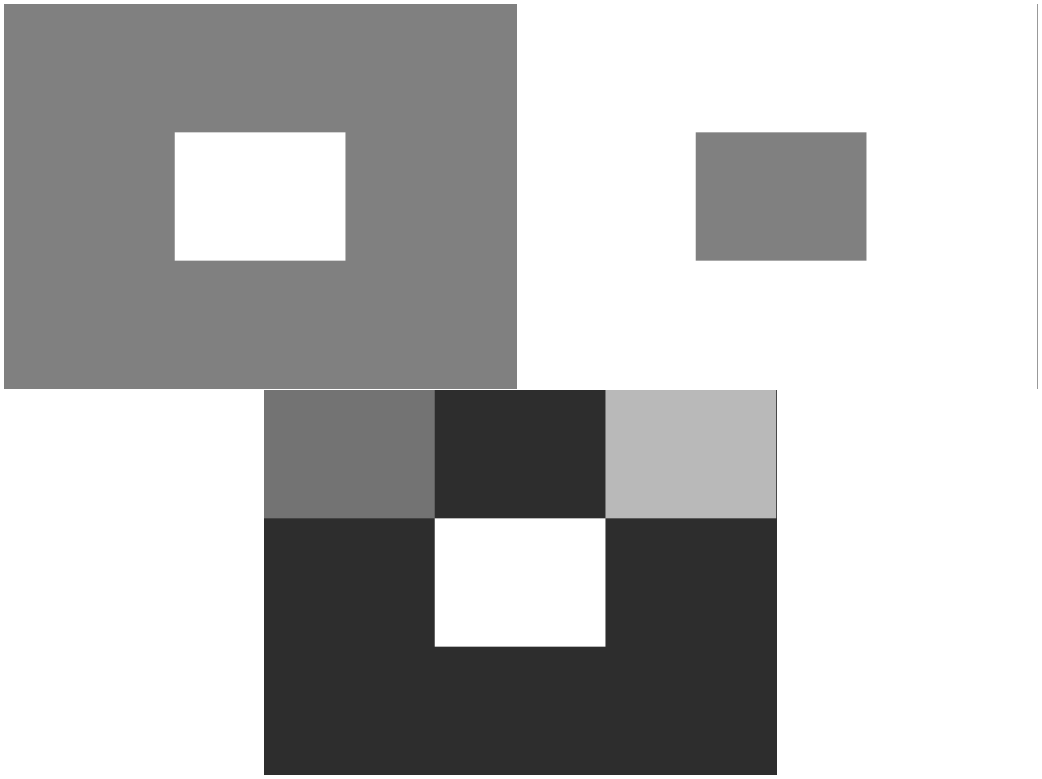
A short orientation was provided for each system immediately before it was first used. This orientation included a short technical description of how the given system works (see Appendix B for the scripts used), followed by an up to five minute "experimentation" period in which the subject was able to listen to the audio signals generated from a training scene described to the user. During this time participants were allowed (and encouraged) to ask questions and explore the system under test. If requested, and if time allowed, participants were allowed to "practice" data entry procedures but were not provided feedback concerning the correctness of their selections. The same initial training image and description were used for all systems. The image and its description are provided in Appendix A of this document.

Throughout the study the labels "System A" , "System B", and "System C" were used

to identify the three system configurations. The vOICe system was referred to as “System A”, AE Mono as “System B” and AE Spatial as “System C” - when communicating with subjects or collecting feedback.

Following the orientation period for each system, participants were given a series of 3 sets of tasks to complete. Each of these task sets required the subject to analyze three depth images (or scenes) individually, for a total of nine images on which to perform a selection - per system. Each image was logically divided into a 3 X 3 tic-tac-toe-like grid of nine regions, with each region occupying approximately 10 degrees of the participant’s field of view both horizontally and vertically. The central region was virtually centered at, or near, eye-level and directly in front of the user. Figure 5.4 provides a visual representation of 3 sample scenes (two above, one below). In these simulated depth maps proximity is indicated by brightness with an increase in distance/depth indicated by a reduction in intensity (i.e. black = distant, white = close).

Figure 5.4: Sample Scene Data



The three images presented during each task were of high, medium and low contrast respectively; contrast being defined as the difference in range between the nearest and furthest regions. For this experiment range values were defined on a linear scale from 0 to 255 with 0 being the lowest allowable rage and 255 the maximum[4] [5] . In this document I will present range values in terms of a percentage of the maximum allowable range. In other words a range of 255 will be reported as 100% of the maximum range etc.

Each of the scenes to be displayed was selected randomly from a subset of relevant images from the image bank described in section 5.1.1.1. The subset of possible images was defined according to task type and contrast level. The three sets of tasks to be performed were defined as follows:

(1) Identify the closest region (simple)

(2) Identify the region furthest away

(3) Identify the closest region (complex)

For task one, each scene contained a single region that was closer than the other eight. The eight "distant" or "far" regions were all at the same simulated distance from the participant.

Scenes for task two had a single region that was more distant than the other eight. The eight "near" regions were all at the same simulated distance from the participant.

Task three scenes contained **three** regions which were closer than the other six. Each of the three "near" regions was at a unique range from the user. One of these regions was therefore the closest - and consequently target region. The six most distant regions were all at the same simulated distance from the participant.

As mentioned above, the three images selected for each task set were of high, medium and low contrast respectively. For tasks 1 and 2, where only two range values were present in an image at a time (i.e. "near" and "far") range values were defined as in table 5.2.

[4] In the actual software implementation these values were inverted, with 255 representing the lowest possible rang, 0 the highest - this was in harmony with grayscale representations of the depth maps which represented increased distance with decreased intensity

[5] In real world applications, where range data are acquired vs. synthesized, these values are scaled and mapped to the range and units of the acquiring sensor.

Table 5.2: Task 1 and 2 Range Values For Near and Far Regions as a Percentage of Maximum Range

Contrast	Near	Far
High	0%	100%
Medium	0%	50%
Low	40%	50%

The ranges of the three near regions and six far regions for task 3 are listed by contrast in table 5.3

Table 5.3: Task 3 Range Values For Regions as a Percentage of Maximum Range

Contrast	Nearest	2nd Nearest	3rd Nearest	Far
High	0%	40%	78%	100%
Medium	0%	27%	55%	82%
Low	0%	15%	30%	57%

For each task participants were allowed up to 30 seconds per scene to complete the specified tasks. A 5 second warning was provided verbally at 25 seconds.

Scenes were randomly selected for each iteration and subject to ensure that subjects could not apply prior knowledge to the tasks, and to avoid potentially biasing the results in the event that certain scene configurations were easier (i.e. scenes with a target region located in the center for instance). To prevent the researcher from influencing the participant, the depth-map being auralized was not visible in any form during the experiment, and results were not made available until the subject had completed all tasks.

Participants were instructed to enter their selections via a ten digit numeric keypad as well as indicate them verbally. Verbal selections could be made by relative column and row descriptions (i.e. "top-left"), or by indicating the corresponding numeric value on a ten digit keypad (i.e. "top-left" = 7). When a participant's verbal response contradicted the data recorded via the number pad, the verbal response was assumed to be correct[6] . In this manner consistent and unbiased timing information could be recorded as well as data integrity ensured - even if users unaccustomed

[6] This "corrective" procedure was explained to subjects before collecting data

to the number pad entered an unintended value. Occasionally subjects would begin to speak a selection before hitting the appropriate key and run out of time to enter their selection via the number pad. In such cases the verbal response was accepted and entered with a timestamp of 30 seconds.

Data collected for the tasks included: failure/success to select a region, the target region, the region selected by the subject, and time to make a selection. If a participant failed to make a selection then the maximum time of 30 seconds was recorded. Subjects were encouraged to do their best on each task but also informed that they had the option to "give up" and move on if they were simply unable to perform a certain task.

Up to a five-minute break was available to subjects when switching between systems. Most participants chose to bypass or shorten this break.

5.1.2 Usability Questionnaire

Upon completing the tasks outlined above each participant was asked to complete a short, one-page questionnaire about their experience with the three systems. The questions on this form were designed to ascertain a subject's impressions and reactions to the systems under test. Participants were asked to comparatively rate the intuitiveness of the three systems, as well as the level of fatigue induced by their individual audio signals. Finally the subject was asked to indicate which of the three systems they would prefer to use. The actual questionnaire is provided in Appendix A of this document. Participants were asked to complete this questionnaire before viewing their performance data.

5.1.3 Debriefing

After completing all aspects of the experiment subjects were fully debriefed. Participants were offered the debriefing document included in Appendix A, as well as given the opportunity to view their data and ask questions about the research. Subjects were asked not to discuss their results and the details of the experiment with other potential participants. Finally, the option of

receiving a summary of the experimental results at the completion of the study was offered.

5.1.4 Complications and Data Analysis

My intent in the study design outlined above was to generate data sets that could be paired for system level comparisons within each group (i.e. AE Mono vs. AE Spatial), while balancing the numbers of males and females in each group to allow for meaningful comparison between them (i.e. ALPHA vs. BETA). This design would have, in many cases, enabled the use of a simple paired samples t-test to look for significant differences between systems and groups.

Unfortunately technical problems at the onset of the experiment required modifications to the protocol, which rendered a significant amount of data unusable - the result was a smaller sample size than I had initially planned for. In addition, a clerical error resulted in a mismatch of the male to female ratios in the two groups; ALPHA was comprised of 5 males and 7 females, BETA contained 3 males and 6 females.

Due to the imbalance in sample size and composition, it was not meaningful to perform intergroup comparisons. Consequently, it became necessary to analyze the data collected from groups ALPHA and BETA separately and look only for significant trends which were present in both groups. Within each group I looked for:

(1) Performance differences between the three ESA systems and theoretical random selection.

(2) Differences in performance when attempting different types of tasks with a given ESA system (i.e. task level intra-system comparisons) .

(3) Performance and user preference trends between AE Mono and AE Spatial.

I present the data collected from the vOICe system principally as an informal secondary reference, against which to "loosely" assess AuralEyes. This data is also useful in establishing a preliminary understanding of the strategies (and implications thereof) employed in scene based systems such as vOICe.

5.1.4.1 Statistical Tests

The primary statistical tools I utilized when analyzing the data collected from the usability study are the well known student's t-test (paired samples) and the exact binomial test. In performing these tests I utilized the statistical software package R in conjunction with Microsoft Excel. I present the results of my analysis in chapter 6.

Chapter 6

Results and Discussion

In this chapter I present the results of the the zero-day usability study described in chapter 5. Eight male and thirteen female participants between the ages of 18 and 34 were recruited. All subjects self-reported normal hearing and normal range of eye motion. Participants were divided into two groups. The first group (ALPHA) performed tasks using vOICe, AuralEyes **without** audio spatialization (AE Mono), and AuralEyes **with** audio spatialization (AE Spatial) in that order. For the second group (BETA), this ordering was reversed.

6.1 Success Rates

Table 6.1 presents the success rate data for groups ALPHA and BETA.[1] Mean values are computed as the mean of the individual success rates[2] for a given group, system and task. The fifth column of the table lists the standard deviation of said means. The sixth column presents the sum of correct selections for the entire group for the given task. An exact binomial test (one-tailed) was applied to test for significance in the difference between the observed success rates and that expected by random selection (1/9). The seventh column lists the resulting p values with significant results (alpha = 0.05) marked with '*'.

For clarity the means listed here are presented again graphically in figure 6.1 in the form of bar charts. Once again, mean values that are statistically significantly better than random are marked with '*' (exact binomial test (on-tailed) p = 0.05).

[1] The raw success data are available in Appendix E.

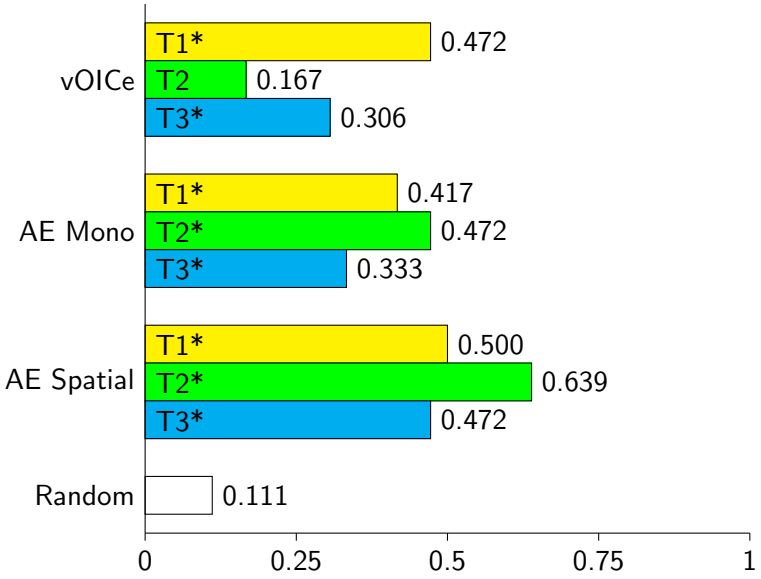
[2] computed as the average number of successes per task

Table 6.1: Mean Success Rates, Standard Deviations and Counts by Group, System and Task

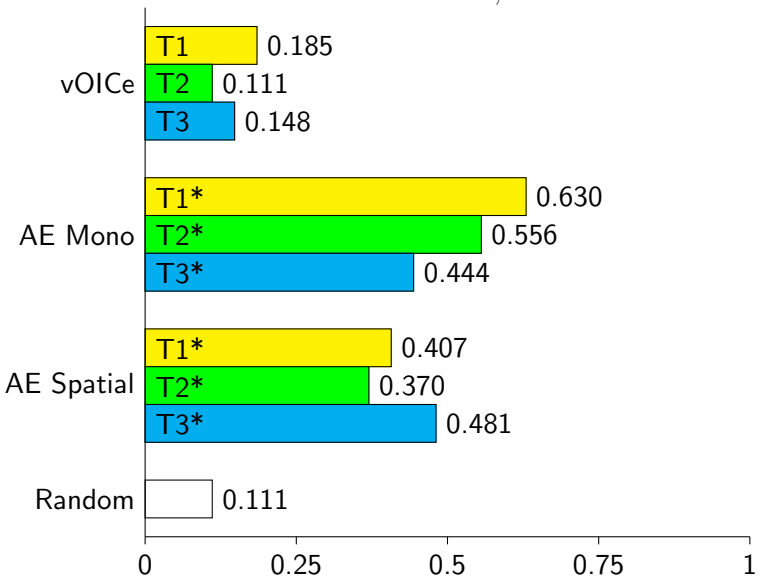
Group	System	Task	Mean	StDev	Correct	p
ALPHA	vOICe	Task 1	0.47	0.44	17	< 0.01*
		Task 2	0.17	0.22	6	0.21
		Task 3	0.31	0.22	11	< 0.01*
	AE Mono	Task 1	0.42	0.25	15	< 0.01*
		Task 2	0.47	0.36	17	< 0.01*
		Task 3	0.33	0.28	12	< 0.01*
	AE Spatial	Task 1	0.50	0.33	18	< 0.01*
		Task 2	0.64	0.36	23	< 0.01*
		Task 3	0.47	0.39	17	< 0.01*
BETA	vOICe	Task 1	0.19	0.24	5	0.17
		Task 2	0.11	0.17	3	0.59
		Task 3	0.15	0.24	4	0.35
	AE Mono	Task 1	0.63	0.42	17	< 0.01*
		Task 2	0.55	0.29	15	< 0.01*
		Task 3	0.44	0.33	12	< 0.01*
	AE Spatial	Task 1	0.41	0.32	11	< 0.01*
		Task 2	0.37	0.39	10	< 0.01*
		Task 3	0.48	0.38	13	< 0.01*

* = statistically significant result using binomial test with alpha = 0.05

Figure 6.1: Group ALPHA Mean Success Rates for vOICe, AE Mono and AE Spatial



Group BETA Mean Success Rates for vOICe, AE Mono and AE Spatial



In all cases the AuralEyes based systems resulted in performance that was significantly better than random selection, supporting one of the primary hypotheses of this work.

Contrary to my expectations, spatialization did not appear to have an effect on performance. It would seem that the participants' sense of eye-position was sufficient to make determinations with equal accuracy with or without audio spatialization - at least in the short term. It is worth noting however, that all of the subjects for this experiment were sighted and using the virtual

blindfold. It remains possible that spatializing the audio signal would have a stronger effect on blind subjects, perhaps especially those with early-onset conditions. For now, I leave this question for a future experiment.

6.2 Task Completion Rate and Time

6.2.1 Completion Rates

Table 6.2: Mean Task Completion Rates and Standard Deviations for vOICe, AE Mono and AE Spatial

Group	System	Task	Mean	StDev
ALPHA	vOICe	Task 1	1.0	0.0
		Task 2	1.0	0.0
		Task 3	1.0	0.0
	AE Mono	Task 1	0.92	0.15
		Task 2	0.86	0.22
		Task 3	0.97	0.10
	AE Spatial	Task 1	0.92	0.21
		Task 2	0.94	0.13
		Task 3	0.97	0.10
BETA	vOICe	Task 1	1.0	0.0
		Task 2	0.93	0.15
		Task 3	1.0	0.0
	AE Mono	Task 1	0.93	0.15
		Task 2	0.93	0.15
		Task 3	0.96	0.11
	AE Spatial	Task 1	0.96	0.11
		Task 2	1.0	0.0
		Task 3	1.0	0.0

Table 6.2 contains the mean individual completion rates and standard deviations for both groups by system and task. Completion rates were high for all systems, with participants making a selection before the 30 second limit at or above 90 percent of the time with but one exception.[3]

Task completion data for all users and all tasks are reported in Appendix C. Applying a paired samples, two tailed t-test (alpha = 0.05) to the data revealed no statistically significant differences in mean completion rates between AE Mono and AE Spatial within either group.

[3] ALPHA group when working with AE Mono on task 2 made a selection 86% of the time

6.2.2 Completion Times

In table 6.3 I present the task completion times for vOICe, AE Mono, and AE Spatial. These times indicate the mean number of seconds subjects spent on a task before making a selection or, in the rare circumstance, running out of time at the 30 second limit.[4] As with task completion rates, a paired samples two tailed t-test revealed no significant differences in completion times between AE Mono and AE Spatial for both groups (alpha = 0.05).

Table 6.3: Mean Task Completion Times and Standard Deviations for vOICe, AE Mono and AE Spatial

Group	System	Task	Mean	StDev
ALPHA	vOICe	Task 1	10.5	4.5
		Task 2	12.3	5.1
		Task 3	11.8	4.7
		Overall	11.5	4.8
	AE Mono	Task 1	20.7	7.2
		Task 2	18.4	8.3
		Task 3	17.8	6.7
		Overall	19.0	7.5
	AE Spatial	Task 1	19.3	6.5
		Task 2	19.1	7.3
		Task 3	16.9	7.0
		Overall	18.4	7.0
BETA	vOICe	Task 1	15.5	7.8
		Task 2	13.9	7.4
		Task 3	13.4	7.3
		Overall	14.3	7.5
	AE Mono	Task 1	20.5	6.8
		Task 2	22.3	6.5
		Task 3	19.9	6.4
		Overall	20.9	6.6
	AE Spatial	Task 1	19.1	7.6
		Task 2	21.2	5.7
		Task 3	20.5	6.2
		Overall	20.3	6.5

units are in seconds

Even a cursory review of table 6.3 will reveal that participants made selections faster, and more consistently (see table 6.2), with vOICe than with the AuralEyes systems. This difference

[4] Task completion times for all users and tasks are reported in Appendix D

is not surprising considering that vOICe presents an entire soundscape once every second whereas AE relies on the speed of the user's deliberate investigation to determine the rate at which the contents of a scene are transmitted. Obviously this increased 'speed of decision' doesn't necessary correspond with 'accuracy of decision' and it seems fruitless to attempt a comparison of these two methods beyond the conjecture that benefits might be made available by combining aspects of both designs.

6.3 Survey Responses

At the conclusion of each subject's participation they were asked to complete the survey described in chapter 5. This survey collected users' opinions about the relative intuitiveness of the three systems, as well as subjective ratings of auditory fatigue, and overall system preference. All of the subjects completed the survey before viewing their performance results. In general the sample sizes involved in the experiment lacked sufficient power to perform hypothesis rejection on this portion of the data. There are however, meaningful trends suggested by the data. I discuss these trends below.[5]

6.3.1 Intuitiveness

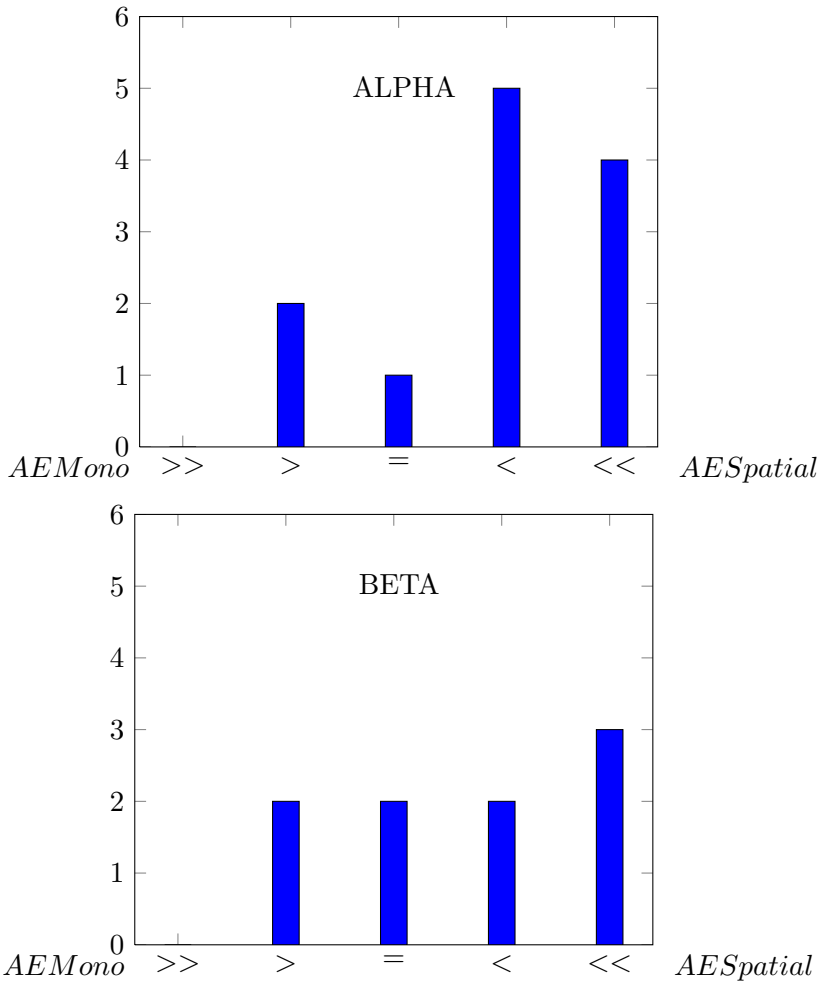
Clearly intuitiveness is an important factor in user adoption of any device. One of my theses in this work is that spatialized audio feedback can provide a reinforced sense of orientation, resulting in a more intuitive user experience. Participants were asked to determine a pairwise comparative intuitiveness rating between each pairing of the three system. Figure 6.2 presents the distributions of the comparative intuitive ratings for AE Mono vs AE Spatial from groups ALPHA and BETA. The labels on the x-axis are representative of the responses available to the participants; namely:[6]

- '>>' AE Mono is much more intuitive than AE Spatial
- '>' AE Mono is somewhat more intuitive than AE Spatial

[5] The raw data compiled from the survey results is available in Appendix F of this document
[6] Note: The actual system names were concealed via pseudonyms during the experiment

- '=' AE Mono and AE Spatial are about equally intuitive
- '<' AE Spatial is somewhat more intuitive than AE Mono
- '<<' AE Spatial is much more intuitive than AE Mono

Figure 6.2: AE Mono vs AE Spatial Comparative Intuitiveness Ratings for Groups ALPHA and BETA

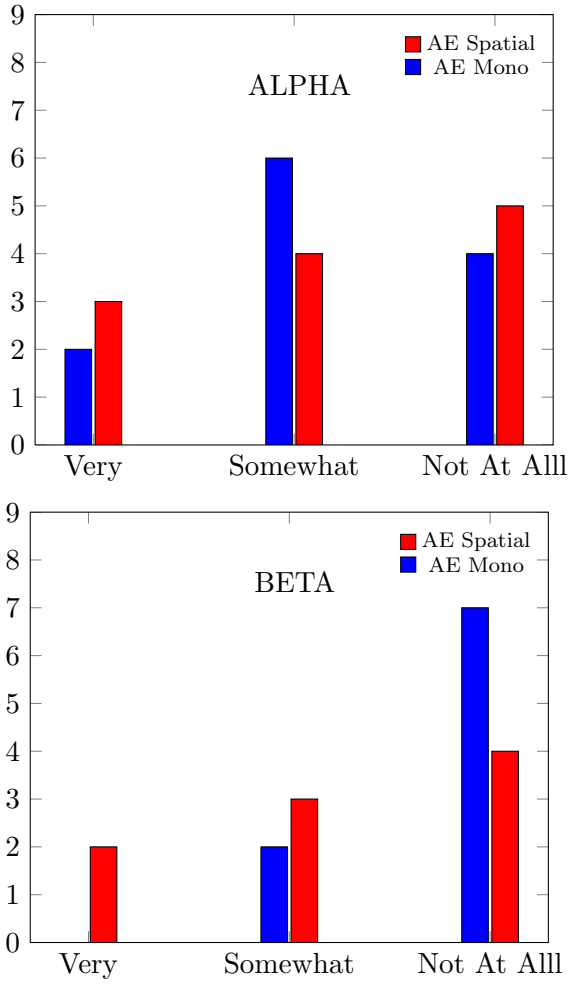


As previously mentioned I am unable to draw any definite conclusions from these data due to insufficient sample size and other problems with the experimental data (as mentioned in chapter 5). Still, a general trend in favor of AE Spatial seems to be present in both groups and may suggest that the intuitiveness of an AuralEyes system is positively affected by spatialization of the audio feedback stream.

6.3.2 Audio Fatigue

Figure 6.3 shows the mean subjective audio fatigue ratings reported by participants. I can detect no discernible and consistent differences between AE Mono and Spatial. These results provide no evidence that audio spatialization has a significant effect on perceived auditory fatigue.

Figure 6.3: Subjective Auditory Fatigue Ratings for ALPHA and BETA Groups

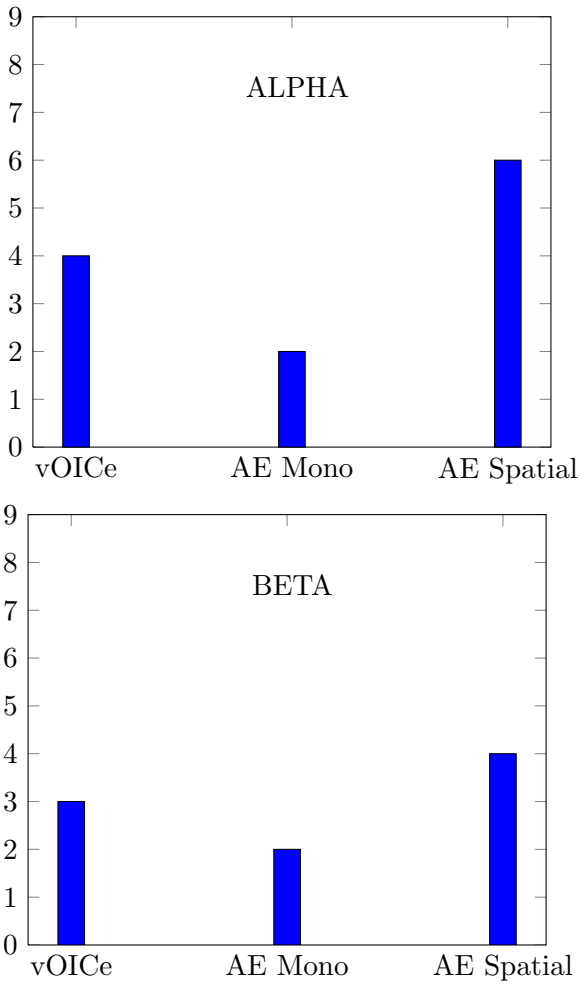


6.3.3 User Preference

The final determination made by each subject was a system preference. All participants were asked which of the three systems they felt they would prefer to use. As mentioned, each participant completed the survey, including this question, before viewing their performance data.

This was done in an attempt to gauge each subjects preference based on their interactions with the user interface independent of a certain knowledge of their actual performance. I theorized that adding spatial qualities to the input audio would have a positive effect on user preference as it would provide another frame of reference from which to evaluate one's interpretation of the input data. In other words, spatial audio provides one more reason for the user to feel comfortable with their decision. The idea was to tease out such possible effects of audio spatialization while also evaluating preference relative to vOICe. After additional consideration and input it became clear that the comparison to vOICe was misguided and served to obscure the comparison between AE Mono and Spatial. Nevertheless, I present the data here with a few observations that I believe will be helpful in future implementations of AuralEyes.

Figure 6.4: User Preference Distribution for ALPHA and BETA Groups



Users appear to prefer AE Spatial more often than AE Mono. Considering only participants who indicated a preference for one of the AuralEyes implementations (2/3 of the subjects in each group), the ratio is 2 to 1 or greater in favor of AE Spatial. The sample size lacks sufficient power to reject the null hypothesis that the two systems are preferred with equal probability; an exact binomial test (one-sided) results in significance values of $p = 0.1445$ and $p = 0.344$ respectively. As with the intuitiveness ratings I presented earlier, these data do not establish conclusively that spatialization has a positive effect on usability - but they do support the suspicion that this is the case.

6.4 Summary

The results presented in this chapter strongly support my first thesis and encourage further validation of my second.

First, the success rates of subjects in both groups (or sub-samples) demonstrated that a user interface that fuses eye tracking with audio feedback can enable a user to extract meaningful information about their environment through directed exploration. Results showed that participants correctly selected target regions (identified by proximity) significantly more often than expected by random selection. These results were significant at the 95% confidence level(see section 6.1).

Secondly, several trends in the survey data lend support to the hypothesis that spatialized audio feedback provides a reinforced sense of orientation, resulting in a more intuitive user experience. Users in both ALPHA and BETA groups tended to consider AE Spatial 'somewhat' or 'much' more intuitive than AE mono, though the differences were not significant at the 95% confidence level. Likewise users tended to express a preference to use AE Spatial more frequently than AE Mono, but again, these results were not significant at the 95% confidence level. These results will guide the design of future experiments in an attempt to more fully validate this hypothesis.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

In this work I have provided evidence that the fusion of eye-tracking and spatial audio can result in ESA designs that are intuitive and provide meaningful benefits to new users. I have shown that users with no experience and very little training are able to quickly learn to perform depth disparity tasks with such designs. This demonstrates the potential benefit of designing an ESA around such a user interface.

In chapter 4 I proposed AuralEyes, a novel user interface for use in electronic sensory aides for the blind. This system utilizes a user's gaze to determine a region or area of interest from which to gather and report data. In this way priority can be given to data identified intentionally by the user as apposed to transmitting all available information, or automatically identified data - much of which may be irrelevant and distracting.

7.1.1 User Study

In chapter 5 I presented the results of a zero-day usability study in which the usability of two AuralEyes based systems (AE Mono and AE Spatial) were evaluated alongside the vOICe Learning Edition[59]. The only difference between the AuralEyes based systems was that the AE Spatial employed spatialized audio while AE Mono did not. Participants were divided into two groups with the order of exposure to the three systems reversed for the second group.

7.1.1.1 Performance

In this study users performed range disparity tasks of varied complexity and contrast. Users achieved higher success rates with the AuralEyes based systems than expected due to random chance with statistical significance at a 95% confidence interval. These results suggest immediately available benefits to new users of such systems.

7.1.1.2 Subjective Usability Ratings

Users tended to identify AE Spatial as more intuitive than AE Mono, supporting the theory that spatialized audio increases intuitiveness. This result was suggestive but not statistically significant at a 95% confidence interval.

Finally, users also tended to indicate that they would prefer to use AE Spatial as apposed to AE Mono, further supporting the theory that spatialized audio increases intuitiveness. Again, this result was suggestive but not statistically significant.

Both of the above mentioned results encourage further experimental validation of the theory that spatial audio can improve the usability of an AuralEyes interface.

7.2 Future Work

7.2.1 AuralEyes Mark-2

In Appendix G I present a fully functional ESA prototype platform intended to enable further investigation of the efficacy of the AuralEyes interface, as well as facilitate the development of more effective audio displays for the blind. Experiments at The College of Idaho, including student driven projects, are already being designed around this system. For instance, funding has already been secured for an electronic wheelchair allowing "blindfolded" participants to be seated while completing a navigational experiment using the AuralEyes Mark-2 system as a sensory aid.

7.2.2 Early Onset Blindness

It is possible that the feedback loop created by the AuralEyes interface could be used to develop ocular motor control in persons with early onset or congenital blindness. By creating an association between muscle movement and auditory feedback individuals may be able to learn to hold and direct their gaze. This would offer both aesthetic and functional advantages to such individuals as it would allow the to direct there "gaze" in a manner comparable to sighted individuals (thus becoming less conspicuous) as well as enable them to use the type of interfaces I am proposing in this work. I have already identified participants for a pilot study to investigate this idea.

7.2.3 Improvements to AuralEyes

Foveal + Peripheral Data Many open questions remain regarding the construction of an audio display that utilizes the AuralEyes interface. The experimental results I presented in this work suggest that a display that presents attenuated scene data in conjunction with prominent region-of-interest data might enable faster response times with the benefit of higher success rates. It is also possible that such a display would be more versatile in diverse situations. This is an area of future work.

Alternate Modalities In this work I have focussed on a single modality - namely sound - for conveying environmental information based on gaze. Interesting questions remain concerning the use of other senses in lieu of, or in conjunction with hearing. For instance, it may be possible/beneficial to transmit lower resolution peripheral information using one mode (such as haptic) and higher detail information from a limited region using another (such as sound). With adequate funding I intend to pursue these questions also.

7.2.4 HRTF Fitting And Spatialization Fatigue

I would have preferred to use individualized HRTFs in this work, but time and resource costs were prohibitive. My efforts to develop an efficient and accessible method of performing HRTF fitting continue. If I am successful, or if I am able to acquire the necessary resources for direct

measurement, then I plan to revisit this question in future experiments.

7.3 The Role of Mobile Computing in Sensory Augmentation

As mobile computers continue to become smaller, more efficient and less expensive, the role they play in our lives will certainly continue to expand. In this work I have investigated the application of computational sensing to retrieve data from the environment, then modify and transfer that information to a user, as guided by the user's behavior. As mobile computational power becomes increasingly available, this model of directed sensory augmentation may be increasingly leveraged. Beyond assistive technologies for the visually impaired, one can easily imagine these same techniques being employed by other groups, such as the deaf community. Indeed, this approach may lend itself to solutions that benefit the general populace at large. For instance, enhanced perception in safety critical situations, or the superimposition of translated text over non-native characters or audio transmission sources within a user's field of view.

Already, widely publicized commercial efforts such as Google's "glass" project are seeking to leverage and develop the fundamental technologies that will enable such possibilities[34]. Advances in user interfaces and virtual display techniques, such as AuralEyes will be a critical component in the development and advancement of this next generation of wearable computing devices.

Bibliography

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

[2] B Ando. Electronic sensory systems for the visually impaired. IEEE Instrumentation Measurement Magazine, 6(2):62–67, 2003.

[3] David Arnim, Benito S. Piuzzi, Chang S. Nam, and Donghun Chung. Guidelines for the development and improvement of universal access systems for blind students. In Proceedings of the 4th international conference on Universal access in human computer interaction: coping with diversity, UAHCI'07, pages 603–612, Berlin, Heidelberg, 2007. Springer-Verlag.

[4] P Bach-y Rita, K A Kaczmarek, M E Tyler, and J Garcia-Lara. Form perception with a 49-point electrotactile stimulus array on the tongue: a technical note. Journal Of Rehabilitation Research And Development, 35(4):427–430, 1998.

[5] Paul Bach-y Rita. Tactile sensory substitution studies. Annals Of The New York Academy Of Sciences, 1013:83–91, 2004.

[6] Armondo B. Barreto and Choudhury H. Maroof. A sonar-based omni directional obstacle detection system designed for blind navigation. WSEAS Transactions on Circuits and Systems, 3(3):480–485, 2004.

[7] Michel Beaudouin-Lafon and William W. Gaver. Eno: synthesizing structured sound spaces. In Proceedings of the 7th annual ACM symposium on User interface software and technology, UIST '94, pages 49–57, New York, NY, USA, 1994. ACM.

[8] Durand R. Begault. 3-D sound for virtual reality and multimedia. Academic Press Professional, Inc., San Diego, CA, USA, 1994.

[9] J M Benjamin. The laser cane. Bulletin Of Prosthetics Research, pages 443–450, 1974.

[10] Bruce B. Blasch, William R. Wiener, and Richard L. Welsh. Use of reflected sound. In Foundations of Orientation and Mobility, pages 145–152, New York, NY, USA, 1997. AFB Press.

[11] Meera M. Blattner, Denise A. Sumikawa, and Robert M. Greenberg. Earcons and icons: their structure and common design principles. Hum.-Comput. Interact., 4(1):11–44, March 1989.

[12] J. Borenstein. The navbelt - a computerized multi-sensor travel aid for active guidance of the blind. In CSUN's Fifth Annual Conference on Technology and Persons with Disabilities, pages 107–116, 1990.

[13] M. Bousbia-Salah, A. Redjati, M. Fezari, and M. Bettayeb. An ultrasonic navigation system for blind people. In Signal Processing and Communications, 2007. ICSPC 2007. IEEE International Conference on, pages 1003 –1006, nov. 2007.

[14] D A Burgess. Real-time audio spatialization with inexpensive hardware. Rehabilitation, (GIT-GVU-92-20), 1992.

[15] David A. Burgess. Techniques for low cost spatial audio. In Proceedings of the 5th annual ACM symposium on User interface software and technology, UIST '92, pages 53–59, New York, NY, USA, 1992. ACM.

[16] Avi Chaudhuri. Fundamentals of Sensory Perception. Oxford University Press, 2011.

[17] M.H. Choudhury, D. Aguerrevere, and A.B. Barreto. A pocket-pc based navigational aid for blind individuals. In Virtual Environments, Human-Computer Interfaces and Measurement Systems, 2004. (VECIMS). 2004 IEEE Symposium on, pages 43 – 48, july 2004.

[18] Brent Cowan and Bill Kapralos. Spatial sound for video games and virtual environments utilizing real-time gpu-based convolution. In Proceedings of the 2008 Conference on Future Play: Research, Play, Share, Future Play '08, pages 166–172, New York, NY, USA, 2008. ACM.

[19] W Crandall, J Brabyn, B L Bentzen, and L Myers. Remote infrared signage evaluation for transit stations and intersections. Journal Of Rehabilitation Research And Development, 36(4):341–355, 1999.

[20] T. Claire Davies, Catherine M. Burns, and Shane D. Pinder. Mobility interfaces for the visually impaired: what's missing? In Proceedings of the 8th ACM SIGCHI New Zealand chapter's international conference on Computer-human interaction: design centered HCI, CHINZ '07, pages 41–47, New York, NY, USA, 2007. ACM.

[21] F.L. Dimmick and E. Gaylord. The dependence of auditory localization upon pitch. Journal of Experimental Psychology, 17(4):593 – 599, 1934.

[22] Kai-Uwe Doerr, Holger Rademacher, Silke Huesgen, and Wolfgang Kubbat. Evaluation of a low-cost 3d sound system for immersive virtual reality training systems. IEEE Transactions on Visualization and Computer Graphics, 13:204–212, March 2007.

[23] W. Keith Edwards, Elizabeth D. Mynatt, and Kathryn Stockton. Access to graphical interfaces for blind users. interactions, 2:54–67, January 1995.

[24] Pololu Robotics & Electronics. Optical range finders. Feb 2012.

[25] Sparkfun Electronics. Triple-axis digital-output gyro itg-3200 breakout. Feb 2012.

[26] F.A. Everest and K.C. Pohlmann. Master Handbook of Acoustics. McGraw-Hill, 2009.

[27] A. Faro, D. Giordano, C. Spampinato, D. De Tommaso, and S. Ullo. An interactive interface for remote administration of clinical tests based on eye tracking. In Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications, ETRA '10, pages 69–72, New York, NY, USA, 2010. ACM.

[28] R M Fish. An audio display for the blind. IEEE Transactions on Biomedical Engineering, 23(2):144–154, 1976.

[29] Miguel Angel Garcia-Ruiz and Jorge Rafael Gutierrez-Pulido. An overview of auditory display to assist comprehension of molecular information. Interact. Comput., 18:853–868, July 2006.

[30] William W. Gaver. Auditory icons: using sound in computer interfaces. Hum.-Comput. Interact., 2:167–177, June 1986.

[31] William W. Gaver. The sonicfinder: an interface that uses auditory icons. Hum.-Comput. Interact., 4:67–94, March 1989.

[32] Lalya Gaye. A flexible 3d sound system for interactive applications. In CHI '02 extended abstracts on Human factors in computing systems, CHI EA '02, pages 840–841, New York, NY, USA, 2002. ACM.

[33] GloPos. Glopos. February 2012.

[34] Google. Glass. March 2013.

[35] Point Grey. Cameras: Stereo vision products. Feb 2012.

[36] Matti Gröhn, Tapio Lokki, and Tapio Takala. Comparison of auditory, visual, and audiovisual navigation in a 3d space. ACM Trans. Appl. Percept., 2:564–570, October 2005.

[37] Zoltan Haraszy, Sebastian Micut, Virgil Tiponut, and Titus Slavici. Multi-subject head related transfer function generation using artificial neural networks. In Proceedings of the 14th WSEAS international conference on Systems: part of the 14th WSEAS CSCC multiconference - Volume II, ICS'10, pages 399–404, Stevens Point, Wisconsin, USA, 2010. World Scientific and Engineering Academy and Society (WSEAS).

[38] Wilko Heuten, Daniel Wichmann, and Susanne Boll. Interactive 3d sonification for the exploration of city maps. In Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles, NordiCHI '06, pages 155–164, New York, NY, USA, 2006. ACM.

[39] F. Wai-ling Ho-Ching, Jennifer Mankoff, and James A. Landay. Can you see what i hear?: the design and evaluation of a peripheral sound display for the deaf. In Proceedings of the SIGCHI conference on Human factors in computing systems, CHI '03, pages 161–168, New York, NY, USA, 2003. ACM.

[40] ASUSTeK Computer Inc. Xtion pro live: The world's first pc exclusive motion sensing development solution. December 2012.

[41] Maxbotics Inc. Ultrasonic sensors by maxbotix. Feb 2012.

[42] Dean P. Inman, Ken Loge, and John Leavens. Vr education and rehabilitation. Commun. ACM, 40:53–58, August 1997.

[43] Yukio Iwaya. Individualization of head-related transfer functions with tournament-style listening test: Listening with other's ears. Acoustical Science And Technology, 27(6):340–343, 2006.

[44] B. Kapralos, M. R. Jenkin, and E. Milios. Virtual audio systems. Presence: Teleoper. Virtual Environ., 17:527–549, December 2008.

[45] Dean Inman Ken, Ken Loge, and Aaron Cram. Teaching orientation and mobility skills to blind children using computer generated 3-d sound environments. In Proc. ICAD 2000, pages 1–5, 2000.

[46] Thomas Koelewijn, Adelbert Bronkhorst, and Jan Theeuwes. Auditory and visual capture during focused visual attention. Journal of Experimental Psychology: Human Perception and Performance, 35(5):1303–1315, 2009.

[47] Ron Kupers and Maurice Ptito. "seeing" through the tongue: cross-modal plasticity in the congenitally blind. International Congress Series, 1270(0):79 – 84, 2004. ¡ce:title¿Frontiers in Human Brain Topology. Proceedings of ISBET 2004¡/ce:title¿.

[48] GT Sonification Lab. Swan: System for wearable audio navigation. January 2012.

[49] Yunjae Lee, Youngjin Park, and Youn-Sik Park. Newly designed hrtf measuring system. System, pages 1781–1784, 2009.

[50] Dongheng Li, Jason Babcock, and Derrick J. Parkhurst. openeyes: a low-cost head-mounted eye-tracking solution. In Proceedings of the 2006 symposium on Eye tracking research & applications, ETRA '06, pages 95–100, New York, NY, USA, 2006. ACM.

[51] Tapio Lokki and Matti Grohn. Navigation with auditory cues in a virtual environment. IEEE MultiMedia, 12:80–86, April 2005.

[52] Jack M. Loomis, Reginald G. Golledge, and Roberta L. Klatzky. Navigation system for the blind: Auditory display modes and guidance. Presence: Teleoper. Virtual Environ., 7:193–203, April 1998.

[53] W Loughborough. Talking lights. Journal of Visual Impairment and Blindness, 73:243, 1979.

[54] Roberto Manduchi and James Coughlan. (computer) vision without sight. Commun. ACM, 55(1):96–104, January 2012.

[55] J.R. Marston. Towards an accessible city: empirical measurement and modeling of access to urban opportunities for those with vision impairments using remote infrared audible signage. University of California, Santa Barbara, 2002.

[56] Tara Matthews, Janette Fong, and Jennifer Mankoff. Visualizing non-speech sounds for the deaf. In Proceedings of the 7th international ACM SIGACCESS conference on Computers and accessibility, Assets '05, pages 52–59, New York, NY, USA, 2005. ACM.

[57] David K. McGookin and Stephen A. Brewster. Multivis: improving access to visualisations for visually impaired people. In CHI '06 extended abstracts on Human factors in computing systems, CHI EA '06, pages 267–270, New York, NY, USA, 2006. ACM.

[58] P.B.L. Meijer. An experimental system for auditory image representations. Biomedical Engineering, IEEE Transactions on, 11(2):112 –121, feb. 1992.

[59] P.B.L. Meijer. The voice learning edition - synthetic vision software for the blind. December 2012.

[60] Lotfi B Merabet and Alvaro Pascual-Leone. Neural reorganization following sensory loss: the opportunity of change. Nature Reviews Neuroscience, 11(1):44–52, 2010.

[61] Microsoft. Kinect for windows. Feb 2012.

[62] Ross Miller. Kinect for xbox 360 review. November 2010.

[63] A W Mills. On the minimum audible angle. Journal of the Acoustical Society of America, 30(4):237–246, 1958.

[64] OpenKinect. Openkinect. Feb 2012.

[65] World Health Organization. Visual impairment and blindness. April 2011.

[66] Philip E. Pedley and Robert S. Harper. Pitch and the vertical localization of sound. The American Journal of Psychology, 73:447–449, September 1959.

[67] Phil D. Picton and Michael D. Capp. Relaying scene information to the blind via sound using cartoon depth maps. Image Vision Comput., 26:570–577, April 2008.

[68] Shane D. Pinder and T. Claire Davies. Exploring direct downconversion of ultrasound for human echolocation. In Proceedings of the 8th ACM SIGCHI New Zealand chapter's international conference on Computer-human interaction: design centered HCI, CHINZ '07, pages 49–54, New York, NY, USA, 2007. ACM.

[69] C.C. Pratt. The spatial character of high and low tones. Journal of Experimental Psychology, 13(3):278 – 285, 1930.

[70] Maurice Ptito, Solvej M Moesgaard, Albert Gjedde, and Ron Kupers. Cross-modal plasticity revealed by electrotactile stimulation of the tongue in the congenitally blind. Brain: A journal of neurology, 128(Pt 3):606–614, 2005.

[71] Rameshsharma Ramloll, Wai Yu, Stephen Brewster, Beate Riedel, Mike Burton, and Gisela Dimigen. Constructing sonified haptic line graphs for the blind student: first steps. In Proceedings of the fourth international ACM conference on Assistive technologies, Assets '00, pages 17–25, New York, NY, USA, 2000. ACM.

[72] Christophe Ramstein, Odile Martial, Aude Dufresne, Michel Carignan, Patrick Chassé, and Philippe Mabilleau. Touching and hearing gui's: design issues for the pc-access system. In Proceedings of the second annual ACM conference on Assistive technologies, Assets '96, pages 2–9, New York, NY, USA, 1996. ACM.

[73] Timothy E. Roden, Ian Parberry, and David Ducrest. Toward mobile entertainment: A paradigm for narrative-based audio only games. Sci. Comput. Program., 67:76–90, June 2007.

[74] S K Roffler and R A Butler. Localization of tonal stimuli in the vertical plane. Journal of the Acoustical Society of America, 43(6):1260–1266, 1968.

[75] Daniel Sanabria, Salvador Soto-Faraco, Jason S Chan, and Charles Spence. When does visual perceptual grouping affect multisensory integration? Cognitive, Affective, And Behavioral Neuroscience, 4(2):218–229, 2004.

[76] Bernhard U. Seeber and Hugo Fastl. Subjective selection of non-individual head-related transfer functions. Proceedings of the 2003 International Conference on Auditory Display, Boston, July 2003.

[77] S Shoval, J Borenstein, and Y Koren. Mobile robot obstacle avoidance in a computerized travel aid for the blind. Proceedings of the 1994 IEEE International Conference on Robotics and Automation, 8:2023–2028, 1994.

[78] Steven W Smith. The Scientist and Engineers Guide to Digital Signal Processing, volume 3. California Technical Publishing, 1997.

[79] Sound Foresight Technology Limited, 21L Evans Business Centre, Marson Business Park, Tockwith, York, Y026 7QF. UltraCane User Guide, 2011.

[80] Martin Talbot and William Cowan. On the audio representation of distance for blind users. In Proceedings of the 27th international conference on Human factors in computing systems, CHI '09, pages 1839–1848, New York, NY, USA, 2009. ACM.

[81] Terrie Terlau and William M. Penrod. 'K' Sonar Curriculum Handbook. American Printing House for the Blind Inc., 1839 Frankfort Avenue, Louisville, Kentucky 40206-0085, 2008.

[82] Lore Thaler, Stephen R. Arnott, and Melvyn A. Goodale. Neural correlates of natural human echolocation in early and late blind echolocation experts. PLoS ONE, 6(5):e20162, 05 2011.

[83] Toshiba. Semiconductors & components. Feb 2012.

[84] TranSafety. Study compares older and younger pedestrian walking speeds. Road Engineering Journal, October 1997.

[85] Bijal Trivedi. Sensory hijack: rewiring brains to see with sound. NewScientist, 2773, 2010.

[86] Ramiro Velzquez, Flavien Maingreaud, and Edwige E. Pissaloux. Intelligent glasses: A new man-machine interface concept integrating computer vision and human tactile perception. In in Proceedings of EuroHaptics 2003, pages 456–460, 2003.

[87] Ramiro Velzquez. Wearable assistive devices for the blind. In Aim Lay-Ekuakille and Subhas Chandra Mukhopadhyay, editors, Wearable and Autonomous Biomedical Devices and Systems for Smart Environment, volume 75 of Lecture Notes in Electrical Engineering, pages 331–349. Springer Berlin Heidelberg, 2010. 10.1007/978-3-642-15687-8_17.

[88] Bruce N Walker and Jeffrey Lindsay. Navigation performance with a virtual auditory display: effects of beacon sound, capture radius, and practice. Human Factors, 48(2):265–278, 2006.

[89] Hua Wang, Mark Chignell, and Mitsuru Ishizuka. Empathic tutoring software agents using real-time eye tracking. In Proceedings of the 2006 symposium on Eye tracking research & applications, ETRA '06, pages 73–78, New York, NY, USA, 2006. ACM.

[90] D Waters and H Abulula. The virtual bat: echolocation in virtual reality, volume 2001, pages 191–196.

[91] Dean A. Waters and Husam H. Abulula. Using bat-modelled sonar as a navigational tool in virtual environments. Int. J. Hum.-Comput. Stud., 65:873–886, October 2007.

[92] Gareth R. White, Geraldine Fitzpatrick, and Graham McAllister. Toward accessible 3d virtual environments for the blind and visually impaired. In Proceedings of the 3rd international conference on Digital Interactive Media in Entertainment and Arts, DIMEA '08, pages 134–141, New York, NY, USA, 2008. ACM.

[93] Wikipedia. Kinect. February 2012.

[94] Wikipedia. Wearable computer. January 2012.

[95] J. Wilson, B.N. Walker, J. Lindsay, C. Cambias, and F. Dellaert. Swan: System for wearable audio navigation. In Wearable Computers, 2007 11th IEEE International Symposium on, pages 91 –98, oct. 2007.

[96] Fredrik Winberg and John Bowers. Assembling the senses: towards the design of cooperative interfaces for visually impaired users. In Proceedings of the 2004 ACM conference on Computer supported cooperative work, CSCW '04, pages 332–341, New York, NY, USA, 2004. ACM.

[97] Song Xu, Zhizhong Li, and Gaviriel Salvendy. Individualization of head-related transfer function for three-dimensional virtual auditory display: a review. In Proceedings of the 2nd international conference on Virtual reality, ICVR'07, pages 397–407, Berlin, Heidelberg, 2007. Springer-Verlag.

[98] Wai Yu and Stephen Brewster. Multimodal virtual reality versus printed medium in visualization for blind people. In Proceedings of the fifth international ACM conference on Assistive technologies, Assets '02, pages 57–64, New York, NY, USA, 2002. ACM.

[99] Dmitry N. Zotkin, Jane Hwang, Ramani Duraiswami, and Larry S. Davis. Hrtf personalization using anthropometric measurements. In Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, page 157160. IEEE Computer Society, 2003.

Appendix A

Protocol Documents

The following research was conducted with the approval, and under the oversight, of the Institutional Review Board at The College of Idaho in Caldwell Idaho. The IRB approval number for this study is 1097.1.

Please note that page numbers have been removed and appendix designations modified from the original document.

IRB rev 04Nov11

TITLE: *Comparing the "Zero Day" Usability of Two Audio Displays for the Blind*

Version/Date: *1.1 / 09-10-12.*

PRINCIPAL INVESTIGATOR: Frank Jones

Name: Frank Jones
Address: 2115 Ison Court #101
Telephone: 208-459-5320
E-mail address: fjones@collegeofidaho.edu

ADDITIONAL KEY PERSONNEL

None

OBJECTIVES

The primary purpose of this study is to investigate and compare the performance and usability of two audio displays for the blind among users who are employing the technologies for the first time. Though they will not be actively recruited for this study, volunteers who have used one of the technologies being evaluated in this work may provide important pilot information for subsequent investigations – a secondary objective of this work – and thus will not be precluded from participation. The impact of spatial audio on user performance and usability in the context of the AuralEyes interface will also be investigated.

Advances in microelectronics engineering and manufacturing have made high quality sensors and powerful mobile computing resources both affordable and widely available. These advances enable the construction of relatively low-cost assistive devices capable of sensing the environment, processing the acquired data and then presenting it to the user in an application appropriate manner. Broadly speaking such devices can be considered electronic sensory aids (ESAs).

This work is part of ongoing investigations pertaining to the efficacy of a new user interface paradigm for ESAs, called AuralEyes. AuralEyes fuses eye tracking with spatial audio in a novel user interface paradigm that enables a user to directly control the behavior of an ESA at runtime. This level of control enables users to perform active, intentional investigations of their environment – exerting real-time control over the sensory input obtained and reported by the ESA.

This work investigates 4 hypotheses:

1) *A user interface that fuses eye/gaze tracking with spatial audio can enable a user to extract meaningful spatial information about their environment through directed exploration.*
2) *Spatialized, gaze-directed audio feedback provides a reinforced sense of orientation resulting in a more intuitive user experience relative to gaze-directed audio feedback alone.*
3) *The combination of the AuralEyes user interface with a timing based encoding of distance information is more accessible (easier to learn) for new users than the volume and frequency based audio code employed in the vOICe SSD.*

4) *A user interface that enables a user to quickly (i.e. with little training) investigate their environment through intentional, directed exploration results in enhanced usability in the short term when compared with a scene-based automated interface.* Prolonged training times have been cited as a primary reason that ESAs fail to be adopted by the blind community.[Davies] Enabling users to quickly perceive benefit from the use of an ESA may be key to the adoption and long-term use of such devices.

BACKGROUND

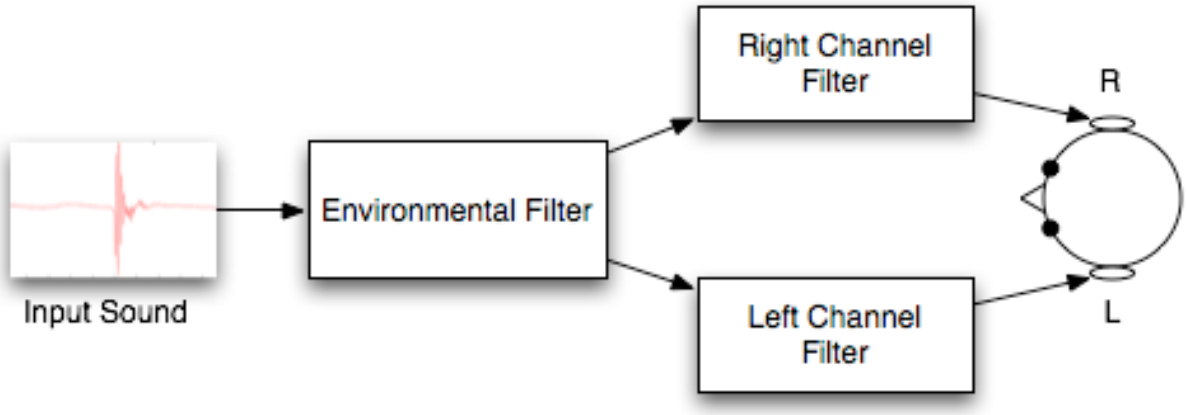
At least as early as 1945 researchers have sought to utilize electronic audio devices to communicate spatial and environmental information to the blind [Benjamin]. Despite significant research and development efforts over the past sixty years the number of audio based electronic sensory aids (ESAs) actively utilized by the blind community remains small. This section presents a cursory overview of spatial audio, its potential and drawbacks, as well as examples of alternative techniques employed in ESAs. The section closes with a description of the AuralEyes interface and its potential benefits.

Spatial Audio

Spatial audio has been shown to be an effective medium for conveying environmental and contextual information to the blind. Spatial audio is the equivalent of 3-D vision in the auditory domain. By mimicking the naturally occurring temporal and spectral differences between signals arriving at the right and left ears, the perception of location in 3-dimensional space can be synthetically added to an artificial audio signal. Spatial audio has greater capabilities than traditional stereo or “surround sound” which convey a limited sense of space. The listener of spatialized audio can identify the azimuth and elevation that a sound is emanating from in a manner similar to naturally occurring sounds (i.e. the guitarist is located five degrees up and twenty degrees to the left of my current position). Spatial audio has numerous applications in video games and other forms of multimedia and entertainment as well as safety critical environments.

Rendering of spatial audio requires the use of filters which when convolved with the original audio signal insert the appropriate effects to create the perception of space for the listener. These filters are known as Head Related Impulse Responses (HRIRs) or Head Related Transfer Functions (HRTFs) (see Figure 1).

Figure 1: Audio Spatialization Process



HRTFs vary from one position to another and across different individuals, meaning that a large number of filters specific to each listener are required for the highest quality spatial audio. The problem is simplified somewhat by the fact that beyond 1 meter the filtering effects of the upper body and outer ear become approximately proportional to distance, and therefore a single HRTF measured at > 1m from the listener can be reasonably used to approximate all locations at a specific azimuth and elevation beyond 1m. The requirement that these measurements be individualized for each listener in order to achieve optimal results remains. [Begault, Kapralos]

Measuring of HRTFs can be done in a variety of ways. A common approach is to seat a subject in the center of an echo-free environment placing high-quality microphones at the opening of each ear canal. A sound source is then moved through space around the subject playing specially designed signals. Recording these signals at each ear and comparing them to the original allows the filtering effects of the subject's shoulders, head and outer ear to be extracted as an impulse response. An impulse response describes the effects of a given system on a pure impulse that has been passed through the system. A pure impulse contains all frequencies at a common amplitude, so the information contained in an impulse response describes the effect of the system on all frequencies. Thus the impulse responses measured in the aforementioned fashion (also known as HRIRs) are sufficient filters for the creation of spatial audio. Variations of this technique exist, but the basic concepts described here are generally applicable.

Studies have shown that using their own HRTFs individuals are capable of determining the "virtual location" of a synthetic sound source with a similar degree of accuracy to localizing actual sound sources in the "real world"[Begault]. Unfortunately processes such as the one described above require expensive facilities and equipment, and several hours of sitting in place for each subject. In short, direct measurement of HRTFs is expensive and time consuming.

An alternative to directly measured HRTFs is to utilize a "generic" set of HRTFs. Such sets may be created by extracting mean characteristics from a large database of individuals' HRTFs, or by measuring the impulse responses generated by a "model human" that accurately embodies the average physical characteristics of a population. The KEMAR manikin is such a model and

HRTF measurements collected for KEMAR manikins using both large and small pinnae (outer ear) have been made publicly available. [Algazi]

The drawback to utilizing generic HRTFs is that localization accuracy suffers - particularly when estimating elevation. This fact limits the application of spatial audio to domains that do not require high levels of localization accuracy, or to unique settings where individual HRTFs can be acquired. This may explain why the numerous ESAs which utilize spatial audio have failed to escape the boundaries of academia in significant numbers.

Alternatives to Spatial Audio Displays

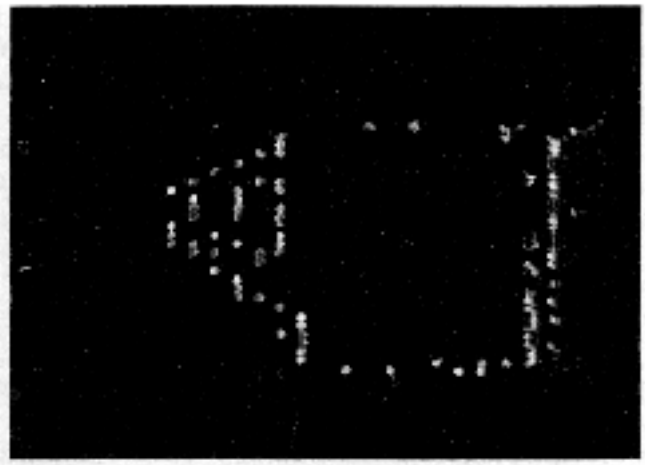
Creating a sensory aid that maps spatial information to audio feedback with a high degree of precision requires the ability of the user to accurately and consistently map the audio information being presented to the appropriate orientation in space. Due at least in part to the difficulties associated with spatial audio, alternative "audio codes" have been proposed that do not rely on HRTF filtering to convey orientation information. Such techniques typically map characteristics of the environment to audio features in the feedback signal, such as pitch and volume.

Audio Codes

A pioneer of such audio codes was Raymond Fish. In 1975 Fish utilized the reported association between pitch and elevation[Pedley, Pratt, Roffler] in his work developing one of the first audio displays for the blind[Fish]. Fish demonstrated two variants of his code. The first approach maps the bright regions of a scene to sound pulses. Scanning horizontally and then vertically (row-wise), bright regions are indicated by deliberately selected tones. The frequency of a tone is determined by the current elevation in the scan, with higher elevations being mapped to higher frequencies. Stereo interaural level difference (ILD)(i.e. volume panning) is applied to the sound to encode horizontal position within the scene. Reported frame rates for this approach are on the order of seconds per frame.

The second variant of Fish's code is similar to the first but seeks to communicate edges rather

Figure 2: Edge Detection of a Cup [Fish]



than bright regions by producing a tone when a light-to-dark or dark-to- light transition occurs. Edges are detected both vertically and horizontally. Figure 2 demonstrates the kind of data

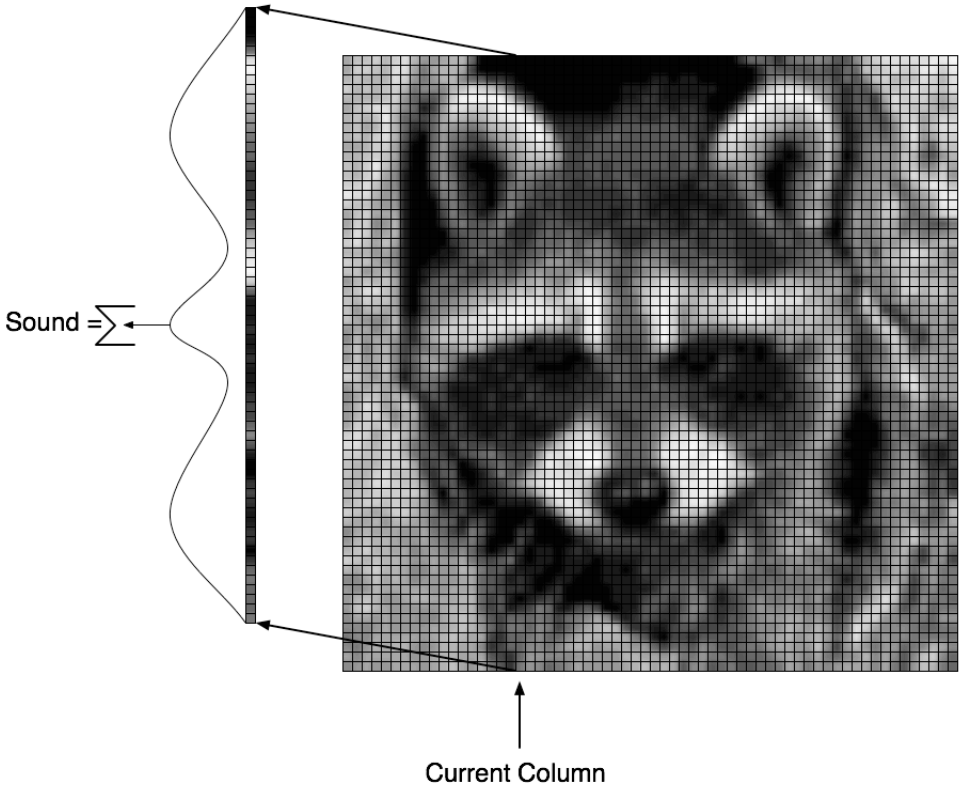
resulting from this process when applied to an image of a coffee cup. Tone pulses would be generated corresponding to the bright "dots" in this image. This approach was found to be effective in conveying scene information while often allowing the system to transmit frames more quickly than the first variant.

Using the technology of the time (namely photoelectric cells, oscilloscopes, tone generators, TV cameras, etc.) and a lot of ingenuity, Fish constructed four different prototype systems. These systems (termed systems I - IV) successfully demonstrated the potential of his audio code. Users of these systems were able to perform tasks such as identify simple and complex shapes, describe shapes they had not been exposed to previously, and (using one of the camera based systems) successfully navigate high-contrast obstacles in a room.

Though the bulkiness of the technology involved, and the operating frame rates of the systems were not suitable for everyday use; Fish's work established important foundational principles for transmitting visual information through an audio display.[Fish]

Building upon the work of Fish and others, in 1992 Peter B. L. Meijer proposed a low-cost portable system built around the idea of conveying images through an audio code with similarities to that proposed by Fish[Meijer]. Like Fish, Meijer leveraged the psychoacoustic nature of pitch to represent elevation. However, instead of performing a 2-dimensional pixel by pixel raster scan of an effectively black and white image Meijer's system combines the intensity values of an entire column of a grayscale image into a single sound comprised of multiple frequencies - somewhat like a musical chord comprised of multiple notes. The volume of each of

Figure 3: Illustration of the vOICe Audio Code



the constituent frequencies is determined by the intensity value of the pixel it is associated with. Thus no sound is emitted for a completely black pixel and the maximum volume is employed for an all white pixel. An image is auralized by playing the combined frequencies for each column in sequence repetitively scanning from left to right(see Figure 3).

To emphasis the boundary between successive scans of a scene a ”synchronization click” is produced between frames.

In general Meijer’s approach allows a higher frame rate (on the order of 1 Hz) than those reported in Fish’s work, especially in the case of complex images. Meijer’s system does not utilize ILD cues to enforce the current horizontal position of the scanline. Instead the synchronization click and constant scan rate are utilized to indicate horizontal scanning position.

The basic ideas of Meijer’s original paper have resulted in the vOICe system, an affordable, portable augmented reality system for the blind. Though still considered under development by Meijer, consistent users of the vOICe system are reporting fascinating results. Beyond simply learning to identify objects and an increased sense of their surroundings, some users are reporting a limited restoration of the *perception* of sight[Merabet, Trivedi].

By providing vOICe with a depth map (acquired via stereo cameras or some other means) encoded as a grayscale image, Meijer's system can be utilized as an electronic travel aid akin to the AuralEyes device utilized in this work (described below). This is the arrangement that will be evaluated in this study.

The success of the vOICe system demonstrates the viability of audio as a sensory substitute for sight. However, the most exciting implications of the system are only manifested after training and regular use of the device for a significant period of time. This is a challenge because anecdotal evidence suggests that many consider the audio signals generated by the device oppressive; consequently user fatigue is likely very high. In addition, the tones generated by vOICe sound very unnatural and can occupy a wide spectrum, raising concerns of overloading the auditory sense or at the very least cluttering the existing audio environment.

Though recent efforts have been made to improve upon the basic idea of vOICe with audio signals that are more flexible (i.e. color can be represented) and less oppressive (the sounds are somewhat more musical in nature) , the basic drawbacks of the approach remain [Tzedek].

User Interface

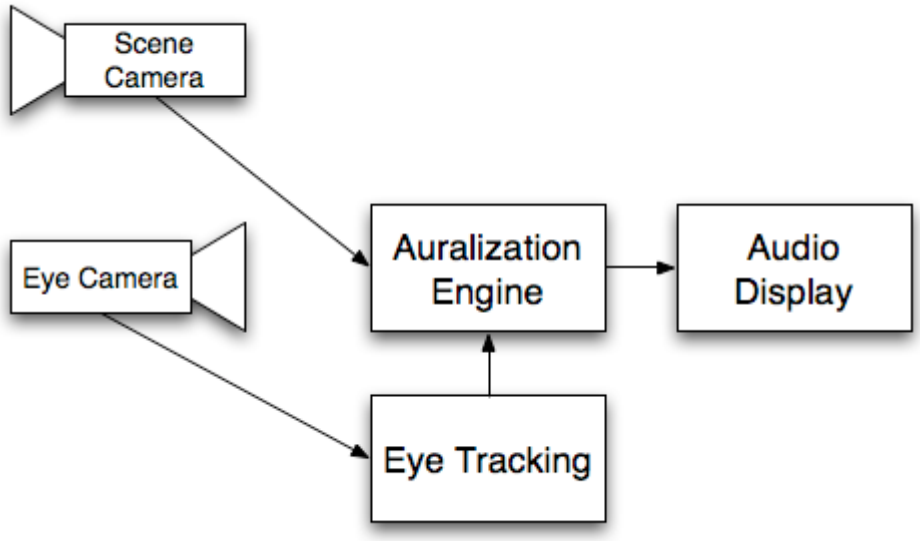
Another major challenge to the adoption of ESAs has been the ineffective nature of the interfaces such devices present to the user. Most glaring is the lack of a hands-free method of affecting the information presented to the user. Such deficiencies in assistive devices limit the usability of otherwise promising technologies, and likely affect both long and short-term adoption rates.

AuralEyes

This work is part of an ongoing effort to investigate the efficacy of AuralEyes, a novel gaze-driven spatial audio interface for assistive devices. This interface is intended to address the aforementioned challenges to ESA design and adoption. The basic concept of the AuralEyes interface is to employ a users gaze as a sort of "joystick" to direct an ESA's behavior. Principles of spatial audio are used to obtain and/or reinforce the orientation of data presented to the user in accordance with their gaze. AuralEyes does not represent a specific device, but rather an interface design that may be utilized in any number of specific implementations.

Potential Benefits of the AuralEyes Interface

Figure 4: A Possible Implementation of AuralEyes



A Better "Fit"

Allowing a user to indicate a region of interest with greater precision than simple head orientation makes possible the investigation of a large "scene" of data via a focused and reduced stream of data. When viewing a complex image humans do not achieve full comprehension "at a glance" by processing all aspects of the environment within the current field of view equally. Rather, the viewer's foveal gaze (approximately 2 degrees wide) is directed from one region of interest to another as a mental image of the scene is created from the high-resolution data acquired therewith. Throughout this process lower resolution periphery vision provides cues of possibly interesting regions.

It seems reasonable to assume that SSDs intended to mimic the visual perception of one's environment through audition, would benefit from a similar model of operation, as it would more closely mimic the sense being substituted. This work investigates this possibility.

Generalizable Spatialized Audio

The coupling of a user's deliberate gaze orientation with directional audio feedback may enhance their perception of orientation with respect to the part of the scene being investigated. It is hoped that this enhancement will be sufficient to override any “ambiguities” introduced by the use of generic HRTFs, removing the need to gather individualized HRTFs from a user.

More Appropriate Audio Codes

Because spatialization is used to encode orientation information in the audio signal, the audio attributes of timing and pitch are available for alternative uses in the audio code. Timing information has been shown to be more effective at conveying range information than volume and pitch. AuralEyes based ESAs that communicate range are able to take advantage of this more natural mapping.

System Design

Figure 4 presents a possible implementation of an AuralEyes equipped system. A depth camera captures spatial information about the environment. This information is fed into an auralization engine along with gaze data from an eye-tracking camera. These data are used to generate an audio display containing spatial information for the user. In this work such a system is modeled by simulating the scene-camera with pre-recorded “still-frame” depth maps. The other elements of the system are fully implemented in the experiment.

Numerous audio codes can be employed in the AuralEyes system. In this work the auralization engine maps range data to timing information in the form of a pulse train. The range to the nearest object within a target area in the depth image (determined by the user's gaze) is represented by the frequency of the pulse train. Small delays between pulses (high frequency) indicate close proximity, longer delays (lower frequency) indicate greater range. The auralization engine in this experiment operates in two modes. In the first mode the pulse train is delivered, unaltered to the audio display (presented through headphones). In the second mode the auralization engine spatializes the audio feedback to give the impression that the pulse train is emanating from the direction of the user’s gaze. Both of these modes will be investigated in this study to better understand the effects of spatial audio on usability and user preference in this study.

STUDY DESIGN

30 – 45 participants will be recruited to participate in this study. The data collection portion of the study consists of an approximately 35 minute long usability study consisting of three sets of task repeated three times, and a questionnaire.

Zero Day Usability Evaluation

This part of the experiment evaluates a user’s ability to perform localization and identification tasks on depth maps using an AuralEyes device (in two modes) and the learning edition of the vOICe system. The purpose of this part of the study is to understand how quickly new users are able to adjust to and utilize the two different approaches, and garner subjective feedback and impressions.

Volunteers are asked how much (if any) time they have spent using an electronic sensory aid, and specifically how much time spent with the vOICe system. Though this study is primarily focussed on new users, volunteers who are familiar with vOICe are not excluded from participation – though there data will be analyzed separately if a sufficient number of participants fall into this category. This determination is made based on two factors: First, it is doubtful than many volunteers will fall into this category and thus there participation will not significantly effect the principle data collection needs of the project. Second, users of vOICe or other ESAs are among the population of individuals who may benefit from this work and their participation in limited numbers may provide useful pilot information for subsequent studies. See the attached questionnaire for more information on this subject.

During the experiment participants are seated in a chair at a table with their head steadied by an adjustable chin rest. Initially A PC monitor is positioned a fixed distance in front of the subject's face. The monitor is used to calibrate the eye-tracker utilized by the AuralEyes system.

Once the eye-tracker is calibrated a "blinder" is placed between the user and the computer monitor. This blinder consists of a nondescript, white, tri-fold foamboard positioned approximately twelve inches away from the participants face. The purpose of the blinder is to eliminate visual reference points for sighted participants (i.e. simulate blindness).

Throughout the experiment a head mounted eye-tracking camera and "ear bud" headphones are appropriately affixed to the subject's head and ears. For consistency subjects wear the headphones and eye-tracker whenever tasks are being performed, whether with AuralEyes or vOICe. As the experiment progresses the performance of the eye-tracker is evaluated and is recalibrated as necessary to ensure consistent accuracy.

Each participant performs three iterations of tasks with both the AuralEyes and vOICe systems. Half of the subjects utilize the vOICe system first, the other half begin with AuralEyes. AuralEyes is operated in two modes, the first with audio spatialization disabled (gaze directed mono audio only), the second with audio spatialization enable. The subjects are therefore partitioned into two sub-pools that perform three iterations of tasks as follows:

	15 Subjects	15 Subjects
Iteration 1	vOICe	AuralEyes Mode 2
Iteration 2	AuralEyes Mode 1	AuralEyes Mode 1
Iteration 3	AuralEyes Mode 2	vOICe

A short orientation is provided for each system immediately before it is first used. This orientation includes a short technical description of how the given system works, followed by an up to 5 minute "experimentation" period in which the subject is able to listen to the audio signals

generated from a training scene that is described to the user. The same training image and description are used for both systems. The image and its description are provided in Appendix AA of this document.

Throughout the study the labels “System A” , “System B”, and “System C” will be used to identify the three system configurations. The vOICe system will be referred to as “System A”, AuralEyes Mode 1 as “System B” and AuralEyes Mode 2 as “System C” when communicating with the subject.

Following the orientation period for each system the participant is given a series of 3 sets of tasks to complete. For each of these tasks the subject must specify one of nine regions as the target region. The regions are organized in a 3 X 3 tic-tac-toe-like grid. The columns are identified as left, center and right. Rows are top, middle and bottom. A specific region is identified by column and row (i.e. “top-left”). Participants are instructed to enter their selection via a numeric keypad as well as call it out verbally. If the participants verbal response does not agree with the data recorded by the number pad the verbal response is assumed to be correct. In this manner consistent and unbiased timing information can be recorded as well as data integrity ensured even if users unaccustomed to the number pad enter the wrong value.

The task sets are as follows:

1) Identify the closest region. (Single “near” region, 3 scenes – high, medium, and low contrast)

2) Identify the region furthest away. (Single “distant” region, 3 scenes – high, medium, and low contrast)

3) Identify the closest region. (3 near regions, 3 scenes – contrast between “near” regions: high, medium and low; contrast between near regions and background medium)

Scenes are randomly selected for each iteration to ensure that the subject cannot apply prior knowledge to the tasks.

Data collected for the tasks include: failure/success to select a region, the region selected by the subject, whether or not the correct region is selected, and time to completion (if applicable).

The purpose of this study is not to train subjects, but rather to evaluate the level of difficulty associated with initial use of these two systems. Subjects will be encouraged to do their best on each task but will also be informed that for each task they can “give up” and move on if they are simply unable to perform the task.

Subjects are allotted up to thirty seconds per scene for each of the three tasks. Completing the three sets of tasks for a single system is anticipated to take up to approximately ten minutes including time to orient/evaluate/calibrate the eye-tracker as needed. Up to a five-minute break is available to the subject when switching between systems.

This portion of the experiment is anticipated to last approximately 35 minutes.

Questionnaire

Upon completing the tasks outlined above each participant will be asked to complete a short, one-page questionnaire about their experience with the three systems. The questions on this form are designed to ascertain a subject's impressions and reactions to the systems under test. The questionnaire is attached to this protocol as Appendix AB.

Completing the questionnaire should talk less than 5 minutes.

The entire experiment is anticipated to last around 45 minutes.

Data Analysis

The success rates, and task completion times (as applicable for each task) for each system will be compared using a paired T-test. The data collected from the questionnaire will be examined in a similar manner.

ABOUT THE SUBJECTS

The inclusion criteria for this study are that:

1) The volunteer is capable of giving informed consent.
2) The volunteer is between the ages of 18 and 34 years of age (inclusive). These age criteria are designed to ensure that the participant capable of giving informed consent, and to limit the age range of the subject pool to minimize variations in hearing performance introduced by the (possibly unnoticed) effects of age related hearing loss.
3) The volunteer self-reports having normal hearing.
4) The volunteer has (self-reported) normal range of eye motion and can intentionally direct and hold eye position. (i.e. the volunteer can intentionally "stare" at something)

VULNERABLE POPULATIONS

No vulnerable populations are anticipated in this study.

RECRUITMENT METHODS

Participants will be recruited from the College of Idaho including the psychology department undergraduate pool. Advertisements seeking participants for the study will be posted on bulletin boards around the campus where permission from the building proctors can be obtained. In the printed materials seeking participants (See Appendix AC for Flyer), a short description of the study and how participants can volunteer is given.

CONSENT PROCESS

Participants will be given consent forms (submitted with this protocol) at the beginning of the study. Participants will be given ample time to read the form and will be given the opportunity to ask questions before the forms are collected.

PROCESS TO DOCUMENT CONSENT IN WRITING

Consent in writing will be provided to all participants before beginning the study. The consent form is being submitted with this protocol.

PROCEDURES

Each participant will be taken through the following procedures. No audio or video recordings will be made at any point in the experiment. All of these procedures are to be carried out in room B-10 in Boone hall on the C of I campus.

1) The participant will read and sign the attached consent form. The consent form describes the experiment and the tasks to be completed. Any questions the participant has will be answered before continuing.
2) The participant is assigned a random subject ID and completes the attached participant questionnaire.
3) The participant is seated at a table and a pair of "earbud" headphones placed in their ears.
4) A head mounted Eye-Guide Assist is placed on the participants head and adjusted for proper operation.
5) The participant is asked to place there chin on a chin-rest and look forward at the computer monitor while the eye-tracker is calibrated. This process requires looking at a shrinking dot that appears at several locations on the screen. Once the calibration is complete the monitor is replaced with the virtual blindfold described above.

6) The following procedures are repeated three times, once for each system:(approx. 35 minutes anticipated)
    a. A five-minute or less orientation for System A/B/C is presented.
    b. The participant is asked to perform the 3 sets of tasks described in the Study Design section above. This portion of the experiment lasts around 10 minutes.
    c. The participant is given a 5 minute break.
7) The headphones and eye-tracker are removed.
8) The participant is asked to complete a short survey (see Appendix AB for survey).
9) The participant is debriefed (see Appendix AD for debriefing document).

The expected total time commitment for each subject is anticipated to be around 45 minutes.

SPECIMEN MANAGEMENT

We are not collecting biological specimens.

IRB rev 04Nov11

DATA MANAGEMENT

The hardcopy questionnaires, surveys and consent forms will be stored in a locked cabinet in the office of Professor Frank Jones.

All data generated from subject responses to the in-person questionnaires will be stored on a password-protected server on the College of Idaho campus. Second and third party subject information will not be asked of the subject.

All of the individual data collected from participants will be kept completely confidential in a database that is stored on a password-protected server located on the College of Idaho campus and will not be shared with any other people. Any and all accesses to the collected data will only be used for statistical analyses and reporting for this research. Only the principle investigator, Professor Frank Jones, will possess confidential knowledge of participant's identification via the assignment of a random Subject ID number. The subject ID number is used only for data organization and analysis purposes and is not linked to any identifiable information (i.e. there is no key linking the ID to a consent form etc.).

The data generated during the evaluation phases of the experiment do not contain any personally identifiable information; measurement and performance data are recorded according to the random subject ID which is not linked to any identifiable information. These data are initially recorded on a password-protected laptop and then transferred to a secure server at the completion of each experiment. The stored data will be destroyed after a period of three years subsequent to the completion of the study.

WITHDRAWAL OF PARTICIPANTS

Subjects who qualify for the experiment will not be withdrawn without their consent. Subjects who are disruptive for any reason will be asked politely to focus on the tasks at hand. In all studies subjects may withdraw from the experiment at any time. Data from any subject who withdraws from the experiment will be thrown out. If early withdrawals cause the sample size to drop bellow 30 participants then additional participants will be recruited. Subjects who withdraw will not be contacted about this study again.

RISKS TO PARTICIPANTS

During the study audio signals are played through in-ear headphones. Excessive sound volume could be uncomfortable or (in extreme cases) dangerous. To remove this risk, the volume levels for both the headphones will be calibrated at the beginning of the experiment and inline hardware volume limiters will be used to ensure that sound levels never exceed 85 db (the threshold for dangerous noise levels [7]) – even if the participant requests increased volume.

Use of the virtual blindfold shutters described above could be uncomfortable/unnerving for some participants - especially those who suffer from claustrophobia. To remove this risk participants

will be given the option to proceed with or without the shutters and they will be removed at the participants request at any point in the experiment.

POTENTIAL BENEFITS TO PARTICIPANTS

There is no direct benefit to the subject.

PROVISIONS TO MONITOR THE DATA FOR THE SAFETY OF PARTICIPANTS

Performance metrics and questionnaires will be collected for each participant. There will be no identifiable information associated with the data. The data will be protected on a server with a secure password. Completed consent forms and questionnaires will be stored in a locked cabinet in the office of Professor Frank Jones. See the Data Management section above for further details.

PROVISIONS TO PROTECT THE PRIVACY INTERESTS OF PARTICIPANTS

This study will have very little impact on a participant's privacy interests. There will be no identifiable information associated with the data collected. The data will be protected on a server with a secure password. Completed consent forms will be stored in a locked cabinet in the office of Professor Frank Jones. See the Data Management section above for further details.

MEDICAL CARE AND COMPENSATION FOR INJURY

This study involves no risk of injury – subjects are simply sitting in a chair and moving their eyes while listening to sounds at safe volume levels. The activities are equivalent to the everyday activity of sitting and listening to music with headphones (albeit while staring out a dark window). There are no plans for medical care or compensation for injury.

COST TO PARTICIPANTS

There will not be any direct cost to participants, since we will schedule our meeting with them (to sign the Consent form and perform the experiment) on the College of Idaho campus and at a time of their convenience.

SHARING OF RESULTS WITH PARTICIPANTS

Participants will be fully debriefed. Please see the additional submitted forms in Appendix AD.

References

V.R. Algazi, R.O. Duda, D.M. Thompson, and C. Avendano. The cipic hrtf database. In Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the, pages 99 –102, 2001.

Durand R. Begault. 3-D sound for virtual reality and multimedia. Academic Press Professional, Inc., San Diego, CA, USA, 1994.

J M Benjamin. The laser cane. Bulletin Of Prosthetics Research, pages 443–450, 1974.

T. Claire Davies, Catherine M. Burns, and Shane D. Pinder. Mobility interfaces for the visually impaired: what's missing? In Proceedings of the 8th ACM SIGCHI New Zealand chapter's international conference on Computer-human interaction: design centered HCI, CHINZ '07, pages 41-47, 2007

R M Fish. An audio display for the blind. IEEE Transactions on Biomedical Engineering, 23(2):144–154, 1976.

Yukio Iwaya. Individualization of head-related transfer functions with tournament-style listening test: Listening with other's ears. Acoustical Science And Technology, 27(6):340–343, 2006.

B. Kapralos, M. R. Jenkin, and E. Milios. Virtual audio systems. Presence: Teleoper. Virtual Environ., 17:527–549, December 2008.

Yunjae Lee, Youngjin Park, and Youn-Sik Park. Newly designed hrtf measuring system. System, pages 1781–1784, 2009.

P.B.L. Meijer. An experimental system for auditory image representations. Biomedical Engineering, IEEE Transactions on, 11(2):112 –121, feb. 1992.

Lotfi B Merabet and Alvaro Pascual-Leone. Neural reorganization following sensory loss: the opportunity of change. Nature Reviews Neuroscience, 11(1):44– 52, 2010.

NIH. National Institute on Deafness and Other Communication Disorders, How Loud is Too Loud? Bookmark, National Institutes of Health, http://www.nidcd.nih.gov/health/hearing/pages/ruler.aspx , May 2009

Philip E. Pedley and Robert S. Harper. Pitch and the vertical localization of sound. The American Journal of Psychology, 73:447–449, September 1959.

C.C. Pratt. The spatial character of high and low tones. Journal of Experimental Psychology, 13(3):278 – 285, 1930.

S K Roffler and R A Butler. Localization of tonal stimuli in the vertical plane. Journal of the Acoustical Society of America, 43(6):1260–1266, 1968.

Bernhard U. Seeber and Hugo Fastl. Subjective selection of non-individual head- related transfer functions. Proceedings of the 2003 International Conference on Auditory Display, Boston, July 2003.

Bijal Trivedi. Sensory hijack: rewiring brains to see with sound. NewScientist, 2773, 2010.

Levy Tzedek, s. Abboud, S. Maidenbaum and A. Amedi. Fast, Accurate Reaching Movements with a Visual-to Audiotory Sensory Substitution Device. Neuroscience, 30:4, July 2012

IRB rev 04Nov11

Appendix AA: Training Image and Description

During an orientation phase audio signals are generated based upon this hypothetical scene in which a rectangular region at eye level is situated in close proximity to the participant, the surrounding area is more distant.

IRB rev 04Nov11

Appendix AB: Post Study Survey

Subject ID #_____

Key:
System A: Frequency + Volume; Scans Left to Right; No Eye Tracking; No Spatial Audio
System B: Pulse Train (Tick); Eye-Tracking; No Spatial Audio (Sound Does Not Move)
System C: Pulse Train (Tick); Eye-Tracking; Spatial Audio (Sound Moves With Eyes)

Place an X in the box beneath the statement you most closely agree with (one for each row).

Intuitiveness (i.e. natural to use)

System A vs. System B:

System **A** is Much More Intuitive Than System **B**	System **A** is Somewhat More Intuitive Than System **B**	System **A** and System **B** Are About Equally Intuitive	System **B** is Somewhat More Intuitive Than System **A**	System **B** is Much More Intuitive Than System **A**

System A vs. System C:

System **A** is Much More Intuitive Than System **C**	System **A** is Somewhat More Intuitive Than System **C**	System **A** and System **C** Are About Equally Intuitive	System **C** is Somewhat More Intuitive Than System **A**	System **C** is Much More Intuitive Than System **A**

System B vs. System C:

System **B** is Much More Intuitive Than System **C**	System **B** is Somewhat More Intuitive Than System **C**	System **B** and System **C** Are About Equally Intuitive	System **C** is Somewhat More Intuitive Than System **B**	System **C** is Much More Intuitive Than System **B**

Audio Fatigue (i.e. tiring to listen too)

System A

The Sounds Generated By System **A** Are **Very Fatiguing**	The Sounds Generated By System **A** Are **Somewhat Fatiguing**	The Sounds Generated By System **A** Are **Not Fatiguing**

System B

The Sounds Generated By System **B** Are **Very Fatiguing**	The Sounds Generated By System **B** Are **Somewhat Fatiguing**	The Sounds Generated By System **B** Are **Not Fatiguing**

System C

The Sounds Generated By System **C** Are **Very Fatiguing**	The Sounds Generated By System **C** Are **Somewhat Fatiguing**	The Sounds Generated By System **C** Are **Not Fatiguing**

Please Circle One Response

Which System Would You Prefer To Use?

System A / System B / System C

Additional Comments: _____

_____

_____

_____

Appendix AC: Recruitment Flyer

ASSISTIVE DEVICE USABILITY STUDY!

Researchers at the College of Idaho want to compare the effectiveness of multiple user interfaces employed in assistive devices for the blind. This study seeks to clarify how design decisions in user interfaces for sensory substitution devices (i.e. communicate sight through sound) affect a user's ability to quickly adjust to a device and perform useful tasks, and it's overal usability.

Research is always voluntary!

Would this study be a good fit for me?

This study might be a good fit for you if:

- ☐ You are 18 – 34 years old
- ☐ You have normal hearing
- ☐ You have normal range of motion with your eyes

What would happen if I took part in the study?

If you decide to take part in the research study, you would:

- ☐ Perform simple localization tasks using headphones and a wearable eye tracker
- ☐ Complete a short questionairre at the beginning of the study and a short survey at the end
- ☐ Take between 45 and 60 minutes to complete your participation

NOTE: If you chose to participate in the study and you wear glasses or contact lenses you will be asked to remove them during the study as they may interfere with the eye-tracker

To take part in the assistive device user study or for more information, please contact Frank Jones at 208-459-5320, or at fjones@collegeofidaho.edu.

The principal researcher for this study is Frank Jones at The College of Idaho at Caldwell.

Appendix AD: Debriefing Document

The purpose of this experiment is to evaluate the performance of a new interface technique for electronic assistive devices (ESAs), and to compare this technique to an existing approach. Many ESAs have been proposed but few are released commercially and virtually none have seen widespread adoption and use. One of the major factors that individuals report prevents them from utilizing an ESA is the amount of training time required to effectively use such devices. Anecdotally, it appears that the user interfaces of many existing ESAs make them difficult to use and adjust to. Improving our understanding of what kinds of interfaces/technologies are more natural/easy to use will contribute to our ability to serve individuals who can benefit from assistive technologies.

The proposed interface technique ties eye-tracking to spatial audio (3-d sound) feedback, enabling the user to direct the "attention" of an ESA, and perceive the direction the feedback is coming from (i.e. the sounds appear to come from the desired location). This is in contrast to many systems which automatically present the user with information the system deems valuable - with no control given to the user. This work investigates the hypothesis that giving the user such control results in a device that is more natural and intuitive to learn and use.

An additional problem with ESAs is that the sounds generated are often very unnatural sounding and generate fatigue in the user. The user interface investigated in this work allows for more flexible audio "codes", enabling the design of possibly more "pleasant" and/or natural sounds. This study investigates two possible audio codes and compares their performance to an existing approach, namely the frequency and volume based approach used by the vOICe sensory substitution device. It is expected that the codes proposed in this study will prove to be more effective at communicating depth, and less tiring to listen to.

30 - 45 participants, ranging in age from 18 to 34 are scheduled to participate. All individuals will perform the same tasks but with differences in the order in which the ESA models are employed.

The data we collect will be analyzed to determine the average user performance in completing the assigned tasks using each system. User feedback will also be analyzed in order to understand user preferences and impressions. It is expected that the interface proposed in this study will show an increase in performance and user experience for new users relative to the vOICe system.

Thank you for the time you spent participating in this experiment and for trying hard to do the tasks.

If you have any questions or concerns, please contact the experimenter, Frank Jones (fjones@collegeofidaho.edu, (208) 459-5320). If you would like a summary of the results when research is completed, please leave your name and address with the experimenter. Thank you for participating.

IRB rev 04Nov11

If you have concerns about the experiment, you may also contact the College of Idaho Institutional Review Board (IRB) responsible for regulating research involving human participants (information below).

Meredith Minear, PhD
Assistant Professor of Psychology
Dept. of Psychology,
College of Idaho
2112 Cleveland Blvd. Caldwell, ID 83605
Phone: (208)-459-5171
Email: mminear@collegeofidaho.edu

IRB rev 04Nov11

Appendix AE: Participant Questionnaire

Subject ID #_____

Thank you for you willingness to participate in this study. The purpose of this questionnaire is to obtain some background information about you.

Age: _____

Gender: _____

Rate Your Hearing: Normal Hearing / Mild Hearing Loss / Moderate Hearing Loss / Severe hearing Loss / Profound Hearing Loss

Rate The Range of Motion of Your Eyes: Normal Range of Motion / Slightly Impaired Range of Motion / Severely Impaired Range of Motion

Rate Your Ability to Direct Your Gaze:

Normal Ability	Somewhat Limited Ability	Severely Limited Ability
able to stare at a fixed point or follow a moving target	some difficulty staring at a fixed point or following a moving target	unable to stare at a fixed at a fixed point or follow a moving target

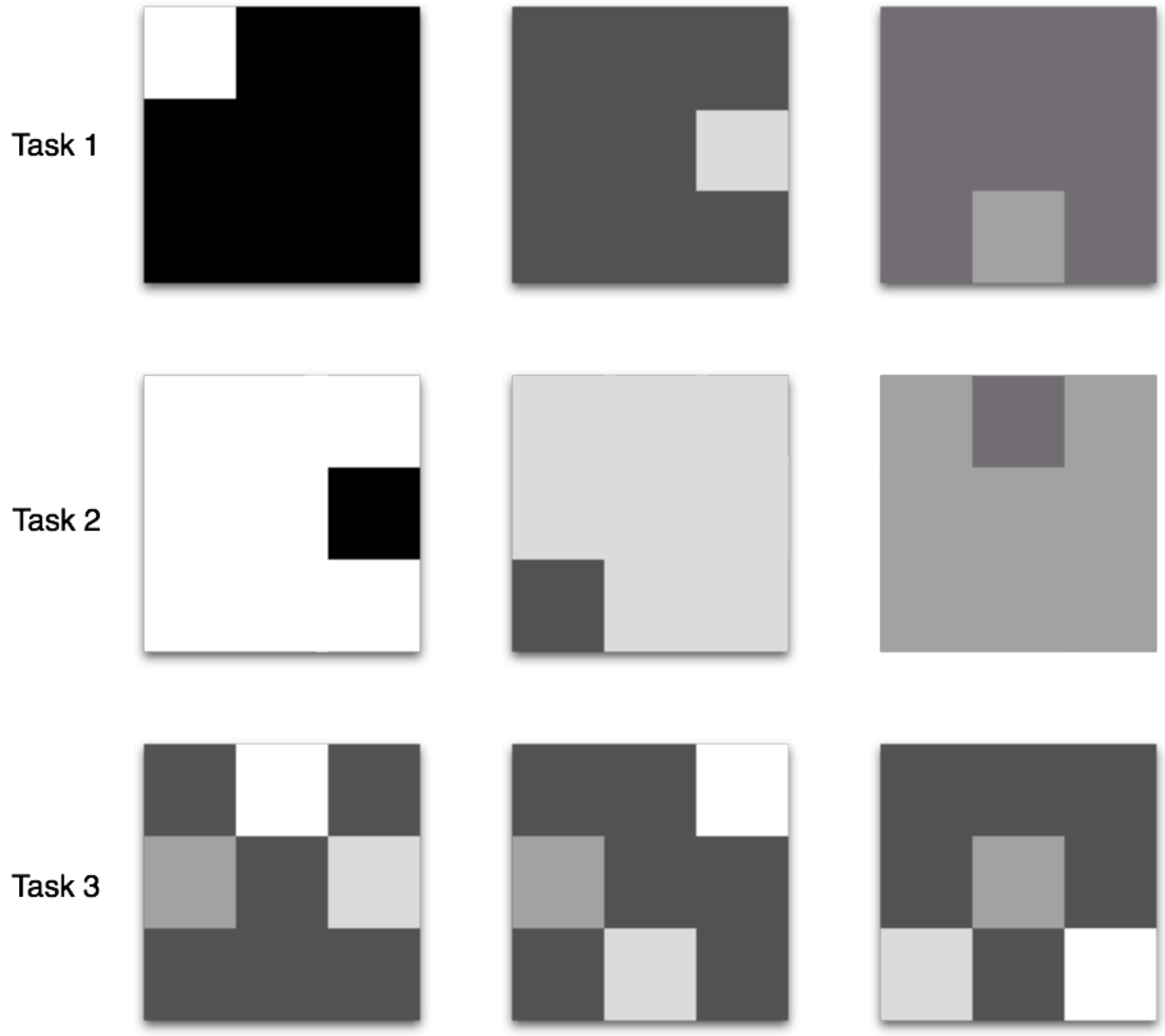
Have you ever used a sensory substitution device? Y / N
(i.e. a device that maps one sense to another - sight to sound for example?)

<u>If yes</u> please name (if possible) or describe the device: _____
_____

Approximately how much time did you spend using the device? _____
(i.e. 5 minutes, 1 hour, "I use it all the time, etc.)

Appendix AF: Example Stimuli

Example scenes for tasks 1 - 3. Bright regions represent surfaces that are near, darker regions further away. These scenes are obviously scaled to fit on the page. Each task includes scenes/scene elements with high, medium and low contrast - which equates to surfaces that have large, medium, or minimal differences in proximity.



Appendix B

Orientation Documents

System A

“ [System A], is a new ... approach towards seeing with sound. It is not based on sonar or echolocation, but uses real … input from a [digital sensor]. ... Through special software this lets blind people ... hear live views from their environment through their stereo headphones, thus hearing the very same shapes and things that their sighted friends see with their eyes. The software translates images from the [sensor] on-the-fly into closely corresponding sounds. For instance, [if the sensor is a depth or distance sensor, a nearby point or pixel in the scene] gives a short beep. If this [point] is on your left, you will hear the beep on your left side, and if it is on your right, you will hear it on your right side. If the [point] moves up, you will hear the pitch of the beep move up, and if the [point] moves down, so does its pitch. With two [points] you get two beeps, with three [points] three beeps, and so on. A [nearby] horizontal line yields a long tone, because the [points] that make up this line give a corresponding concatenation of beeps in time, sounding as a pure tone. Again, if the whole line moves up or down, so does its pitch. A vertical line is a stack of [points], sounding all at the same time but all with different pitches since they are at different heights. Together this sounds like a brief noise burst. ...

The above [depth]-image to sound mapping allows for sounding any visual scene, but the more complex the view, the more complex the sound will be. It takes about a second to sound the entire content of a view, and every second the sound will be "refreshed" to reflect any changes in the [spatial] content of scenery as captured by the [sensor]. These one-second sounds that contain the whole view are called "soundscapes", and they sound the [spatial] content via a left to right scan with pitch indicating elevation and loudness indicating [nearness]...

Mapping principles: It is essential that you … first obtain a thorough understanding of the principles of [system A's] sound mapping, consisting of three simple rules, each rule dealing with one fundamental aspect of ([near and far]) vision: rule 1 concerns *left* and *right*, rule 2 concerns *up* and *down*, and rule 3 concerns *[near]* and *[far]*. The actual rules are

Left and *Right*. Video is sounded in a left to right scanning order, by default at a rate of one image snapshot per second. You will hear the stereo sound pan from left to right correspondingly. Hearing some sound on your left or right thus means having a corresponding [depth] pattern on your left or right, respectively.

Up and *Down*. During every scan, pitch means elevation: the higher the pitch, the higher the position of the [depth] pattern. Consequently, if the pitch goes up or down, you have a rising or falling [depth] pattern, respectively.

Near and *Far*. Loudness means [nearness]: the louder the [closer]. Consequently, silence means [out of range], and a loud sound means [close up], and anything in between is [scaled accordingly].

In other words, [system A] scans every [depth-map] from left to right, while associating height with pitch and brightness with nearness.

Active Demo:

Systems B & C

Systems B and C introduce a new user interface designed for use (among other things) in assistive devices for the blind. The system gathers information from the environment using one or more digital sensors (cameras, depth sensors etc.) and converts that information into sound. Both systems use eye tracking to determine the user's area of interest within the scene data being gathered by the sensor(s). Much like a computer mouse can be used to move a pointer to a desired region of an image on a computer screen, the user's eyes determine the portion of the scene that the device should report information about. Specifically, information about the distance to the nearest surface or object within the region of interest is encoded as a series of audio "pulses". The pulses indicate range according to the time between pulses - the more time, the greater the distance. Thus, objects or regions that are far away will generate a slow series of pulses (long pauses between pulses) and regions or objects that are near will result in more rapid pulses (shorter pauses between pulses). Fast pulses = near; slow pulses = far away. In the event that the nearest surface within a region of interest is too far away to be detected by the sensor then the pulses will slow down to one every four seconds - indicating simply that the nearest obstacle is "far away" and out of range.

By moving their eyes up, down, left and right, users can investigate a scene, somewhat like scanning a dark area with a very narrow flashlight beam or running one's finger over an unseen surface in order to understand its features.

System C differs from System B by also incorporating principles of spatial audio (3-D sound). In addition to generating the necessary audio pulses, System C performs special filtering of the sounds to create the impression that they are coming from the direction of the region of interest. In other words the sounds generated "follow" the user's eye. For instance, if the user looks up and to the left, not only will the audio pulses be modified to reflect that part of the scene but the audio will "sound" as though it is coming from a region up and to the left relative to the user.

Active Demo:

Appendix C

Task Completion Data

Table C.1: Participant Task Completion Performing Three Range Disparity Tasks With System A at Three Levels of Contrast

Group	Gender	System A								
		1H	1M	1L	2H	2M	2L	3H	3M	3L
ALPHA	F	1	1	1	1	1	1	1	1	1
	F	1	1	1	1	1	1	1	1	1
	F	1	1	1	1	1	1	1	1	1
	F	1	1	1	1	1	1	1	1	1
	F	1	1	1	1	1	1	1	1	1
	F	1	1	1	1	1	1	1	1	1
	F	1	1	1	1	1	1	1	1	1
	M	1	1	1	1	1	1	1	1	1
	M	1	1	1	1	1	1	1	1	1
	M	1	1	1	1	1	1	1	1	1
	M	1	1	1	1	1	1	1	1	1
	M	1	1	1	1	1	1	1	1	1
	Average	1.0			1.0			1.0		
BETA	F	1	1	1	1	1	1	1	1	1
	F	1	1	1	1	0	1	1	1	1
	F	1	1	1	1	1	1	1	1	1
	F	1	1	1	1	1	1	1	1	1
	F	1	1	1	1	0	1	1	1	1
	F	1	1	1	1	1	1	1	1	1
	M	1	1	1	1	1	1	1	1	1
	M	1	1	1	1	1	1	1	1	1
	M	1	1	1	1	1	1	1	1	1
	Average	1.0			0.93			1.0		

1 = Successfully Made A Selection, 0 = Failed To Make A Selection

Table C.2: Participant Task Completion Performing Three Range Disparity Tasks With System B at Three Levels of Contrast

Group	Gender	System B								
		1H	1M	1L	2H	2M	2L	3H	3M	3L
ALPHA	F	1	1	1	0	1	1	1	1	1
	F	1	1	1	1	1	0	1	1	1
	F	1	1	1	1	1	1	1	1	1
	F	1	1	1	1	1	1	1	1	1
	F	0	1	1	1	1	0	1	1	1
	F	1	1	1	1	1	1	1	1	1
	F	1	1	1	1	1	1	1	1	1
	M	1	1	0	1	0	0	1	1	1
	M	1	1	0	1	1	0	1	0	1
	M	1	1	1	1	1	1	1	1	1
	M	1	1	1	1	1	1	1	1	1
	M	1	1	1	1	1	1	1	1	1
	Average	0.92			0.83			0.94		
BETA	F	1	1	1	1	1	1	1	1	1
	F	1	1	0	1	1	0	1	1	1
	F	1	1	1	0	1	1	1	1	1
	F	1	1	1	1	1	1	1	1	1
	F	1	1	1	1	1	1	1	1	1
	F	1	1	0	1	1	1	1	1	1
	M	1	1	1	1	1	1	1	1	0
	M	1	1	1	1	1	1	1	1	1
	M	1	1	1	1	1	1	1	1	1
	Average	0.93			0.93			0.96		

1 = Successfully Made A Selection, 0 = Failed To Make A Selection

Table C.3: Participant Task Completion Performing Three Range Disparity Tasks With System C at Three Levels of Contrast

Group	Gender	System C								
		1H	1M	1L	2H	2M	2L	3H	3M	3L
ALPHA	F	0	1	1	1	1	1	1	0	1
	F	1	1	1	1	1	1	1	1	1
	F	1	1	1	1	1	1	1	1	1
	F	1	1	1	1	1	1	1	1	1
	F	1	1	1	1	1	1	1	1	1
	F	1	1	1	1	1	1	1	1	1
	F	1	1	1	1	1	1	1	1	1
	M	1	1	1	1	1	1	1	1	1
	M	0	1	0	1	1	0	1	1	1
	M	1	1	1	1	1	0	1	1	1
	M	1	1	1	1	1	1	1	1	1
	M	1	1	1	1	1	1	1	1	1
	Average	0.92			0.94			0.97		
BETA	F	1	1	1	1	1	1	1	1	1
	F	1	1	1	1	1	1	1	1	1
	F	1	1	1	1	1	1	1	1	1
	F	1	1	1	1	1	1	1	1	1
	F	1	1	1	1	1	1	1	1	1
	F	1	1	1	1	1	1	1	1	1
	M	1	1	0	1	1	1	1	1	1
	M	1	1	1	1	1	1	1	1	1
	M	1	1	1	1	1	1	1	1	1
	Average	0.96			1.0			1.0		

1 = Successfully Made A Selection, 0 = Failed To Make A Selection

Appendix D

Task Completion Times

Table D.1: Participant Task Completion Times For Three Range Disparity Tasks With System A at Three Levels of Contrast

Group	Gender	System A								
		1H	1M	1L	2H	2M	2L	3H	3M	3L
ALPHA	F	16.8	10.5	12.4	6.3	11.2	7.2	12.8	11.5	16.6
	F	12.4	10.2	11.5	8.6	16.1	14.2	8.5	17.5	17.5
	F	11.8	8.5	14.2	7.4	9.3	7.9	12.1	14.5	7.1
	F	13.5	22.7	10.6	8.6	18.9	22.2	14.0	11.1	11.8
	F	6.9	5.0	4.9	9.5	6.3	12.9	4.6	5.0	8.7
	F	4.8	5.1	3.6	3.5	8.8	10.2	7.5	11.4	10.7
	F	7.5	9.8	13.5	10.1	21.9	14.8	12.7	11.5	8.9
	M	9.4	7.0	7.6	7.0	8.8	8.8	9.4	9.0	7.7
	M	7.8	11.3	19.4	25.3	21.9	16.7	30.0	19.4	8.2
	M	5.0	10.8	11.5	12.1	15.4	13.0	18.1	8.3	9.0
	M	16.8	7.8	11.6	14.0	9.9	10.6	12.2	10.2	14.2
	M	9.2	7.8	19.0	15.9	13.0	14.8	8.1	13.8	11.8
	Average	10.50			12.32			11.82		
BETA	F	10.3	15.6	9.2	5.1	6.8	10.0	8.8	7.4	5.6
	F	11.3	9.3	7.8	7.9	6.3	10.5	12.5	8.2	5.7
	F	23.0	28.2	29.4	29.0	30.0	16.4	28.8	27.9	29.4
	F	11.4	19.9	19.2	12.7	11.0	18.2	23.3	12.2	11.6
	F	29.0	16.0	13.7	12.7	12.2	15.0	17.6	9.5	9.9
	F	8.9	7.1	12.2	4.1	11.1	10.1	6.5	9.1	7.2
	M	24.4	25.2	30.0	20.6	30.0	21.2	19.6	23.4	17.4
	M	8.1	14.0	11.7	4.0	15.8	11.9	10.5	9.2	10.1
	M	7.1	8.4	9.5	11.8	20.0	10.0	9.9	10.5	8.8
	Average	15.55			13.87			13.37		

Time is in seconds

Table D.2: Participant Task Completion Times For Three Range Disparity Tasks With System B at Three Levels of Contrast

Group	Gender	System B								
		1H	1M	1L	2H	2M	2L	3H	3M	3L
	F	19.2	7.1	7.9	30.0	11.7	30.0	17.2	26.9	30.0
	F	27.3	17.5	24.0	17.9	28.7	30.0	20.0	22.4	22.9
	F	24.0	15.2	19.0	20.3	13.4	11.1	11.6	7.8	10.3
	F	24.3	29.9	28.3	23.2	18.0	30.0	25.4	24.8	17.8
A	F	30.0	13.9	30.0	7.3	16.7	30.0	6.1	11.6	17.8
L	F	21.9	28.1	15.5	9.2	10.6	23.0	16.7	14.0	18.8
P	F	19.4	27.9	29.0	11.6	14.0	21.0	8.8	10.6	9.8
H	M	10.9	22.0	30.0	7.2	30.0	30.0	11.5	13.4	13.5
A	M	15.3	17.6	30.0	22.9	2.6	0.0	24.6	30.0	26.1
	M	12.1	10.2	15.9	21.6	19.9	17.0	15.7	9.1	18.9
	M	11.6	13.9	29.3	11.2	16.8	21.2	20.6	18.4	12.8
	M	25.5	15.4	25.4	23.6	16.8	14.4	22.6	24.6	26.8
	Average	20.67			18.41			17.77		
	F	24.7	20.7	18.6	25.2	22.3	16.4	28.1	16.9	18.6
	F	30.0	16.9	20.0	15.8	14.2	27.8	16.1	12.2	17.7
	F	26.2	24.3	23.3	29.1	13.0	18.5	25.2	15.0	25.9
B	F	20.7	15.0	30.0	24.7	22.1	21.8	15.8	5.0	23.3
E	F	26.0	20.8	27.8	21.6	29.2	25.0	25.1	20.3	30.0
T	F	9.4	18.2	9.0	14.0	26.2	14.7	11.7	13.0	13.6
A	M	17.5	21.9	30.0	30.0	29.1	30.0	21.5	14.3	30.0
	M	10.5	8.7	17.0	30.0	16.7	20.2	25.2	25.5	16.6
	M	10.9	25.3	30.0	7.2	28.0	28.4	27.8	18.3	24.3
	Average	20.50			22.27			19.89		

Time is in seconds

Table D.3: Participant Task Completion Times For Three Range Disparity Tasks With System C at Three Levels of Contrast

Group	Gender	System C								
		1H	1M	1L	2H	2M	2L	3H	3M	3L
	F	30.0	14.3	21.0	30.0	27.6	24.4	30.0	0.0	12.9
	F	21.3	17.6	22.0	25.2	22.3	22.0	15.6	15.2	16.6
	F	23.9	9.6	7.6	15.4	9.4	7.4	18.1	8.1	6.6
	F	14.6	10.3	22.1	13.9	25.0	25.5	16.5	21.9	14.5
A	F	22.2	13.2	17.7	8.1	24.8	30.0	4.6	12.5	9.6
L	F	15.4	16.8	25.1	17.1	15.4	16.0	19.2	17.3	21.0
P	F	14.2	28.1	20.4	10.3	11.0	12.0	10.4	10.0	7.2
H	M	27.0	18.9	22.6	6.3	21.1	26.5	13.8	24.6	20.0
A	M	30.0	28.6	30.0	18.9	30.0	30.0	21.4	21.8	20.5
	M	11.7	11.8	16.5	23.2	11.4	30.0	17.9	17.8	16.7
	M	18.8	15.5	15.0	11.4	17.0	17.5	16.0	25.0	28.1
	M	20.2	10.3	30.0	12.4	18.1	20.2	19.0	29.1	27.5
	Average	19.29			19.09			16.85		
	F	3.9	23.6	27.5	19.4	19.3	30.0	21.2	13.4	14.6
	F	15.9	20.4	30.0	9.6	18.0	15.2	12.5	19.9	17.3
	F	27.6	20.2	24.8	26.7	19.2	17.8	30.0	16.8	16.0
B	F	17.1	12.5	14.0	19.1	30.0	19.0	30.0	27.6	24.3
E	F	25.5	21.1	30.0	20.4	28.9	20.0	21.1	30.0	21.3
T	F	6.7	21.0	5.0	21.4	25.3	18.0	21.1	8.3	20.8
A	M	22.6	21.2	28.8	30.0	30.0	30.0	28.8	30.0	12.8
	M	15.2	16.3	15.9	14.5	17.5	20.7	17.1	23.6	19.7
	M	9.9	12.3	28.0	18.8	13.1	20.2	15.1	14.3	25.1
	Average	19.14			21.18			20.47		

Time is in seconds

Appendix E

Accuracy Data

Individual success data by system, group and task. A '1' indicates the participant correctly identified the target region, a '0' indicates failure to do so.

Table E.1: Participant Accuracy Performing Three Range Disparity Tasks With vOICe at Three Levels of Contrast

Group	Gender	vOICe								
		1H	1M	1L	2H	2M	2L	3H	3M	3L
ALPHA	F	0	0	0	0	0	0	0	1	0
	F	1	0	1	0	0	0	0	0	0
	F	0	0	0	0	0	0	1	0	0
	F	0	0	0	0	0	0	0	0	0
	F	0	0	0	0	0	0	1	0	0
	F	1	1	1	0	0	1	1	1	0
	F	1	1	0	0	0	0	0	0	1
	M	1	1	1	1	0	0	0	0	0
	M	1	0	1	0	0	1	0	0	1
	M	0	0	0	1	1	0	0	1	0
	M	1	1	0	0	0	0	0	1	1
	M	1	1	1	1	0	0	1	0	0
BETA	F	0	0	0	0	0	0	0	0	0
	F	0	1	0	0	0	0	0	1	0
	F	1	0	0	0	0	0	0	0	0
	F	0	0	0	0	0	0	0	1	0
	F	0	0	0	0	1	0	0	1	1
	F	0	0	0	0	0	1	0	0	0
	M	0	0	0	0	0	0	0	0	0
	M	1	0	0	0	0	0	0	0	0
	M	1	1	0	1	0	0	0	0	0

Table E.2: Participant Accuracy Performing Three Range Disparity Tasks With AE Mono at Three Levels of Contrast

Group	Gender	AE Mono								
		1H	1M	1L	2H	2M	2L	3H	3M	3L
ALPHA	F	1	0	0	0	0	0	1	0	0
	F	0	1	0	1	1	0	1	0	0
	F	1	1	0	0	0	1	0	0	0
	F	1	0	0	0	1	0	1	0	1
	F	0	1	0	1	1	0	0	0	0
	F	1	0	0	1	1	0	0	1	0
	F	0	0	0	0	1	0	0	0	0
	M	0	1	0	0	0	0	0	0	0
	M	0	1	0	0	0	0	1	0	1
	M	1	1	1	1	1	0	1	0	0
	M	1	0	0	1	1	1	1	1	0
	M	1	0	1	1	1	1	1	1	0
BETA	F	1	1	1	0	1	1	0	1	0
	F	0	0	0	0	0	0	0	0	0
	F	1	1	1	0	1	1	1	1	0
	F	1	0	0	0	0	1	0	1	0
	F	1	0	1	1	1	1	1	1	0
	F	1	0	1	0	1	1	1	1	0
	M	0	0	0	0	1	0	0	0	1
	M	1	1	1	0	1	1	0	0	0
	M	1	1	1	1	1	0	1	1	1

Table E.3: Participant Accuracy Performing Three Range Disparity Tasks With AE Spatial at Three Levels of Contrast

Group	Gender	AE Spatial								
		1H	1M	1L	2H	2M	2L	3H	3M	3L
ALPHA	F	0	0	0	0	0	0	0	0	0
	F	1	0	0	0	0	1	1	0	0
	F	0	0	0	0	0	0	0	0	1
	F	1	1	0	1	1	1	1	0	0
	F	1	0	1	1	1	0	1	0	1
	F	0	1	1	1	1	1	1	1	1
	F	0	0	1	1	1	0	0	1	0
	M	1	0	1	1	0	1	0	0	0
	M	1	0	0	1	1	0	0	0	0
	M	1	1	1	1	1	0	1	1	1
	M	1	0	0	1	1	1	1	1	0
	M	1	1	1	1	1	1	1	1	1
BETA	F	1	0	0	1	1	1	1	1	1
	F	0	0	0	0	1	0	1	0	0
	F	0	1	0	0	0	0	0	0	0
	F	1	1	0	0	0	0	0	0	1
	F	1	0	0	0	0	0	1	0	1
	F	0	0	0	0	0	1	0	0	0
	M	1	0	0	0	0	1	0	0	1
	M	1	1	0	1	1	1	1	1	0
	M	1	1	1	0	0	1	1	1	1

Appendix F

Survey Data

Table F.1 uses the following representations in it's column headings.

- '>>' AE Mono is much more intuitive than AE Spatial
- '>' AE Mono is somewhat more intuitive than AE Spatial
- '=' AE Mono and AE Spatial are about equally intuitive
- '<' AE Spatial is somewhat more intuitive than AE Mono
- '<<' AE Spatial is much more intuitive than AE Mono

Table F.1: Comparative Intuitiveness Ratings For vOICe, AE Mono and AE Spatial

Group	Pairing	>>	>	=	<	<<
ALPHA	vOICe vs AE Mono	3	1	2	2	4
	vOICe vs AE Spatial	1	2	0	1	8
	AE Mono vs AE Spatial	0	2	1	5	4
BETA	vOICe vs AE Mono	1	1	2	3	2
	vOICe vs AE Spatial	1	2	0	2	4
	AE Mono vs AE Spatial	0	2	2	2	3

Table F.2: Subjective Auditory Fatigue Ratings For vOICe, AE Mono and AE Spatial

Group	Pairing	Very	Somewhat	Not at All
ALPHA	vOICe	2	6	4
	AE Mono	2	6	4
	AE Spatial	3	4	5
BETA	vOICe	1	4	4
	AE Mono	0	2	7
	AE Spatial	2	3	4

Table F.3: User Preference Selection Counts For vOICe, AE Mono and AE Spatial

Group	vOICe	AE Mono	AE Spatial
ALPHA	4	2	6
BETA	3	2	4

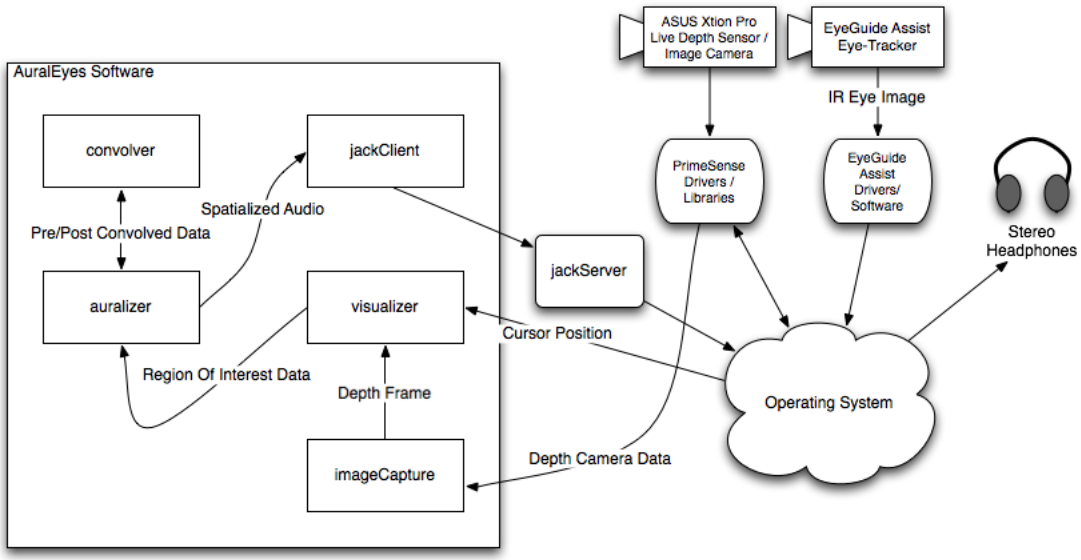
Appendix G

AuralEyes Mark-2

G.1 Aural Eyes Mark-2

The experimental results I presented in chapter 6 indicate promise in the AuralEyes approach, but also suggest that much work remains before it can be incorporated into the design of an effective ESA audio display for the blind. In order to facilitate further experimentation with AuralEyes, enable the rapid evaluation of more sophisticated audio codes, and to make possible more meaningful experimental conditions, I have developed a new prototype ESA that fully employs the AuralEyes interface.

Figure G.1: AuralEyes Mark-2 System Diagram



AuralEyes Mark-2 leverages a modified version of the software developed for the simulation and test apparatus described in chapter 5. Instead of a simple web-cam, the Mark-2 utilizes an ASUS Xtion Pro Live sensor to collect live depth maps from the user's environment. As in the experimental apparatus used in this work, an Eyeguide Assist is again used for eye-tracking. This device was selected due to its low-cost (<$1000.0 US), its ability to be head-mounted, and its wireless connectivity (see figure G.2). For sighted research subjects the eye-tracker can be calibrated using a computer monitor or projector temporarily attached to the system. For blind participants a "calibration-less" mode for the device can be utilized. The entire Mark-2 system (including mobile PC) can be constructed for around $1500.00 US.

Figure G.2: Head-Mounted Depth Sensor and Eye-Camera

To ensure portability, the multi-threaded AuralEyes software implemented for this ESA interacts with the operating system primarily through the multi-platform Jack audio server and multi-platform PrimeSense libraries. Also for reasons of portability, as well as performance, the software is equipped with its own multi-channel and multi-threaded convolution engine. This provides the potential to spatialize audio from multiple locations or regions simultaneously.

G.1.1 Principle Modules

The three most significant modules in the AuralEyes software are the auralizer, visualizer and imageCapture modules.

G.1.1.1 'auralizer' Module

The 'auralizer' module performs the necessary data-to-audio mapping and audio spatialization. This module is easily extended to accommodate different audio displays. For instance I have experimented with the spatialization of music based upon user's gaze direction. This configuration may be useful in ocular mobility training of the blind (see section 7.2.2)

G.1.1.2 'visualizer' Module

The 'visualizer' module generates a point cloud from the depth image obtained from the ASUS sensor. The coordinates of each vertex in this cloud are projected from the sensor coordinate space[1] into the user's coordinate space[2] . Using these projected coordinates, azimuth and elevation angles relative to the user's eyes are calculate for each point in the cloud. This enables the scene data to be analyzed by region/region of interest relative to the user's gaze.

Assuming the distance from the center of the user's eyes to the center of their head to be 3 inches, azimuth and elevation angles relative to the center of the user's head are determined for the position of the data to be auralized. These angles are used to select the appropriate HRTF filters for audio spatialization by the 'auralizer'.

For debugging and evaluation purposes, the 'visualizer' module also renders the current depth map from the sensor as a grayscale image on the screen or projector if present.

Like the auralization module, 'visualizer' is also easily extended, enabling experimentation with different audio-displays and data selection algorithms.

[1] origin located at the center of the sensor and aligned with the sensor body
[2] origin located at the center of the user's eyes and aligned with the axes of the head

G.1.1.3 'imageCapture'

The 'imageCapture' module is responsible for obtaining scene data from the external sensor(s). Currently this module interfaces with the PrimeSense libraries to obtain depth map data from the sensor. Future extensions may include live video capture for image and text processing applications. This module can be independently replaced or modified to accommodate alternate sensing solutions.

G.1.2 Summary

The AuralEyes Mark-2 system is a fully functional implementation of the AuralEyes user interface in a mobile ESA. As a research platform the low-cost of the hardware, and the modular and extensible design of the software make the system accessible and flexible. This system will enable further exploration and development of the AuralEyes interface, as well as audio display technologies and approaches at large.