Interface Design Implications for Recalling the Spatial Configuration of Virtual Auditory Environments

by

Kyla A. McMullen

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Computer Science and Engineering)
in The University of Michigan
2012

Doctoral Committee:

Associate Professor Gregory Wakefield, Chair
Professor David Kieras
Professor Benjamin Kuipers
Associate Professor Jason Corey

Lovingly dedicated to the memories of:

Grandmama Lottie,

Mr. Lamont Toliver,

Ms. Imade Asemota,

Ms. Kamilah Neighbors,

2nd Lieutenant Emily Perez,

and Apostle Robert A. Hill

ACKNOWLEDGEMENTS

"With every victory, let it be said of me. My source of strength, my source of hope, is Christ alone."

First and foremost I'd like to thank my Lord and Savior Jesus Christ, without whom, none of this would be possible.

I'd like to thank my dissertation committee for their help and suggestions. Whether our discussions were about the memory capacity of bumblebees or plane mapping, you have always been available.

I'd like to thank my advisor Dr. Greg Wakefield for his constant support, guidance, and mentoring. Thank you so much for teaching me what it means to be a true researcher. As unpredictable as my journey has been, he has always believed in me and helped me to acquire the tools that I needed to succeed.

This dissertation would not have been possible without financial support from the Rackham Graduate School, the College of Engineering, and the Office of Naval Research.

I share the honor of this accomplishment with:

- My *Grandmama Lottie* who had to stop attending school to take care of her family, but always dreamed of the day when her children could go to school as long as they wished.

- *Mr. Lamont Toliver*, my undergraduate mentor and friend during my tenure at UMBC. He encouraged me to be confident in my abilities, humble in my accomplishments, and to always give back to the community. His words of wisdom and encouragement will always be with me. Rest in Peace Mr. T !

- *Ms. Imade Marie Asemota*, my friend and mentee, a bright and shining star at Michigan who was pursuing a PhD in Chemical Engineering when she was called home.

- *Ms. Kamilah Neighbors*, one of my first friends at Michigan who was pursuing a PhD in Health Management & Policy. Her warm, bright presence will be missed!

- *2nd Lieutenant Emily Perez*, my high school friend (and easily one of the smartest people I've ever met). She was the first female minority Cadet Command Sergeant Major in the history of the United States Military Academy at West Point. She was called home while serving her country in Al Kifl, Iraq.

- *Apostle Robert A. Hill*, my spiritual father at Christian Love Fellowship. I have grown so much under his leadership and he will always remain a part of me.

I'd like to thank my parents (Rita and James McMullen) and family for their constant support and prayers, even when they had no idea what I was working on.

In no particular order, I'd like to thank my wonderful friends for their constant support, hugs, and advice throughout this entire process. Even when the road became harder than expected, you all were the glue that held me together. I could not assemble a better group of friends if I tried. I am truly blessed to know each of you.

I'd like to thank SMES-G, MUSES, and SCOR student groups for the social and academic support. I'd like to thank Dance Theatre Studio for helping me to maintain my sanity during this entire process. I'd like to thank my family at Christian Love Fellowship for their constant love and support.

These inanimate objects also contributed to the completion of this dissertation.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Interface Design Implications for Recalling the Spatial Configuration of Virtual Auditory Environments

by

Kyla A. McMullen

Chair: Gregory Wakefield

Although the concept of virtual spatial audio has existed for almost twenty-five years, only in the past fifteen years has modern computing technology enabled the real-time processing needed to deliver high-precision spatial audio. Furthermore, the concept of virtually walking through an auditory environment did not exist. The applications of such an interface have numerous potential uses. Spatial audio has the potential to be used in various manners ranging from enhancing sounds delivered in virtual gaming worlds to conveying spatial locations in real-time emergency response systems.

To incorporate this technology in real-world systems, various concerns should be addressed. First, to widely incorporate spatial audio into real-world systems, head-related transfer functions (HRTFs) must be inexpensively created for each user. The present study further investigated an HRTF subjective selection procedure previously developed within our research group. Users discriminated auditory cues to subjectively select their preferred HRTF from a publicly available database. Next, the issue of training to find virtual sources was addressed. Listeners participated in a localization training experiment using their selected HRTFs. The training procedure

was created from the characterization of successful search strategies in prior auditory search experiments. Search accuracy significantly improved after listeners performed the training procedure.

Next, in the investigation of auditory spatial memory, listeners completed three search and recall tasks with differing recall methods. Recall accuracy significantly decreased in tasks that required the storage of sound source configurations in memory. To assess the impacts of practical scenarios, the present work assessed the performance effects of: signal uncertainty, visual augmentation, and different attenuation modeling. Fortunately, source uncertainty did not affect listeners' ability to recall or identify sound sources. The present study also found that the presence of visual reference frames significantly increased recall accuracy. Additionally, the incorporation of drastic attenuation significantly improved environment recall accuracy. Through investigating the aforementioned concerns, the present study made initial footsteps guiding the design of virtual auditory environments that support spatial configuration recall.

CHAPTER I

Introduction

As a part of our everyday lives, we humans navigate the world around us. Initially, this vast world is unfamiliar territory, but somehow, we manage to find our way (most of the time). Using distinct sources of information, visual and vestibular, we are excellent at perceiving and remembering the environment we traverse.

Consider navigating the environment only using a different information source, the auditory channel. This would be a difficult, but feasible task. Fortunately, humans are familiar with using auditory cues to navigate. From birth we have learned to use our ears to perceive our position relative to those of acoustic sources in the world around us. This is particularly true of visually-impaired individuals who can be trained to navigate their environment using acoustic cues. Sound is already used for navigation in many contexts. For example, at sea, the sounds of buoys and foghorns are used to guide ships (*Sobey* (2006)).

To further complicate the task, consider navigating while receiving vestibular information virtually. Indeed, this may sound like a far-fetched situation, however many systems use virtual audio to convey spatial information to a listener. For example, sonar operators use headphones to listen to a representation of their underwater surroundings. Additionally, researchers from many areas including gaming, music, and psychoacoustics have encouraged studying how to display a virtual auditory environ-

ment (VAE) to a listener.

Unfortunately, few researchers have investigated how listeners acquire spatial information through navigating a VAE. This dearth was partly due to a limitation in adequate technology. Although the concept of virtual spatial audio has existed for almost twenty-five years, only in the past fifteen years has modern computing technology enabled the real-time processing needed to deliver high-precision spatial audio to render sonic cues as a listener moves. Characterizing this behavior is important because the ability to walk in a VAE is needed in systems where visual cues are degraded or direct sensation of the physical world may not be possible or desirable (*Wenzel et al.* (1988b)).

1.1 Objective

The objective of this dissertation is to further understand listener acquisition of spatial information when navigating a VAE. We created VAE in which the listener used a mouse to search for multiple sound sources. The system outputs the spatial sound over headphones using a customized filter for each listener. The filter characterizes how spatial sounds are heard; however, creating veridical filters is very costly. In the absence of direct measurement, we successfully used a method to combine pre-measured filters to create each listener's personalized filter.

Because virtual sound localization is not innate, a search training procedure was developed. Due to the lack of research in this area, observations from the search behaviors of successful navigators were used to develop the tasks of the training procedure. We found that listeners' performance significantly improved after training. Once listeners were trained to find sounds in the VAE, the present study investigated their ability to remember the locations and identities. Generally, listeners recalled the VAE best when allowed to freely recall the environmental information while minimizing the delay between presentation and response. The present study also assessed

listener performance in three "real-system" conditions: sound source uncertainty, using visual cues as a reference, and non-standard attenuation.

The primary contribution is an understanding of how listeners remember and recall the virtual worlds they navigate. In addition to aiding users to navigate visually degraded or dangerous environments, our research findings could be used to create subsystems to augment navigation of visual or haptic environments, thus creating a multimodal system, which has many advantages such as: error prevention, interface robustness, error correction / recovery, increasing communication bandwidth, and providing alternate communication methods (*Cohen and McGee* (2004)).

1.2 Dissertation Organization

The remainder of this dissertation is organized as follows:

- Chapter 2, *Background*, reviews the necessary literature in spatial audio and auditory interfaces.
- Chapter 3, *HRTF Selection*, describes the process used to select HRTFs for the listeners of the dissertation.
- Chapter 4, *Virtual Auditory Search and Training*, characterizes search strategies used by VAE listeners. From the characterization, a training procedure is described and assessed.
- Chapter 5, *Auditory Spatial Memory*, assesses listeners' memory of auditory spatial objects.
- Chapter 6, *Three Issues in Systems Integration*, discusses the implications of semi-practical system scenarios on the memory of auditory spatial objects.
- Chapter 7, *Conclusion and Future Directions*, reviews the research objectives and suggests areas for future research.

CHAPTER II

Background

2.1 Introduction

This chapter provides a brief overview of relevant spatial audio rendering concepts and the development of virtual auditory environments. In this chapter, we introduce the physical characteristics of sound and how these characteristics influence its perception. When a sound is perceived, the brain automatically determines its position by comparing the cues received by each ear and analyzing the sound's quality. These cues comprise head-related transfer functions (HRTFs) that characterize how the ear receives sound. Only recently has technology enabled the real-time processing of HRTFs to render virtual spatial sounds. With this technology, listeners can explore a virtual auditory environment while receiving real-time sonic cues. To aid the study of listener navigation in VAEs, HRTF creation must be quick and inexpensive. When creating VAEs to study navigation, there are many implementation options to consider that affect the quality of the environment.

2.1.1 Characteristics of sound

Sound is described as a longitudinal wave that is an oscillation of pressure transmitted through a solid, liquid, or gas. Sound propagates through compressible media such as air or water. The variation of pressure as a function of time can be called a

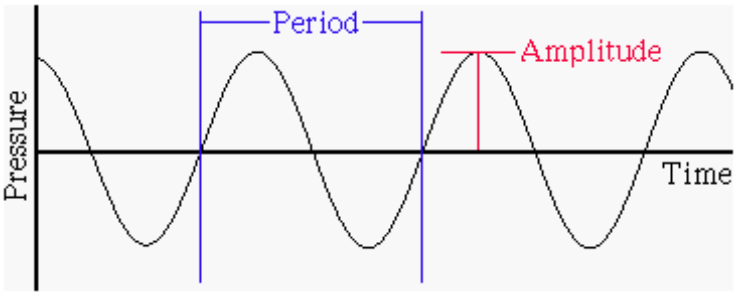


Figure 2.1: Variations in air pressure over time for a pure tone (from *Hass et al.* (2003)).

sound wave (*Green* (1976)). Periodic sound waves, such as sinusoids, are characterized by *frequency*, *intensity*, and *phase*.

Figure 2.1 shows a periodic sound as a function of pressure changes over time. The pitch of a sound depends on its frequency. *Frequency* is the number of times a wave repeats itself (or periods) per unit time. Frequency is usually measure in the number of cycles per second or Hertz (Hz). Humans are sensitive to sound at frequencies between about 20 Hz and 22 kHz (*Shilling and Shinn-Cunningham* (2000)). The amplitude of the wave determines the sound's loudness or intensity. *Intensity* is proportional to a sound wave's amplitude squared. A sound's intensity is measured in decibels (dB). The decibel is not an absolute measure of a sound's intensity; however, it represents the relationship between the intensities of two sounds. A sound's *phase* is the fractional part of a period as measured at any point in time.

Typically, we associate spatial sounds with an azimuth and elevation (Figure 2.2). The azimuth specifies the direction of the sound source, which is the angle of direction on the horizontal plane. Elevation is specified by the angle of direction in the median plane. A sound with 0° azimuth and 0° elevation comes from straight ahead. Humans can localize sustained sounds in terms of their position in this field.

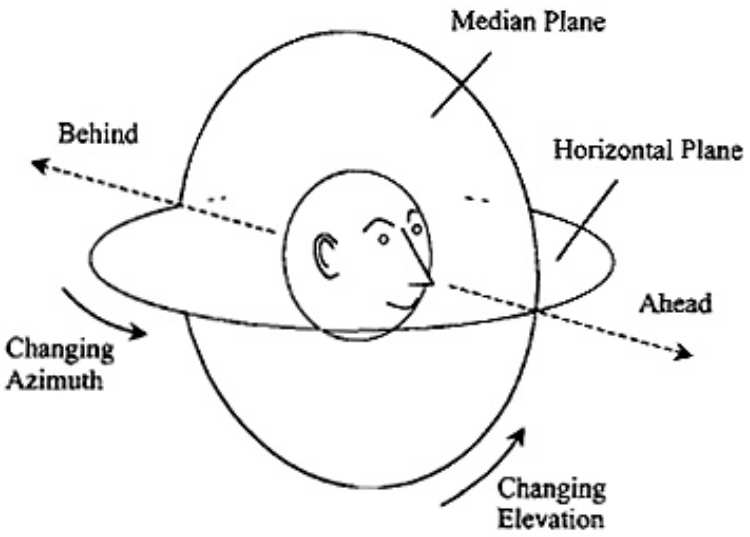


Figure 2.2: Coordinate system for sound direction.

2.1.2 Interaural cues for localization

Because the ears are located at two different places on the head, humans can determine the location of a sound based on these interaural differences. For example, sound waves reach each ear at different times, which is referred to as the interaural time difference, or ITD (Figure 2.3). A sound is perceived to be closer to the ear at which the wavefront first arrives, or the ipsilateral ear. If a source is directly in on the median plane, there is no time difference between the two ears.

Additionally, when a sound is heard, the head shadows the sound received at the farther, or contralateral, ear. The shadowing creates an intensity difference between the right and left ears, which is called the interaural intensity difference, or IID (Figure 2.4).

Lord Rayleigh's foundational model of the head as a sphere explains much of the IID and ITD behavior in the estimation of a sound's location (*Rayleigh* (1907)). His theory, the duplex theory, is based on the assumption that the essential cues of a sound's location are based on the interaural differences between the sound waves at each ear. For the ideal spherical head model (Figure 2.5), physics tells us that the

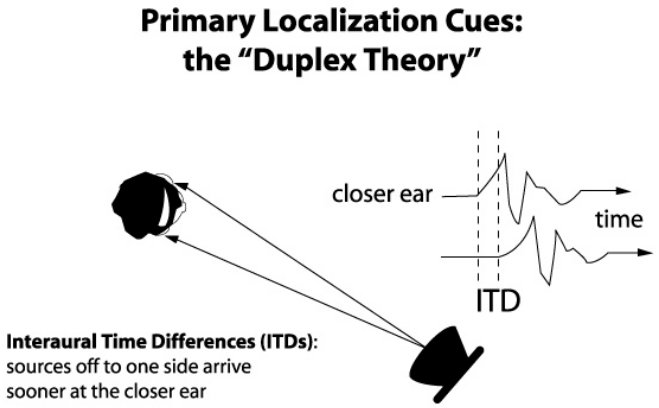


Figure 2.3: Interaural Time Differences of a sound in space. The sound arrives sooner to the ear closest to the sound source.

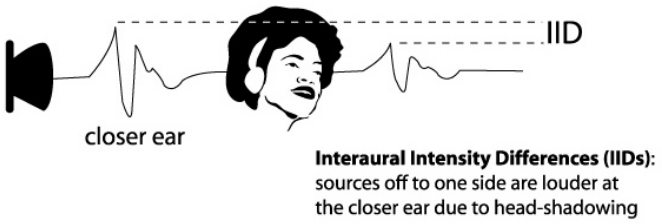


Figure 2.4: Interaural Intensity Differences of a sound in space. The intensity of the sound is higher at the ear closest to the sound source.

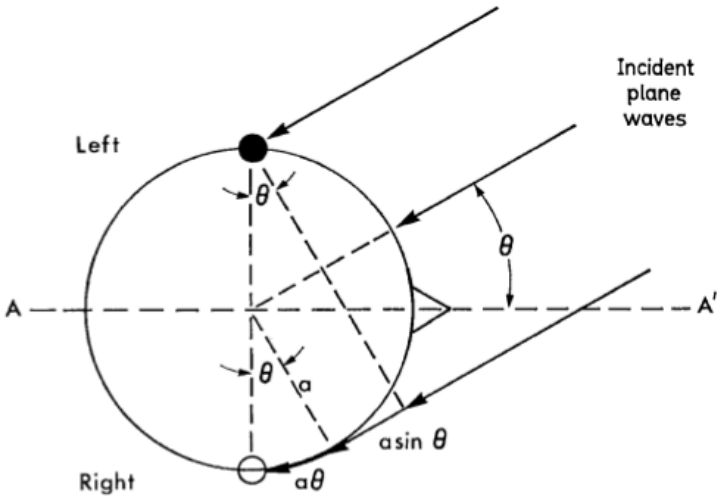


Figure 2.5: A plane wave impinging a spherical approximation of the human head (from *Carlile* (1996))

arrival times and amplitudes of plane waves are different at the two ears at all angles except 0 and 180 degrees.

2.1.2.1 Challenges of binaural localization

IIDs and ITDs primarily mediate the localization of sounds in the horizontal plane (*Grantham* (1995)). However, there are many limitations when using these interaural differences to determine the location of a persistent periodic sound. Rayleigh is also credited with pointing out the ambiguity in using ITD cues to localize pure (sinusoidal) tones. Pure tones lack transients, which are characteristic of most natural sounds. Without transients, the ITD is essentially a phase difference between the two ears. While the time delay is independent of frequency, the corresponding phase delays depend on frequency, as expressed in equations 2.1, 2.2, and 2.3. The following expression represents the difference in path lengths, as shown in Figure 2.5,

$$\Delta = \alpha(\theta + sin\theta) \tag{2.1}$$

where α is the radius of the head in meters and θ is the angle of the sound source

from the median plane in radians. If we assume that the speed of sound is constant across frequency, the ITD is expressed as

$$\tau = \Delta c \tag{2.2}$$

where c is the speed of sound (343 meters/s), and the IPD is defined as

$$IPD = 2\pi * f * \tau \tag{2.3}$$

where f is the frequency in Hz. The equations demonstrate that an observed phase difference is consistent with several or more time delays depending on the wavelength, speed of sound, and size of the head. The phase difference is problematic for higher frequency sounds that are smaller than the size of the head because a given phase difference can correspond to multiple angular locations. For an average ear separation of $8.75cm$ (*Algazi and Duda* (2001)), a phase ambiguity will develop for frequencies above 1960 Hz.

Furthermore, below approximately 1 kHz, IID is no longer effective as a natural spatial hearing cue because longer wavelengths will diffract around the head, thus minimizing the IID (*Begault* (1994)). It is for these reasons that the IID is primarily a high-frequency cue and the ITD is a low-frequency cue.

The use of IID and ITD for localization provides good azimuthal cues on the horizontal plane; however, when duplex theory is applied to free space (including elevation and distance) an additional challenge is introduced. There are many locations in space that have indistinguishable ITD and IID cues. These points form a cone that is centered on the interaural axis (Figure 2.6). All of the points along this *cone of confusion* are identical to listeners. Localization on the horizontal plane, can also result in confusions between sound localization in front of or behind the listener. These confusions are called front-back reversals (*Blauert* (1983)).

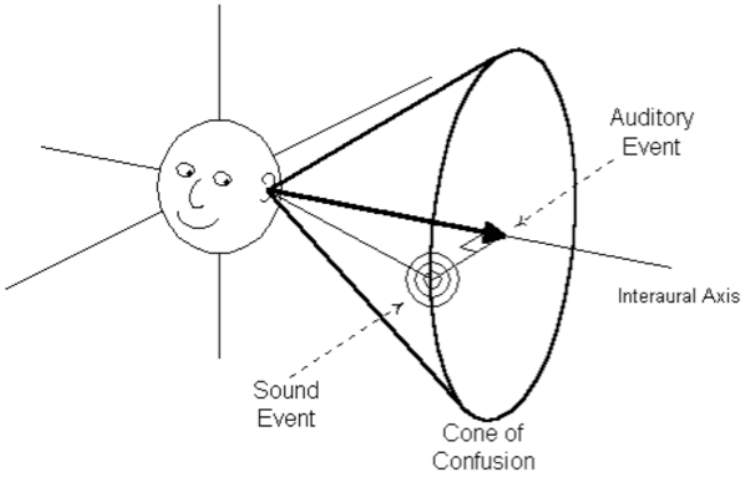


Figure 2.6: Cone of confusion. All points in the cone have ITDs and IIDs that are indistinguishable to listeners.

This problem is less pronounced in the median plane, where the IID and ITD are essentially zero. Interestingly, listeners can distinguish sound sources on the median plane. Additionally, since ITD and IID are indistinguishable in the cone of confusion, other types of cues are needed to explain discrimination between different positions around its circumference. This suggested the existence of monaural spectral cues to aid localization. These cues consist of the spectral filtering of an incoming sound wave by the head, torso, and most importantly pinnae (outer ear) (*Hebrank and Wright* (1974)).

It should also be noted that head movement produces dynamic changes in ITD and IID. These changes are dynamic cues, which disambiguate front/back cues (*Wightman and Kistler* (1999)) and elevation (*Perrett and Noble* (1997)).

Sophisticated measuring technology was developed after Lord Raleigh's foundational model. Subsequent studies showed that by placing a listener in an anechoic chamber and positioning a loudspeaker at a given location, it was possible to measure the entire acoustic transformation. Such measurements are called head-related transfer functions (HRTFs). If a sound is filtered through an HRTF corresponding to a certain location, the sound is heard at that location when delivered over headphones.

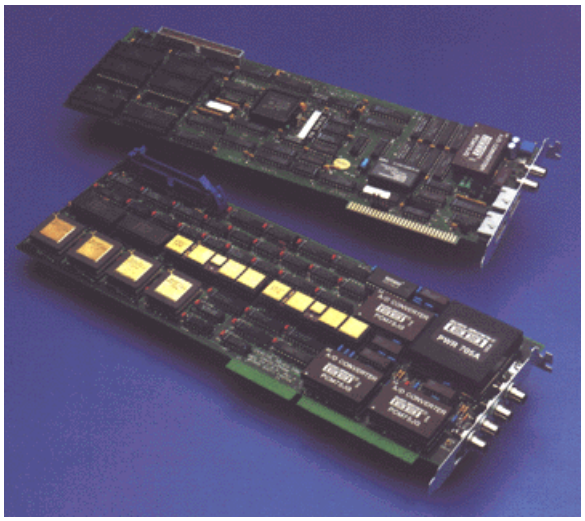


Figure 2.7: Convolvotron

They depend on the azimuth and elevation of the source relative to the listener and characterize not only the ITD and IIDs, per Rayleigh's model, but the additional spectral effects of the pinnae (outer ear).

2.1.2.2 Technological Advancements Aid Spatial Audio Rendering

Technological advances have afforded the real-time rendering of spatial audio cues. The Convolvotron (Figure 2.7) was the first real-time spatial audio processor that used HRTFs to replicate the ITD, IID, and pinnae localization cues. Beth Wenzel and Scott Foster created the Convolvotron in 1992 for NASA's VIEW (Virtual Interactive Environment Workstation) project. The Convolvotron uses the HRTFs computed by *Wightman and Kistler* (1989b).

In the Convolvotron (Figure 2.8), a set of two printed circuit boards converts one or more monaural analog source inputs to digital signals at a rate of 50kHz (16-bit resolution). Each stream is convolved with filter coefficients determined by the coordinates of the desired target locations and the position of the listeners head (using the head-tracker). The resulting data streams are mixed, converted to left and right analog signals, and output over headphones as spatialized sound (*Wenzel* (1992)).

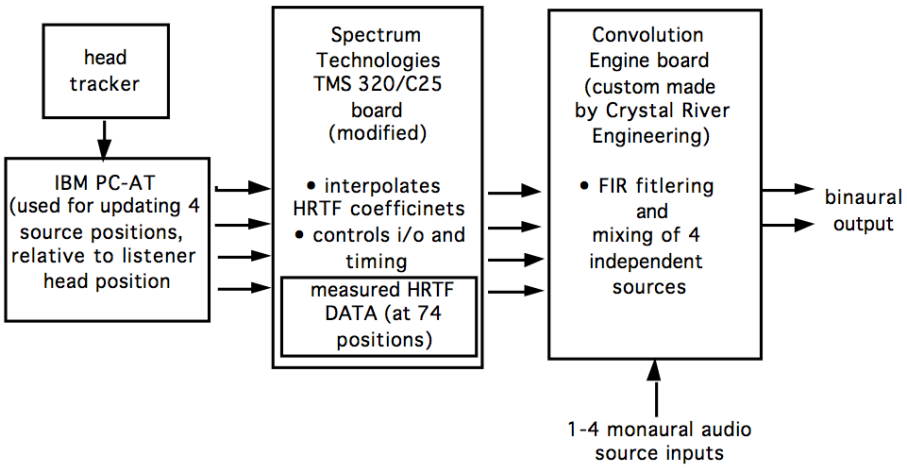


Figure 2.8: Convolvotron block diagram (from *Wenzel* (1992))

For locations where the sound location was known, each sound was spatialized using the specified HRTF filter. For locations where the HRTF was not known, the nearest four points were interpolated to create the filter.

Essentially, the Convolvotron is a piece of special-purpose digital signal processing (DSP) hardware for executing the convolution stage ("Convolution Engine" in Figure 2.8). There are many calculations involved in real time digital filtering. Digital filtering is done by convolving the sampled input signal *x(n)* with the desired filter *f(n)* of length L. The result is the filtered signal *y(n)*.

$$y(n) = \sum_{i=1}^{L-1} f(i)x(n-i) \tag{2.4}$$

The calculation of each element of y(n) (equation 2.4) must be performed in less time than the sample length of the filter. Calculating a particular *y(n)* involves at least L multiplications, additions, and memory shifts, referred to as convolution points or taps. The required amount of convolution points per second (cps) in real time processing is:

$$cps = filter - length * samplerate$$

In real-time headphone rendering, two filtering operations are required (one for each ear). Assuming a filter length of 512 points and a sample rate of 44.1 kHz, the required *cps* is:

$$2 * 512 * 44.1 = 45.16 * 10^6 points/second$$

This is the amount of work needed for one sound source. The same amount of work is required to render each additional sound.

The Convolvotron contained specialized hardware to meet the computational demands of real-time digital filtering. It consisted of two boards: the spectrum board and the convolution engine card. The spectrum board was a modified TMS320-C25 system board, which included a Texas Instruments TMS320 DSP chip. The spectrum board primarily performed the interpolation calculations used for the convolution engine card, as well as performing input data buffering and clock signal generation. The convolution engine was a custom-made board that contained four INMOS A-100 DSPs, each containing 32 parallel 16-by-16 multipliers. The engine could reach a peak convolution speed of 320 million taps/sec (*Begault* (2000)).

The advancement of spatial sound rendering is directly tied to the evolution of DSP chips (*Begault* (2000)). DSPs differ from ordinary microprocessors in that they are specifically designed to rapidly perform the sum of products operation required in many discrete-time signal processing algorithms (*Tretter* (2008)).

With each generation of DSPs, processing performance improved. Texas Instruments produced the first generation of commercially successful DSPs in 1982. The TMS32010 chip used 16-bit data and needed 390ns for each tap while it completed in 5 million instructions per second (MIPS). MIPS were a popular measurement before computers reached gigahertz speeds. The second generation of DSPs, created in 1987, mostly operated on 24-bit data and a typical model only required about 75ns per tap as it processed 13 MIPS. Examples of such chips are the Texas Instru-

ments TMS320C50 AT&T DSP16A, Analog Devices ADSP-2100, and the Motorola DSP56001. These processors generally operate at around 20-50 MHz, and provide good DSP performance while maintaining very modest power consumption and memory usage. DSP speed continued to improve and the third generation of DSPs, created in 1995, was introduced. The chips of this generation, such as the Texas Instruments TMS320C541 and Motorola DSP56301 processed taps in 20ns in 50 MIPS and operate at 100-150 MHz. During this time, architectural innovation also gave rise to multi-processor chips, such as the Texas Instruments TMS320C80 and the Motorola MC68356. The fourth generation, around 1997, gave rise to chips such as the Texas Instruments TMS320C6201, which processed taps in 3ns in 120 MIPS, consuming 32 bits. Specialized multimedia extensions, such as the Intel Pentium with MMX were also created in this period. It processed 466 MIPS. *Eyre and Bier* (2000)

From 1982 to 1997, DSP processor performance increased by a factor of about 150. DSPs became increasingly specialized for applications, however many general-purpose processors were also viable options for DSP applications.

Many general-purpose CPUs, such as Pentiums and PowerPCs, were also enhanced to increase the speed of computations associated with signal processing tasks. The most common modification was the addition of SIMD-based instruction-set extensions, such as MMX and SSE for the Pentium, and AltiVec for the PowerPC. High-performance CPUs typically operate around 3.6 GHz MHz, while the fastest DSP processors operate in the 1.2 GHz range.

As technology advanced, so did the availability of commercial spatial audio systems. Faster spatial audio simulators were developed that could convolve more sounds. Various companies, such as AuSIM, OpenAL and VRSonic, have created commercial products that use a digital sound, an HRTF, and position as input to render spatial audio over headphones as output. The companies differ with respect to how much of the processing is executed on dedicated DSP units as compared to

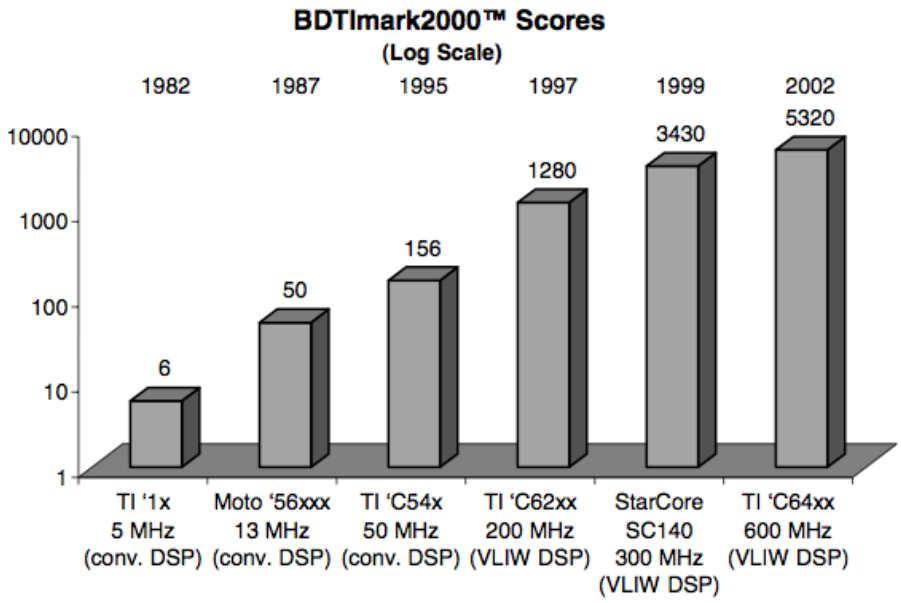


Figure 2.9: The speed increases achieved by DSPs from 1982 to 2002 through a combination of faster clock rates and more powerful DSP architectures. The BDTImark2000 is a DSP speed metric based on a processor's results on BDTI's suite of DSP benchmarks; higher is faster (from *BDTI* (2003)).

general-purpose CPUs. By using these newer spatial audio simulators, many more sound sources can be rendered in virtual locations.

The computational ability to enable real time spatial audio rendering and advancements in our understanding of spatial hearing contributed to the surge of interest to study spatial audio display systems (*Arons* (1992)). We must remember that spatial audio cues for localization are idiosyncratic. Thus, for the widespread study and use of HRTFs to render spatial audio, HRTF specification must be quick and inexpensive.

2.1.3 HRTF Specification

HRTFs differ by individual as a function of head shape, placement of the pinna, and the shape of the pinna and ear canal. When these differences are large, localization accuracy can be degraded (*Middlebrooks* (1999); *Wenzel et al.* (1993); *Zahorik et al.* (2006)). The potential for increased localization error suggests that the HRTFs selected to render sounds in spatial audio be as close as possible to those of the indi-

vidual user. From a practical standpoint, however, requiring individually-measured HRTFs adds a potentially costly layer of complexity to any spatial audio system. Therefore, finding a means to provide some degree of user selection is important in any practical and useful system.

The most accurate HRTFs are created through a costly and time consuming procedure called *explicit measurement* where individualized HRTFs are measured on a human listener's ears (*Wightman and Kistler* (1989b); *Hammershoi et al.* (1992); *Bronkhorst* (1995); *Moller et al.* (1996)). Microphones are placed in the subject's ear canal (open meatus - *Wightman and Kistler* (1989b)) or at the entrance of the plugged ear canal (blocked meatus - *Hammershoi et al.* (1992)). After the microphones are placed, a wideband signal is played through a loudspeaker at a specific azimuth θ, elevation φ, and distance from the subject's head (Figure 2.10). The frequency response at each ear is recorded and one of several system identification techniques is used to estimate the head-related transfer function. The process is repeated for sound sources at various locations. A sound source can be rendered at a given location in space by filtering that source with the left and right HRTFs for that location and playing the output over headphones. Although using individualized HRTFs leads to a more accurate spatial image, the explicit measurement procedure can be both costly, time consuming, and typically requires specialized equipment, such as an anechoic chamber. From a practical standpoint, individual measurements are a poor way to customize the HRTFs for a given listener.

Alternatives to acoustical measurement of the HRTFs have been considered. Many researchers have worked on HRTF approximation, which typically involves using theoretical models or pre-measured HRTFs.

There are two notable theoretical procedures to approximate HRTFs: theoretical computation and active sensory tuning. Individual HRTFs can be developed using *theoretical computation* based on individualized anthropometry (*Algazi et al.* (2002),

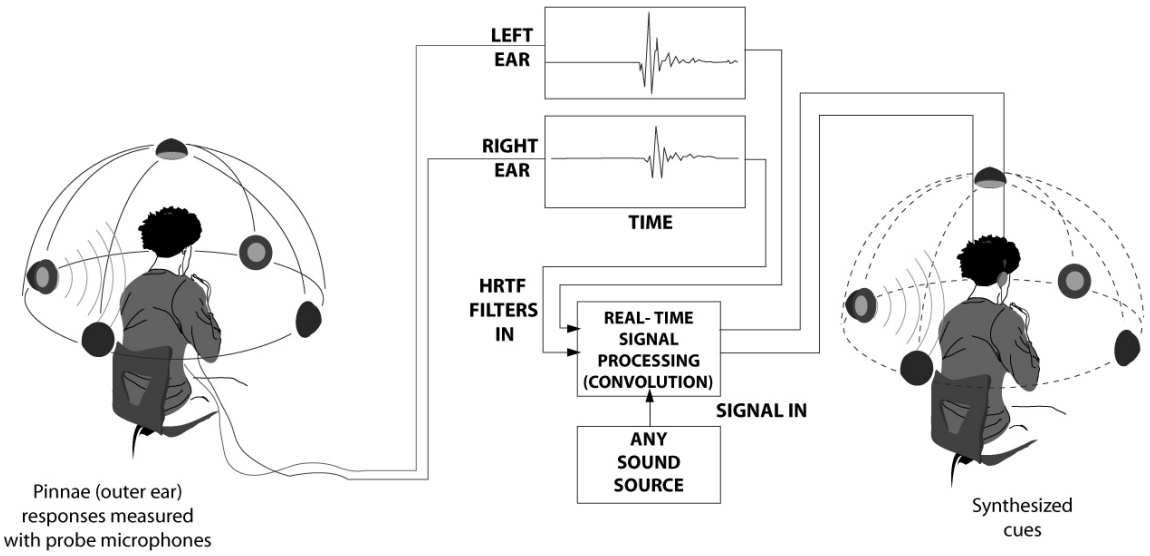


Figure 2.10: HRTF direct measurement process for a single individual.

Terai and Kakuhari (2003)). This method is less costly than individual HRTF measurement. On the surface, this method appears to be ideal for HRTF specification; however, the method of theoretical computation is based upon simplified geometric models, which may lead to inaccuracies in the HRTFs. While useful as a first-order approximation, such models are not easily extended to incorporate more complex anthropometric measurements.

Another approach to customizing the choice of HRTFs is allowing the listener to search through a set of options. Interactive genetic algorithms (IGAs) are one class of psychophysical procedure for searching through large multidimensional parameter spaces. *Runkle et al.* (2000) proposed a low-order pole-zero model of the HRTF and developed an IGA for subjective tuning of the model parameters. In a similar method proposed by *Silzle* (2002), the tuning and selection of HRTF by tuning experts can be done individually for every desired direction.

Pre-measured HRTFs can also be used to approximate HRTFs for a listener. *Standardized measurement* (sometimes referred to as generic measurement) involves measuring the HRTFs of a manikin or selected humans (*Bronkhorst* (1995);*Moller*

Figure 2.11: Knowles Electronic Manikin For Acoustic Research(KEMAR) used for standardized HRTF measurement

et al. (1996);*Wenzel et al.* (1993)). KEMAR as shown in Figure 2.7, is the Knowles Electronic Manikin for Acoustic Research and has become a standard for measuring HRTFs, however, *Moller et al.* (1996) observed that KEMAR HRTFs often render sources with decreased externalization and degraded localization accuracy. Besides KEMAR HRTFs, several research groups have published databases of HRTFs from a large number of subjects, which are available for use in spatial audio systems (*Algazi and Duda* (2001); *Gardner and Martin* (1995); *Blanco-Martin et al.* (2011))

Data from pre-measured HRTFs are used to create HRTFs in *physical feature measurement* and *subjective selection*. In physical feature measurement, the listener's ears and head are analyzed to create their HRTF. *Wenzel et al.* (1988a) studied the differences between listeners during sound localization and found that a listener's elevation accuracy could be predicted by analyzing the acoustic characteristics of the listener's outer ears. Many of the individual differences in localization behavior are due to individual differences in outer-ear acoustics.

Individualized HRTFs can be created by measuring physical features of the pinnae and using the correlation between those features and pre-measured HRTFs to fine-

tune the HRTF (*Inoue et al.* (2005); *Gumerov et al.* (2002); *Zotkin et al.* (2002); *Zotkin et al.* (2003); *Jin et al.* (2000)).

In *subjective selection*, a listener selects HRTFs based on listening to sounds rendered by a variety of HRTFs and choosing the best. For example, in *Seeber and Fastl* (2003), pulses of 30 ms white noise were generated sequentially at -40°, -20°, 0°, 20°, and 40° on the horizontal plane and were played over an electrostatic headphone. The change in azimuth over time created the perception of a moving noise source. Participants compared the quality of the moving source rendered by several pre-measured HRTFs according to externalization, front/back distinction, perceived direction, and source width. Once the listener picked a given set of HRTFs, they were tested in a localization task. Seeber and Fastl observed that localization errors decreased when the listener's selected HRTF set was used when compared with other HRTF sets. The results indicated that subjective selection minimized the variance of the localization responses, the number of inside-the-head localizations, front/back confusion, and localization error.

In *Roginska et al.* (2010a) participants listened to sound cues and chose their preferred HRTFs based on tournament-style listening tests. Listeners were asked to reject HRTFs in which pairs of sources were poorly discriminated on the basis of externalization, elevation, and front/back distinctiveness. Included among the pre-measured HRTFs for each listener was their own HRTF. The findings showed that listeners preferred a subset of standardized HRTFs as often as they preferred their own HRTF.

Among all of the methods for selecting a set of HRTFs, subjective selection appears to be the easiest to implement while still providing reasonable localization performance. Validity experiments by *Iwaya* (2006) and *Saito and Iwaya* (2004) showed that the virtual sound localization performance listening using HRTFs fitted with this method was similar to the performance of subjects using their own HRTFs.

2.2 Implementation of the virtual auditory environment

A virtual auditory environment is more than just a choice of computational machinery and set of HRTFs for rendering. It also must be populated with acoustic sources, the analog to objects in virtual visual environments. These sources should, at a minimum, be located at a given position, but they may also bear other types of information. They may be continuous, in which case they are also generating sound, or may exhibit various degrees of intermittency. How the listener is placed in the VAE must also be determined. Whether sounds are delivered over headphones or loudspeakers influences both the cost and flexibility of the resulting system. Finally, the manner by which the user interacts with the environment must also be determined. Whether they are immobile, can orient the position of their head, or move through the environment are choices that might be dictated by convenience or necessity. In the following, each of these topics is discussed.

2.2.1 Speech vs. non-speech sounds

Sources in a VAE generate sounds. If the goal is to study human performance in navigating through VAEs, it is important to note that not all sounds are equally localizable in VAEs. For example, *Loomis* (1985) suggests using a virtual auditory display that identifies landmarks using synthesized speech sounds. The speech sounds in the interface were spatialized at a given location. Although speech sounds provide clarity when describing the identity of a virtual location, speech cues are harder to localize in a virtual environment than non-speech cues (*Tran et al.* (2000)).

Non-speech sounds are alternatives to speech sounds. They are classified as beacons and auditory icons. Beacons are typically described as "boings, bangs, squeaks, clicks, etc.". They are commonly used in alert or warning systems. For example, *Rutherford and Withington* (2001) explored the use of beacons to aid emergency egress from buildings, ships, oil exploration platforms, and airplanes. Beacon us-

age does not face the same limitations as speech sounds as they can be designed so that their spectrum highlights relevant binaural cues for localization. However, like speech, beacons bear some information content that is separate from their location. The mapping of a beacon's sound to its meaning in the interface may add to the listener's task and possibly interfere with their spatial navigation. One limitation in their usage is that the listener must learn the mapping of each beacon sound to its intended meaning.

Auditory icons are nonspeech, "naturalistic" sound sources located at specific positions within the VAE. They are the building blocks of auditory soundscapes in which a sonic world is created for the listener. These natural sounds have associated meanings that can be mapped onto similar, familiar meanings in the interface (i.e. the sound of water, representing a water fountain). The listener does not face the additional task of learning the mapping of the auditory icon to its intended meaning. Like beacons, auditory icons do not face the same limitations as speech sounds. It is for these reasons that auditory icons were used in the VAE implemented for the experiments of this dissertation.

2.2.2 Intermittent vs. Continuous Sounds

In the absence of sound, a listener cannot determine the location of a source from what they hear. Intermittent sources are relatively brief sonic events over time. Accordingly, should the position of an intermittent source relative to the listener change over time, any changes in position over silent intervals must be inferred from localization cues present during each sonic event. For continuous sources, no such ambiguities exist as relative position changes, as the sources emit sound continuously. Thus, as a listener moves through a VAE, their position with respect to that of a waterfall will always be observable, whereas their position with respect to that of a cricket will be observable only during those times when the cricket chirps. To

eliminate potential errors in positional judgment due to missing the needed time slice of acoustic data, only continuous sources will be used in the thesis.

2.2.3 Loudspeakers vs. Headphones

Typically, VAE sounds sources are played over loudspeakers or through headphones. A challenge in using external speakers to deliver spatial sounds is positioning the listener such that the sound delivered through the speakers give the proper spatial audio cues. Headphones do not face the same challenges as loudspeakers inasmuch as they maintain their position relative to the listener regardless of listener position in the environment. However, headphones require substantial amounts of signal processing to affect the same changes in the acoustic field as occur when a listener rotates their head within a loudspeaker environment. The thesis will use headphones and a spatial audio rendering algorithm to create VAEs.

2.2.4 Exploration freedom

Listeners can interact with a VAE either through directed or free exploration. For example, navigation in the SWAN system (*Walker and Lindsay* (2005)) is an example of directed exploration. In the SWAN system, navigational "beacons" are constructed from intermittent sounds to provide the waypoints that listeners should follow in order to reach a previously determined destination. As the user approaches a waypoint (source), the intermittency of the beacon sound decreases. After the waypoint has been reached, the associated beacon is turned off and the beacon for the next waypoint is turned on. In this interface, listeners cannot freely explore the environment, as they are required to follow a prescribed path in the system. Listeners only hear one beacon at a time so they cannot benefit from the knowledge could be obtained from the detection of other sounds in the environment.

Another type of directed exploration is passive exploration. In passive explo-

ration, spatial knowledge is acquired as the observer is moved through the VAE by the experimenter. *Hahm et al.* (2007) investigated the effects of passive and active exploration on recall. In their study, listeners were less accurate when recalling passively learned objects than actively learned ones. Some studies even suggest that passive navigation causes cybersickness. Cybersickness is a type of motion sickness affecting virtual environment users. The sickness experienced by users can range from a slight headache to a nauseating feeling. One promising approach to moderating cybersickness is through the manipulation of the level of interactive control provided to users (*Stanney and Hash* (1998)). *Reason and Diaz* (1971) and *Casali and Wierwille* (1986) found that crew members and copilots are more susceptible to this form of motion sickness because they have little or no control of the plane's movement. *Lackner* (1990) suggested that the system operator becomes less sick than the passengers because they can control and anticipate the motion.

This thesis is particularly interested in how users learn about sources in their VAE without any experimental intervention. Therefore, rather than prescribing the path along which either the user or display explores the VAE, the dissertation uses free exploration instead. Under free exploration, the listener has complete control over the path they choose to follow when orienting to a VAE.

2.2.5 Navigation Mediation

There are many mediations that listeners can use to facilitate navigation through virtual spaces. For example, subjects in *Holland et al.* (2002) walked in a physical environment that was mapped to a virtual environment and their navigation was mediated through GPS. In this type of interface, the GPS update rate may not be sufficient to indicate fine positional changes. For example the GPS signal can be lost as the user is walking. Also, the GPS update rate may not be sufficient to indicate the types of positional changes that make a difference acoustically.

Head-tracker mediation is another approach to virtual mediation. While head movement is a natural interaction style, using head trackers can be spatially limiting and dysfunctional in certain environments. When sitting in a stationary chair, head motion is limited by the extent to which the listener can turn their head. This can be ameliorated somewhat by having the listener seated in a swivel chair. Head trackers differ with respect to the technology used to sense head position. Magnetic trackers, for example, are particularly susceptible to the presence of ferrous objects in the vicinity of the sensor. Careful calibration of such sensors is necessary, but these settings are known to drift over time. Also, depending on the type of technology used, processing the data from a head tracker can be a costly operation, resulting in an update lag, which also contributes to cybersickness.

Mouse mediation has become a standard means for users to navigate a computer screen. Most adults are familiar with mouse usage and its usage does not face the same limitations as GPS and head-tracker usage. One may propose that mouse interaction is less natural than walking and head movement thus making interaction more difficult than a more "natural" mediation. Fortunately, our work in *Roginska et al.* (2010b) discovered that mouse mediation is equally as effective as a head-tracker mediation to locate sounds in a VAE. Because of these factors, mouse mediation was used to implement navigation in our VAE.

2.2.6 Summary

This chapter reviewed the relevant literature on spatial audio rendering and described the implementation of virtual auditory environments. The chapters that follow, serve to describe the methods used to address the objectives of this dissertation.

CHAPTER III

HRTF Selection

3.1 Introduction

Given that each listener in our study will need an HRTF to perceive sounds in the VAE, we must consider the tradeoffs associated with each HRTF specification method. Undoubtedly, individualized measurement would produce the most accurate HRTF; however, the procedure is not practical, given the costliness of the infrastructure and laboratory personnel often required to measure individual HRTFs (*Begault* (1994); *Wightman and Kistler* (1989a); *Bronkhorst* (1995)).

As you may recall from Chapter II, subjective selection rises as the least expensive, yet most effective solution (*Algazi and Duda* (2001)). For subjective selection, the experimenter needs only to have access to a database of HRTFs. Many HRTF datasets are free and publicly-available (for example, *Algazi and Duda* (2001)). Subjective selection is a desirable method because it eliminates the need for individualized HRTFs, without compromising the quality of the spatial auditory image. Subjective selection satisfies a common goal in this research area: to reduce expensive measurement procedures and substitute them with simple HRTF selection procedures. This procedure leads to greater accessibility and better overall experience of spatial audio (*Roginska et al.* (2010a)).

3.1.1 Prior work in subjective selection

As mentioned in Chapter II, *Seeber and Fastl* (2003), investigated whether subjective selection could be used to improve localization performance when using non-individually measured HRTFs. The HRTFs were from the AUDIS-catalogue of human HRTFs (*Blauert et al.* (1998)).

The experimental procedure consisted of two parts: preselection and final selection. In preselection, listeners evaluated the overall directional impression of sounds generated using 12 sets of HRTFs. The sounds were judged according to externalization quality, minimization of front-back confusions, match of presented and perceived direction, and minimization of perceived source width. The five HRTFs that generated the highest scores were advanced to the final selection phase of the experiment. In this final phase, 1 of the 5 identified HRTFs was selected according to the perceived location, whether the source moved horizontally in equal increments, whether the elevation of the sound remained constant, whether the source was in the frontal plane, whether the sound is perceived at a constant distance and from the head, and whether that distance is heard as far away.

Seeber and Fastl (2003) found that subjective selection minimized the variance of localization responses and increased the degree to which sources appear externalized. Direct access and manipulation of HRTFs was a superior selection method as compared to experimenter-controlled sequential presentation, as the former allowed the subject to directly compare two sounds back-to-back and focus on small acoustic differences.

In a follow-up study to *Seeber and Fastl* (2003), our research group investigated subjective selection of HRTFs to improve the awareness of azimuth and elevation when using non-individually measured HRTFs (*Roginska et al.* (2010a)). For comparison, each listener's own set of HRTFs was included among the options. Twenty six of the twenty eight HRTFs used during subjective selection were from the IRCAM

and CIPIC databases. These twenty six sets of HRTFs are a subset of the IRCAM and CIPIC databases, and were selected by finding those with the greatest spectral disparity across HRTFs that belong to the same cone of confusion. One was measured on KEMAR and one was the listener's directly measured HRTF. The stimulus was a 500ms infrapitch signal made of a repeated pink noise burst.

The experiment was divided into two parts: HRTF measurement and a selection test. The listener's HRTF was measured using the HeadZap measurement system (*Anderson et al.* (2006)) and the ITD was extracted. For each HRTF in the database, a new HRTF was created by cascading the minimum-phase section of the given HRTF with the listener's ITD. The listening test consisted of three stages where listeners judged the externalization, elevation distinction, and front/back discrimination. Externalization was defined operationally by asking whether the listener heard a difference between a spatially-rendered sound and a diotic version of that sound. The latter was constructed by averaging the left and right channels of the rendered sound. To the extent that listener's heard a difference between the two modes of presentation, externalization was inferred. Similarly, elevation was defined operationally by asking whether the listener heard a difference in elevation between two sources that shared the same azimuth but were located at elevations reflected above and below the horizontal plane. Finally, front/back discrimination was inferred by asking whether the listener heard one source to be in front and a second source behind, as opposed to both sources coming from either in front or behind. The selection procedure began by judging against elevation. If the listener rejected a particular set of HRTFs at least twice in three trials, the set of HRTFs was eliminated from the search space. A similar culling was performed following the elevation phase. Thus, those sets of HRTFs that remained following the front/back discrimination phase were judged acceptable according to all three criteria.

The study found that most listeners preferred their own individualized HRTF in

at least two of the three experiment stages. Listeners also preferred many of the other sets of HRTFs in the database as often as their individualized HRTF. The results also suggest that many listeners judge some sets of HRTFs acceptable.

Roginska et al. (2010a) found that listeners found other HRTFs besides their own to be equally acceptable under the three criteria and concluded that VAE listeners can have good auditory perception, using subjectively-selected HRTFs.

3.1.2 The present work

Although listeners selected sets of HRTFs besides their own, the method used by *Roginska et al.* (2010a) for creating HRTFs required the listener's ITD function. The present study investigates whether listeners can select ITD functions as well. Specifically, the method proposed in *Roginska et al.* (2010a) was modified by substituting the ITD function measured on KEMAR (*Cheng and Wakefield* (2001)), which provides an ITD for an average human. From the selected sets of HRTFs, one was selected and listeners repeated the search procedure over the database of ITD functions. As in the original study, we are interested in whether a listener finds any set of HRTFs acceptable, and, if so, whether they also show a preference for particular ITD functions.

Roginska et al. (2010a) observed there were groups of subjects that preferred distinct sub-groups of spectral colorations. The present work seeks to determine if there are groups of subjects who prefer distinct sub-groups of ITDs as well. *Hartmann and Wittenberg* (1996) and *Middlebrooks* (1999) found that listeners tended to prefer HRTFs from subjects with slightly larger heads (larger ITDs). Perhaps we can expect that the listeners of this study will prefer the spatialization cues of their preferred HRTF with the ITD cues from a larger head. It is possible that all of the listeners could prefer the KEMAR ITD, as it was the ITD used while choosing their preferred spectral coloration.

Two experiments (spectral coloration and ITD selection) were designed to investigate the aforementioned questions. In the first experiment, listeners selected their preferred spectral coloration using spatial judgment tasks for externalization quality, elevation discrimination, and front/back discrimination. Repeated measures of experiment 1 were performed to determine the selection reliability. In experiment 2, one of the listener's preferred sets of HRTFs was used and the search was conducted over the sets of ITDs from the HRTF databases. The listener performed the spatial judgment tasks to indicate their preferred ITD. The preferred spectral coloration and ITD were combined to create each listener's customized HRTF. *Roginska et al.* (2010a) observed that listeners preferred a handful of HRTF sets as often as their own. The present study runs the selection task twice and compares results across the two runs to provide an initial assessment of reliability. It is possible that listeners may prefer a different subset of HRTF colorations after repeating the experimental measures.

3.2 Experiment 1: Spectral Coloration Selection

The first experiment, *spectral coloration (SC) selection*, replicated *Roginska et al.* (2010a) with the exception that the individually measured ITD was replaced by the KEMAR ITD and the listener's HRTF was not included in the search space. As in the original study, listeners judged the quality of the spatial rendering according to externalization, elevation differentiation, and front/back differentiation.

3.2.1 Participants

Four women and eleven men ranging from 20 to 43 years old participated in the study. All were undergraduate students at the University of Michigan. The participants were paid $10 per hour. Each participant underwent audiometric screening to make sure their hearing thresholds were within normal range. Before participating,

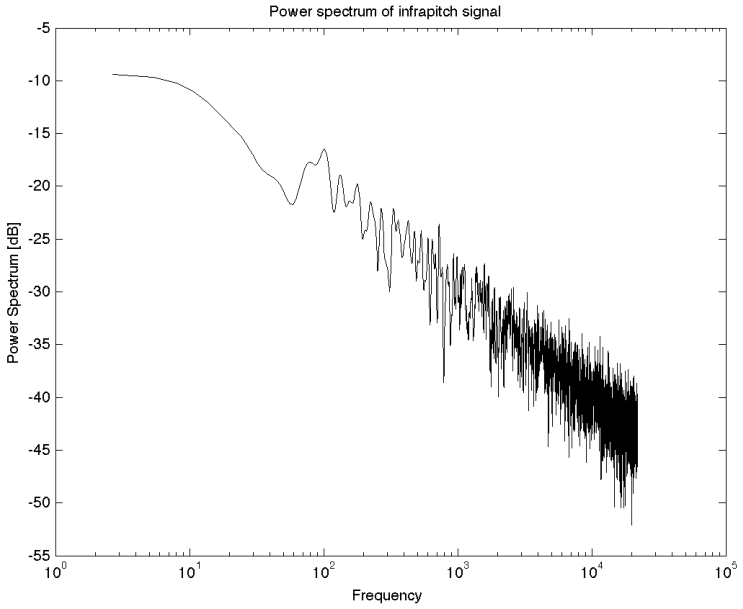


Figure 3.1: Infrapitch stimulus used for preference judgments

each listener read and signed a consent form (Appendix A).

3.2.2 Apparatus

The graphical user interface used in this study was developed using MATLAB. The interface was used to control all phases of the experiment. Before beginning the experiment, the HRTFs datasets were loaded into the system. The experiments were rendered on a 21" iMac system with an Apogee Duet audio interface. All stimuli were delivered using open circumaural Beyerdynamic DT 931 headphones.

3.2.3 Stimuli

A 500-ms infrapitch noise was used as the stimulus. The infrapitch noise was constructed by sampling 200 msec of a pink noise source and repeating the signal 2.5 times. With this period, listeners were able to hear the characteristic small-temporal spectral structures that are associated with infrapitch noise sources (*Warren and Bashford* (1981)). The digital pink noise generator was randomly seeded each time

an infrapitch noise was generated. Figure 3.1 shows the power spectrum of one instantiation of the infrapitch signal that was used in the experiment.

The HRTF datasets used in the present experiment were collected from two publicly-available databases: the database obtained at IRCAM (Institut de Recherche et Coordination Acoustique/Musique) and AKG Acoustics for the Listen project (http://recherche.ircam.fr/equipes/salles/listen) and the CIPIC (Center for Image Processing and Integrated Computing) database measured at UC Davis by *Algazi and Duda* (2001). Additionally, a dataset measured at the Naval Submarine Medical Research Laboratory (NSMRL) on KEMAR by *Cheng and Wakefield* (2001) was included in the database. The IRCAM database contains 51 HRTF and anthropometric subject measurements. The CIPIC database contains 45 HRTF and anthropometric measurements. A measure of spectral contrast between HRTFs along a common cone of confusion was used to eliminate HRTFs that were unlikely to provide sufficient acoustic information for listener's to discriminate front from back (*Roginska et al.* (2010a)). From the 97 sets of HRTFs, 27 were selected: 13 from the IRCAM database, 13 from the CIPIC database and the KEMAR dataset. For each trial, a given set of HRTFs was selected and impulse responses for the left and right ears were created by cascading the minimum-phase head-related impulse response (drawn from the given set) with the all-pass impulse responses for the KEMAR ITD.

3.2.4 Procedure

All listening tests were conducted in a soundproof booth in the Computer Science and Engineering Building at the University of Michigan. Each participant was seated at a table in front of the iMac system/Apogee Duet audio interface with the Beyerdynamic headphones in place.

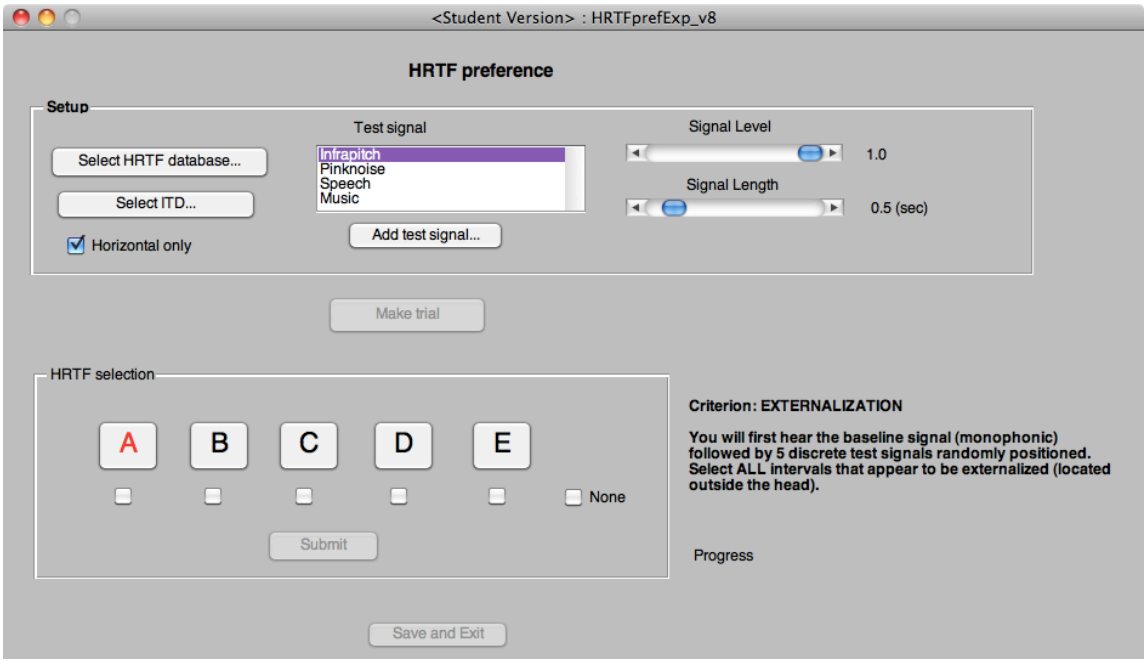


Figure 3.2: Interface seen during coloration selection tasks. A is the interval currently playing during the externalization stage.

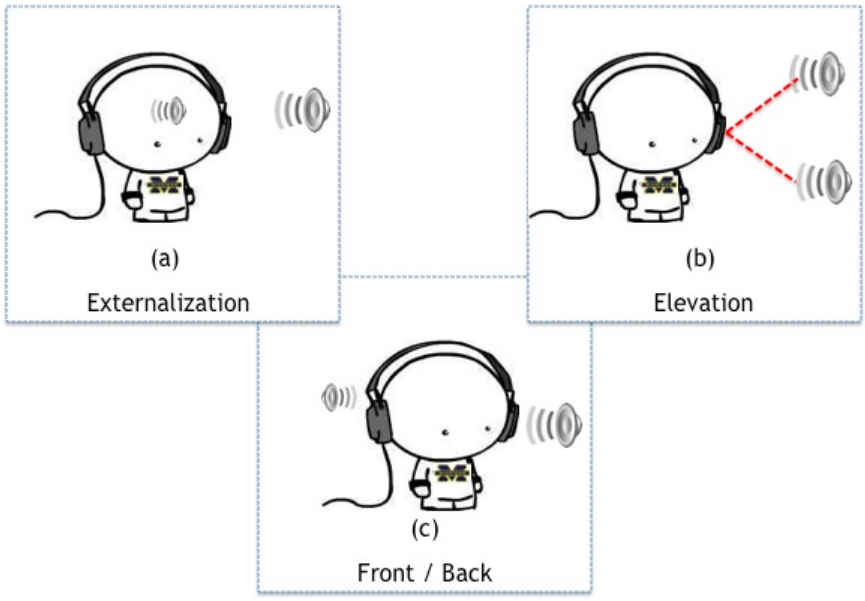


Figure 3.3: Stimulus preference judgments made during experimental procedures: externalization quality (a), elevation (b) and front/back (c) distinction

3.2.5 Methods

The user interface for controlling the experiment is shown in Figure 3.2. At the beginning of a session, the HRTF database, ITD, and test signal were selected. Following this selection, the listener began the first phase of the experiment in which they judged each rendering on the basis of externalization. Upon completion, the database was culled of sets that weren't selected two or more times, and the elevation phase was begun. Similarly, upon completion, the database was culled and the front/back phase was begun. To remind the listener, instructions on the screen included a definition of the criterion and a description of the listener's task.

Each phase of the experiment consisted of a set of trials, each trial of which presented examples of five different renderings. At the beginning of a trial, the listener heard each example in sequence. The visual display was used to cue the listener by highlighting the appropriate option button while the example was played. Before making their judgment, the listener could replay any one option by selecting the appropriate button. Check boxes below each option button were used to indicate which options provided adequate externalization (Phase One), elevation distinction (Phase Two), or front/back distinction (Phase Three).

In the case of externalization (3.3a), the beginning of each trial was preceded by an unspatialized reference signal. Each interval was comprised of a series of sounds. The first sound in the interval was an unspatialized reference signal. This signal served as an in-the-head reference, to which the externalized sounds could be compared. Following the reference signal, the listener heard a series of five spatialized signals that were generated from randomly selected HRTFs at randomly selected azimuths on the horizontal plane: $\pm 150^\circ$, $\pm 120^\circ$, $\pm 90^\circ$, $\pm 60^\circ$, $\pm 30^\circ$. The same sequence of azimuths was used for all intervals in each trial. The listener checked the boxes of the intervals in which externalization was perceived. In some cases, sound sources in the same interval were perceived as externalized and others were not. The listener

was instructed to select the cases in which a majority (three or more) of the sounds was perceived as externalized. If none of the intervals were perceived as externalized, the listener checked the "None" checkbox. After submitting their selections, the results were saved and the overall completion progress was displayed. Next, the listener repeated the externalization discrimination task for a new set of five intervals. Each HRTF in the database appeared in three intervals. Only the preferred HRTFs (selected at least two out of three times) were used in the elevation discrimination task.

The elevation discrimination phase (Figure 3.3b), proceeded along lines similar to the externalization phase. Each interval was comprised of five pairs of stimuli. Each pair was spatialized using a randomly selected HRTF at a randomly-selected azimuth: $\pm$ 150° , $\pm$120°, $\pm$90°, $\pm$60°, $\pm$30° at $\pm$36° elevation. The same sequence of azimuths was used for all intervals in each trial. The listener checked the boxes of the intervals in which they could discriminate the high and low signals in each pair of stimuli. The listener was instructed to select the intervals in which a majority (three or more) of the signal pairs had discriminable elevation differences. If none of the intervals contained pairs in which elevation could be discriminated, the listener checked the "None" checkbox. After submitting their selections, the results were saved and the overall completion progress was displayed. The trials continued in this manner until each set of HRTFs in the database had been evaluated three times. The database was then culled and only those sets of HRTFs that were selected two or more times were advanced to the front/back phase.

The front/back discrimination task (Figure 3.3c) required the listener to judge the front/back distinctiveness of two sounds presented at locations along a common cone of confusion. Each interval was comprised of five pairs of stimuli presented on the horizontal plane. The pairs were spatialized using randomly selected HRTFs from the following azimuths: $\pm$ 150° , $\pm$120°, $\pm$90°, $\pm$60°, $\pm$30°. The same sequence of

azimuths was used for all intervals in each trial. Upon completion of the third phase, the sets of HRTFs that remain after culling of the database have passed the listener's judgment with respect to externalization and two attributes of localization: up/down and front/back.

Each of the 27 HRTFs were ranked with a number that indicated the number of subjects that preferred the HRTF's spectral coloration. The highest-ranking HRTF coloration within a subject's set of preferred colorations was selected as their preferred coloration. In the event of a tie, a spectral coloration was randomly selected from the listener's highest-ranking colorations.

3.3 Results

In the figures that follow, HRTFs #2 - #14 come from the IRCAM database, #15 - #27 come from the CIPIC database, and #28 was the KEMAR HRTF measurement. HRTF #1 is deliberately blank, to allow direct comparison to the results of *Roginska et al.* (2010a), in which it represented the listener's measured HRTF.

3.3.1 Externalization

Figure 3.4 shows the percentage of times that each HRTF of the 27 HRTFs made it through the externalization judgment phase. The CIPIC datasets and the KEMAR dataset were chosen by 86-100% of subjects for their externalization quality. In sharp contrast, the IRCAM datasets were selected by 6-26% of subjects. The selection results follow the trend observed in *Roginska et al.* (2010a) (Figure 3.5). The CIPIC datasets and the KEMAR dataset were chosen by 70-80% of subjects for their externalization quality. In contrast, the IRCAM datasets were selected by 10-40% of subjects.

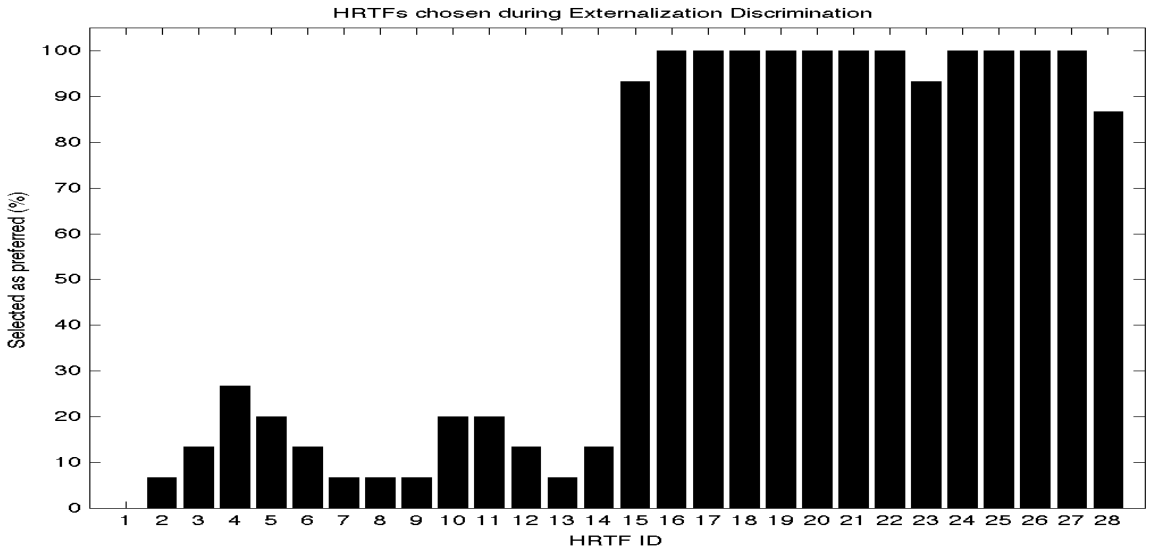


Figure 3.4: Percentage of subjects that selected each HRTF during the externalization discrimination stage of HRTF spectral coloration selection. Along the abscissa is the HRTF ID and along the ordinate is the percentage of participants that perceived each HRTF's externalization cues.

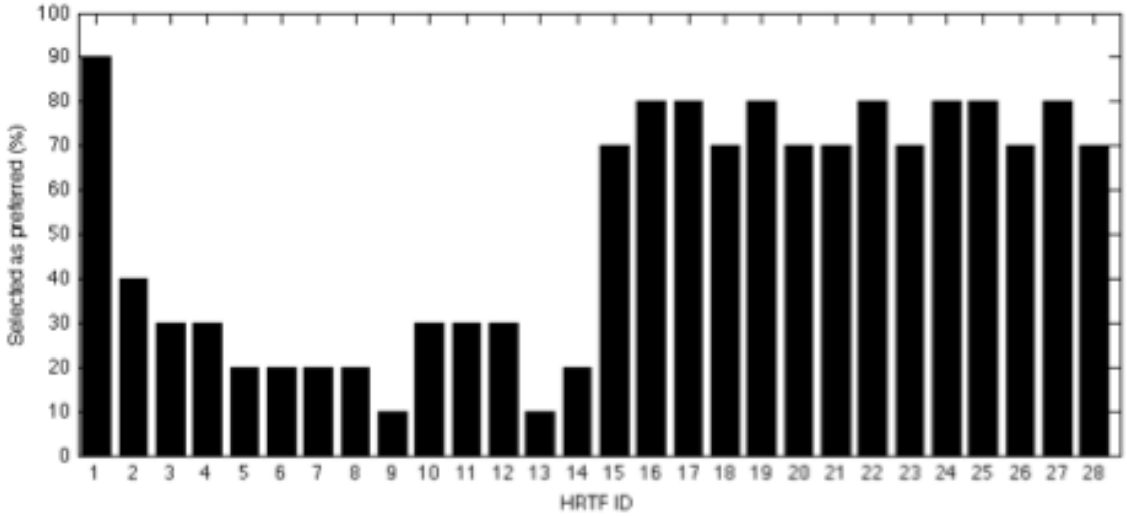


Figure 3.5: Percentage of subjects that selected each HRTF during the externalization discrimination stage of *Roginska et al.* (2010a) (reprinted with permission).

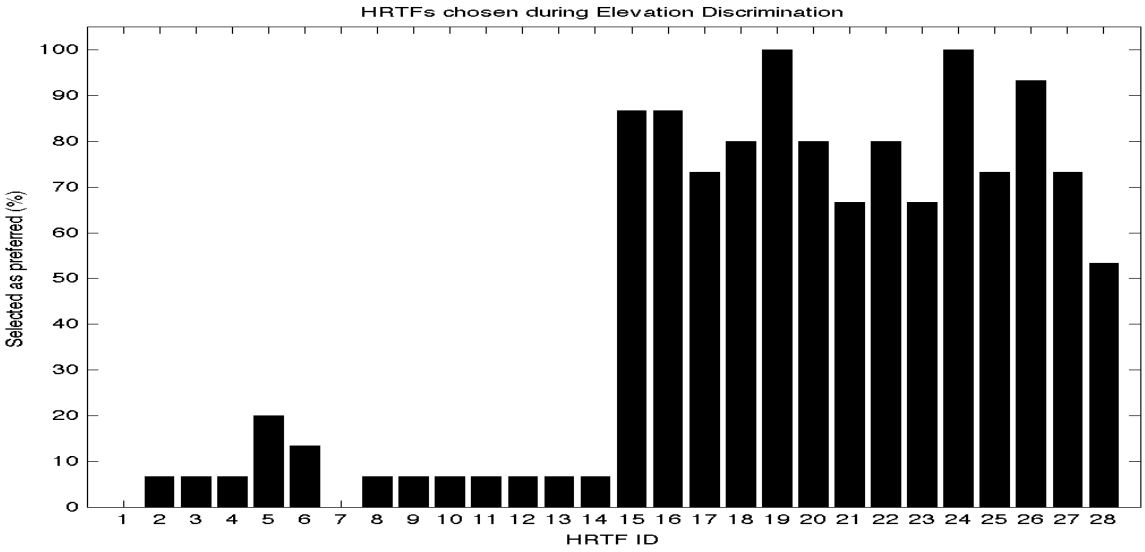


Figure 3.6: Percentage of subjects that selected each HRTF during the elevation discrimination stage of HRTF spectral coloration selection. Along the abscissa is the HRTF ID and along the ordinate is the percentage of participants that perceived each HRTF's elevation cues.

3.3.2 Elevation

In the second stage of SC selection, subjects were asked to discriminate between examples that were rendered at elevations above and below the horizontal plane. Figure 3.6 shows the percentage of subjects that chose each of the 27 HRTFs presented during the elevation discrimination task. Only HRTFs that were chosen in at least 67% of the presentations continued to the next stage . The CIPIC datasets and the KEMAR dataset were chosen by 53-100% of subjects for their elevation quality. In sharp contrast, the IRCAM datasets were selected by 0-20% of subjects. Among the subjects for which #7 was advanced after the externalization phase, none of them judged the examples as good two or more times during the present phase.

Our results follow the trend observed in the previous work (Figure 3.5) in that a higher percentage of subjects preferred the spectral colorations of the CIPIC and KEMAR HRTFs. A smaller percentage of subjects preferred the IRCAM HRTFs. In the present work, HRTF #7 was eliminated at this stage, as it was not preferred by

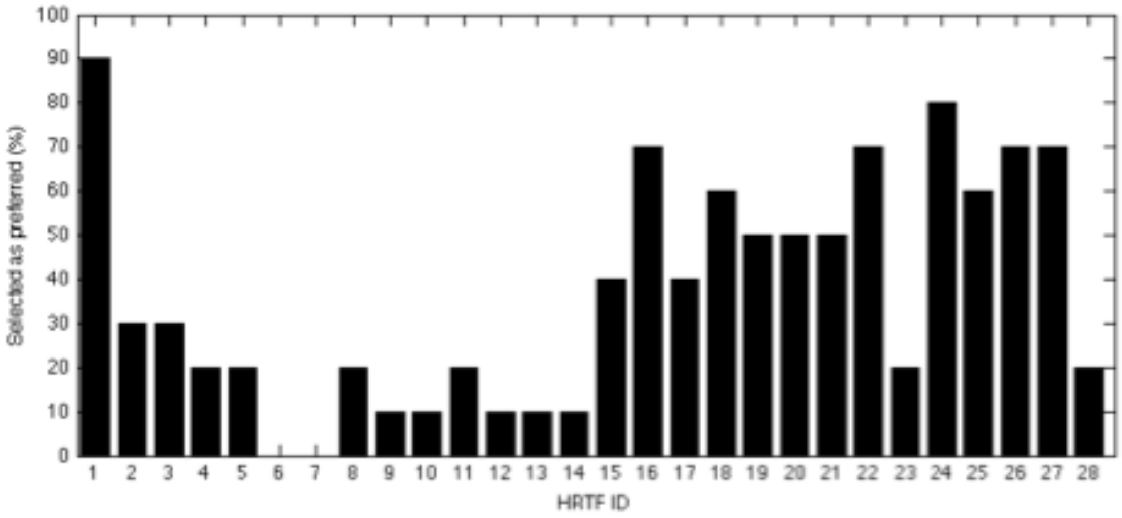


Figure 3.7: Percentage of subjects that selected each HRTF during the elevation discrimination stage of *Roginska et al.* (2010a) (reprinted with permission).

any listeners. In the previous work, the same HRTF was also eliminated at this stage.

3.3.3 Front/Back

Figure 3.8 shows the percentage of subjects that chose each of the 27 HRTFs presented during the front/back phase. The CIPIC datasets and the KEMAR dataset were chosen by 20-86% of subjects for their front/back discernibility. The IRCAM datasets were selected by only 0-20% of subjects. An additional HRTF (#9) from the public datasets has been eliminated at this stage.

As compared to the results of our previous work (Figure 3.9), fewer HRTFs were eliminated at this stage of the listening test. In the present study, HRTF #9 was eliminated, as also seen at this stage in the previous work.

The previous figures have shown data averaged across subjects. Figure 3.10 summarizes the selection process for each subject (ordinate) for each HRTF set (abscissa). The absence of a symbol indicates that the subject did not select that HRTF set two or more times during the externalization phase. The 'x' marker represents externalization selections. HRTFs selected after the externalization and elevation discrimination

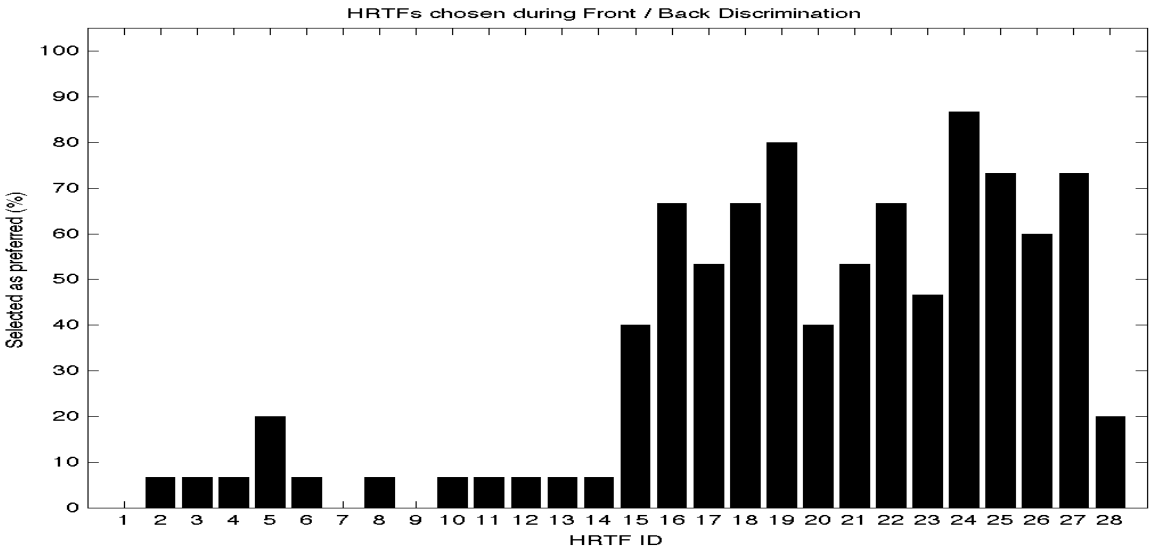


Figure 3.8: Percentage of subjects that selected each HRTF during the front/back discrimination stage of HRTF spectral coloration selection. Along the abscissa is the HRTF ID and along the ordinate is the percentage of participants that perceived each HRTF's front/back cues.

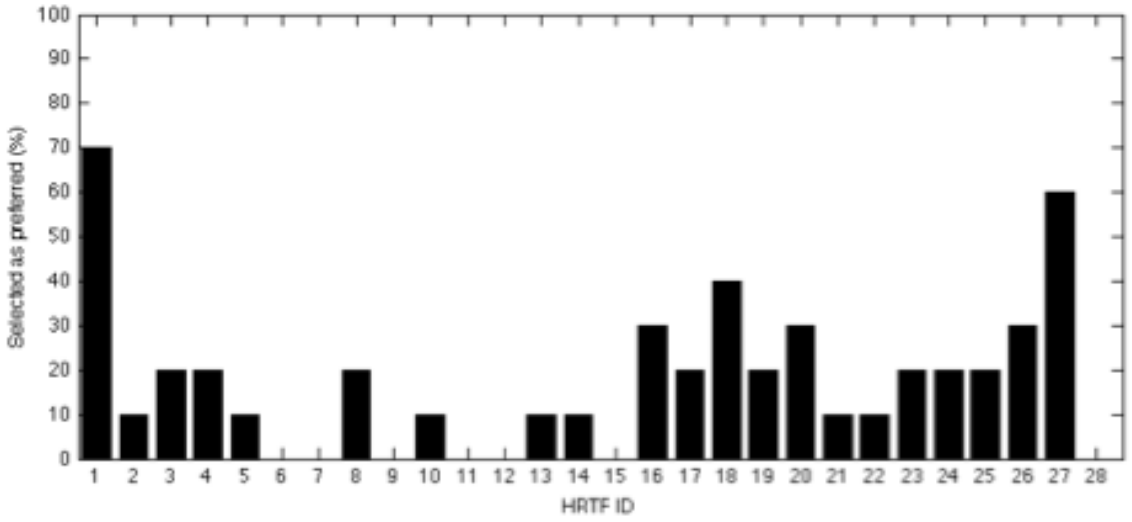


Figure 3.9: Percentage of subjects that selected each HRTF during the front/back discrimination stage of *Roginska et al.* (2010a) (reprinted with permission).

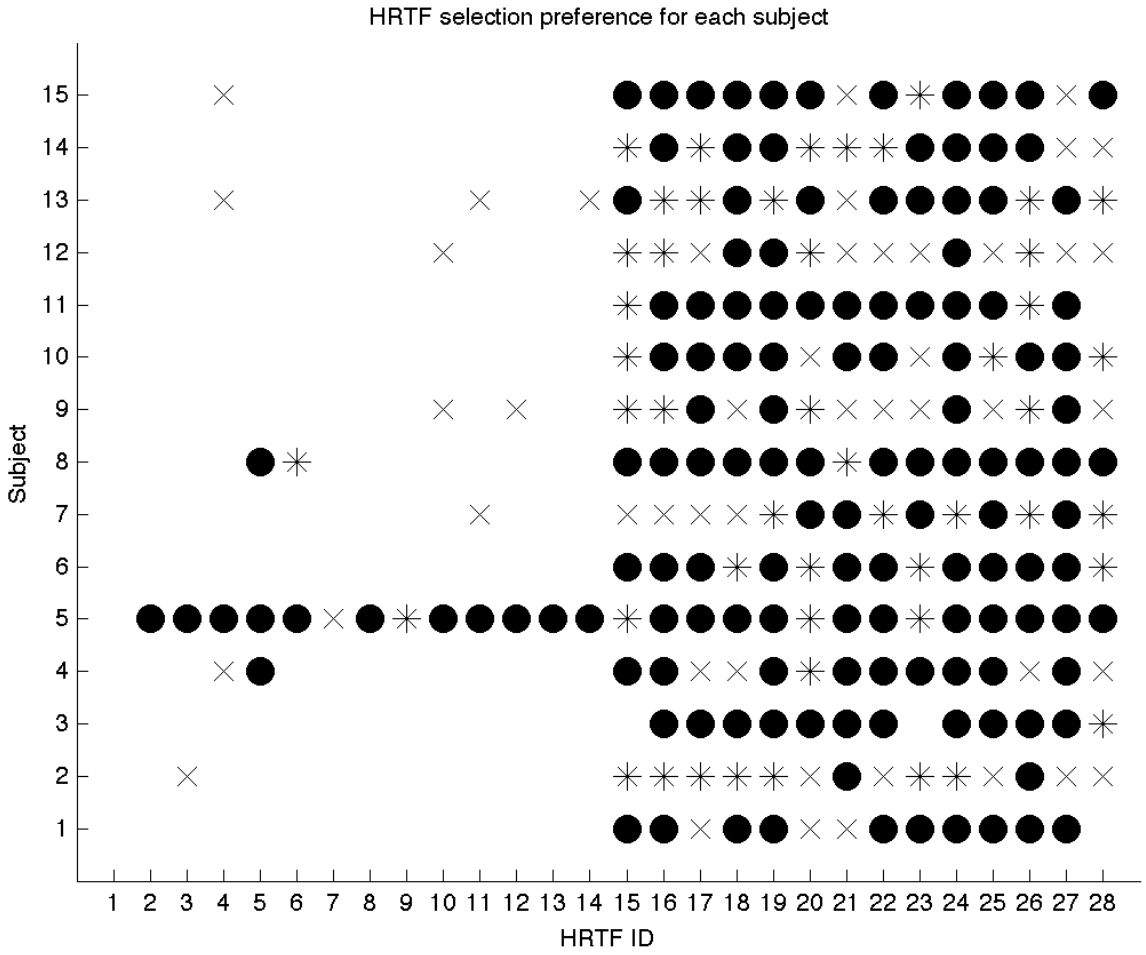


Figure 3.10: HRTF spectral coloration preference for each subject, for the 3 judgment tasks: externalization (x), elevation discrimination (*), and front/back discrimination (filled circles). Along the abscissa is the HRTF ID and along the ordinate is the Subject ID.

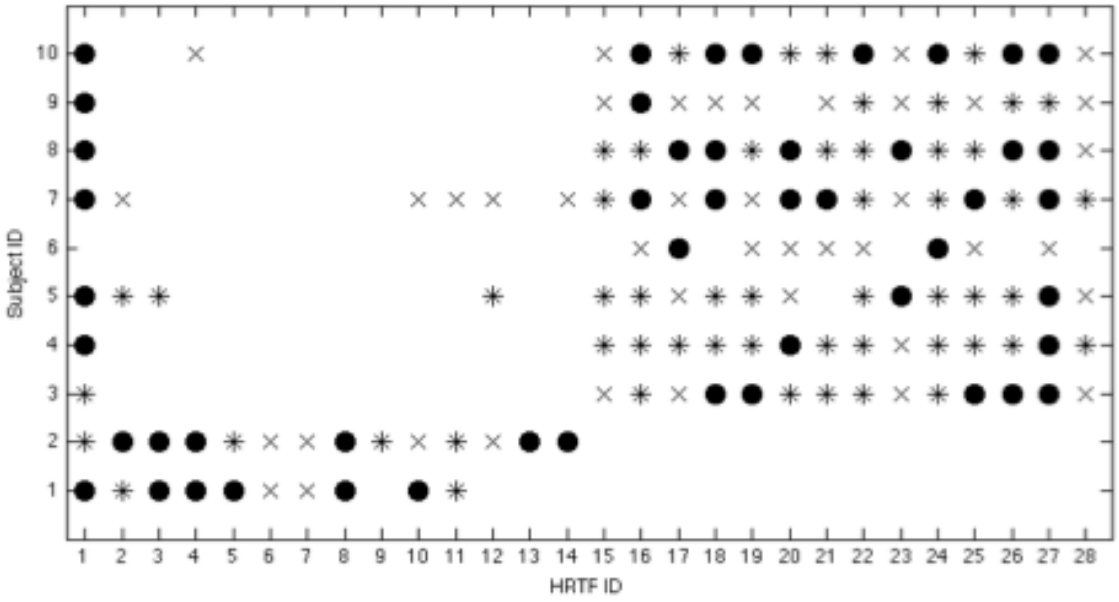


Figure 3.11: HRTF spectral coloration preference for each subject of *Roginska et al.* (2010a) (reprinted with permission).

tasks are represented by the '*' marker and final winners are presented by filled circles. Selection results are very similar across subjects. All of the subjects had a strong preference for HRTFs from the CIPIC database and the KEMAR measurements. In addition, subjects #4, #5 and #8 preferred at least one HRTF from the IRCAM database. In the previous work (Figure 3.11), only two listeners preferred the HRTFs of the IRCAM database; however, in the present study, none of the listeners preferred the IRCAM HRTFs.

3.3.4 Repeated Measures

To determine the reliability of the listener's judgment during the discrimination tasks, eleven participants repeated the SC selection experiment. Figure 3.12 shows the two measures of reliability (*selectivity* and *consistency*). Selectivity (Equation 3.1) is the number of HRTF sets that were selected in both experiments normalized by the number of HRTFs selected in either experiment.

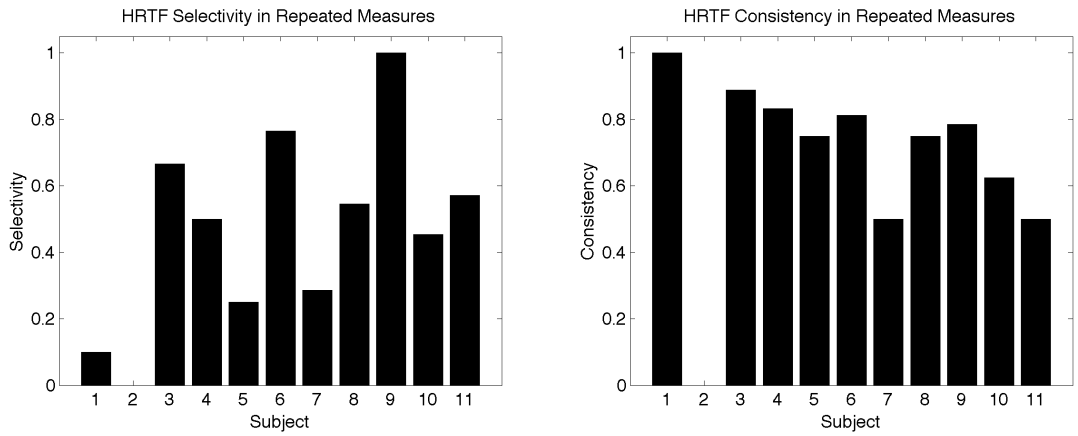


Figure 3.12: On the left is the percentage of HRTFs preferred by each subject during the repeated measures task, compared to all HRTFs preferred in both experiments. On the right is the percentage of HTRFs preferred by each subject during the repeated measures task, as compared to the HRTFs preferred in the initial task.

$$Selectivity = \frac{\mid H_{trial1} \cap H_{trial2} \mid}{\mid H_{trial1} \cup H_{trial2} \mid} \tag{3.1}$$

Consistency (Equation 3.2) compares the HRTFs that were selected in the second experiment to the HRTFs selected in the initial procedure.

$$Consistency = \frac{\mid H_{trial1} \cap H_{trial2} \mid}{\mid H_{trial2} \mid} \tag{3.2}$$

The left panel of Figure 3.12 shows listener selectivity as a function of subject ID. Selectivity ranges from 0 to 1 across the eleven subjects, with a mean selectivity of 0.47. This relatively low value for average selectivity could reflect the fact that listeners make very different choices when performing the task a second time, or that they are becoming more discerning during the second run by accepting some but rejecting more of the choices they made in the first run. The consistency score suggests the latter, rather than the former, explanation for the selectivity results. Consistency scores across subjects are no smaller than 0.5 and show a mean of 0.68. It should be noted that subject #2 has a 0% selectivity and consistency score because

he/she selected two HRTFs in the first task and one (different) HRTF in the repeated measure. The data suggest that listeners, on average, are likely to repeat 70% of their selections when searching through the space of options a second time. While this result is encouraging, stronger conclusions require substantially more data across multiple repetitions of the experiment.

3.4 Experiment 2: ITD Selection

3.4.1 Methods

In the second experiment, *ITD selection*, the listener chose their preferred ITDs by judging the spatial qualities of sounds delivered using different ITDs from a database of HRTFs. The preferred spectral colorations, as determined in Experiment 1 (SC selection), were used to deliver the spatialized sounds. To pick the preferred spectral coloration, the HRTFs were given a numerical score according to the number of subjects that preferred it in the final stage. Of each listener's final set of colorations, The highest scoring ITD was chosen as the listener's preferred spectral coloration. In the case of a tie, one of the highest scoring HRTFs was randomly chosen. In the same manner, a preferred ITD was determined for each participant of the present experiment. Each listener's preferred spectral coloration and preferred ITD were combined to create the set of HRTFs for that listener throughout the rest of the dissertation experiments.

3.4.2 Participants, Apparatus and Stimuli

The fifteen listeners from the previous experiment participated in Experiment 2. The entire experiment was completed in a single session that occurred one to seven days after Experiment 1. The same 500ms infrapitch signal and HRTF datasets from Experiment 1 were used in Experiment 2. Each listener heard all stimuli through

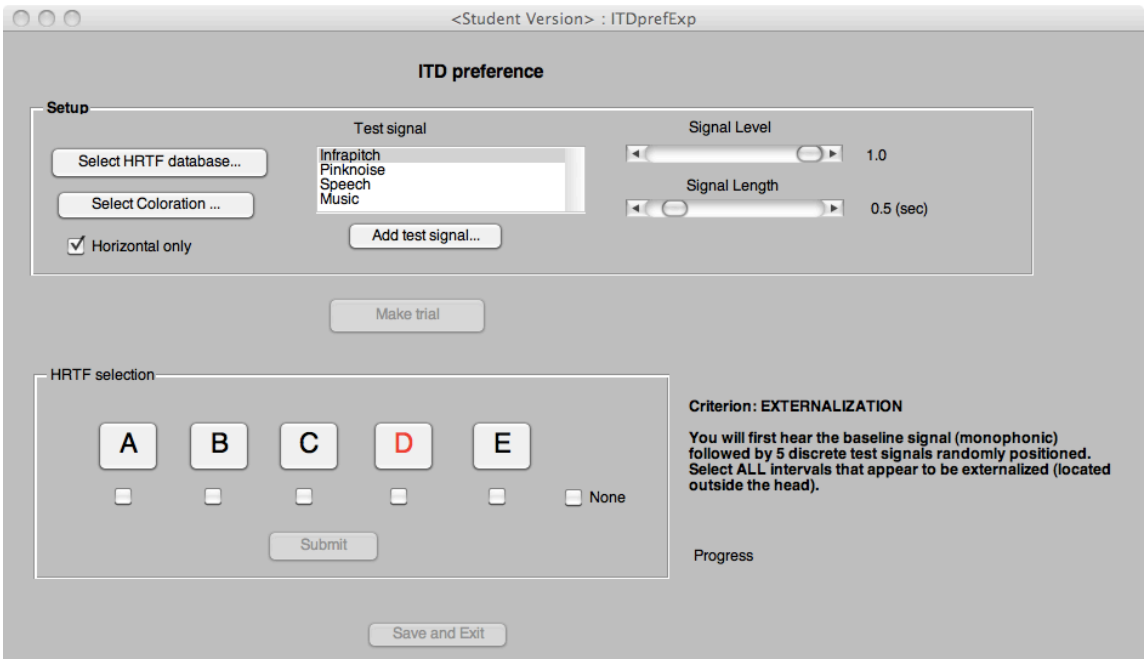


Figure 3.13: Interface seen during ITD selection tasks. D is the interval currently playing.

their preferred spectral coloration.

3.4.3 Procedure

In ITD selection, each listener's preferred ITD was selected from among the ITDs within the HRTF database. At the beginning of the ITD selection experiment (Figure 3.13), the experimenter began by loading the database of HRTFs into the system. Next, the experimenter loaded each listener's preferred spectral coloration into the system. Afterwards the listener pressed a button to begin the trial in which they completed the 3 stages of discrimination tasks in the same manner as that described in Experiment 1.

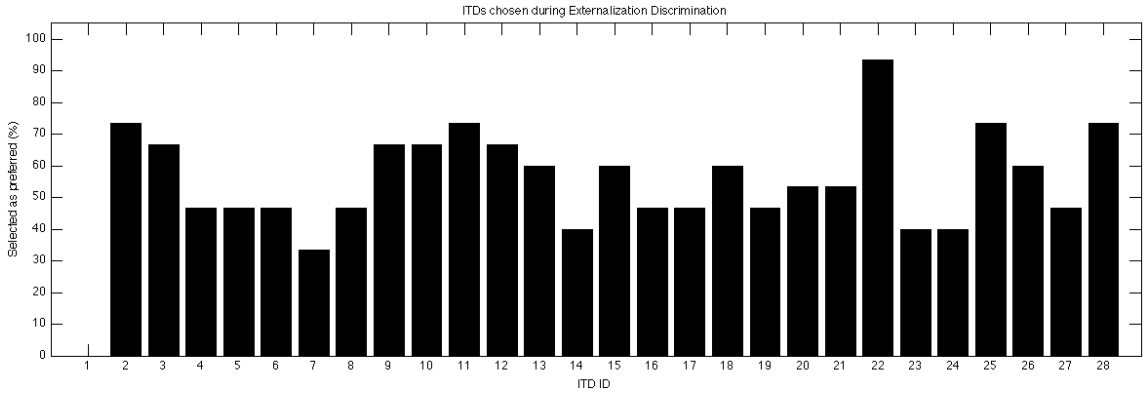


Figure 3.14: Percentage of subjects that selected each ITS during the externalization discrimination. Along the abscissa are the HRTF IDs and along the ordinate is the percentage of participants that selected the ITD of that HRTF

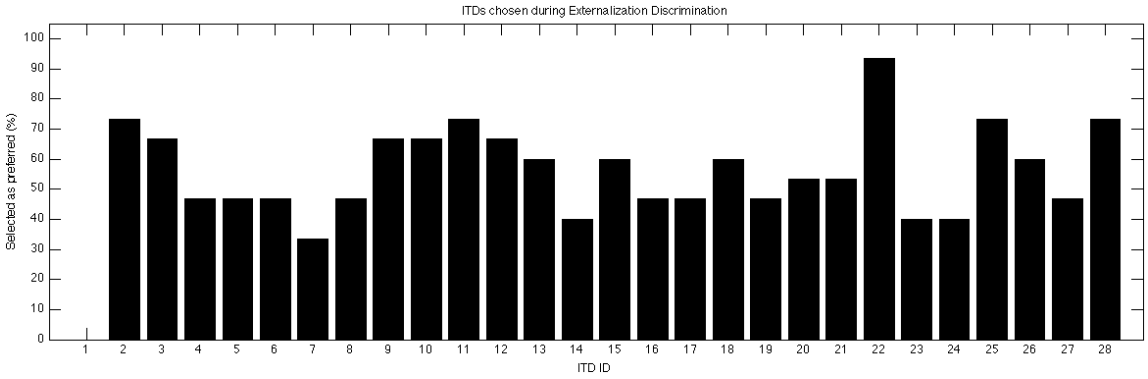


Figure 3.15: Percentage of subjects that selected each HRTF during the elevation discrimination stage of ITD selection. Along the abscissa are the HRTF IDs and along the ordinate is the percentage of participants that perceived each HRTF's externalization cues.

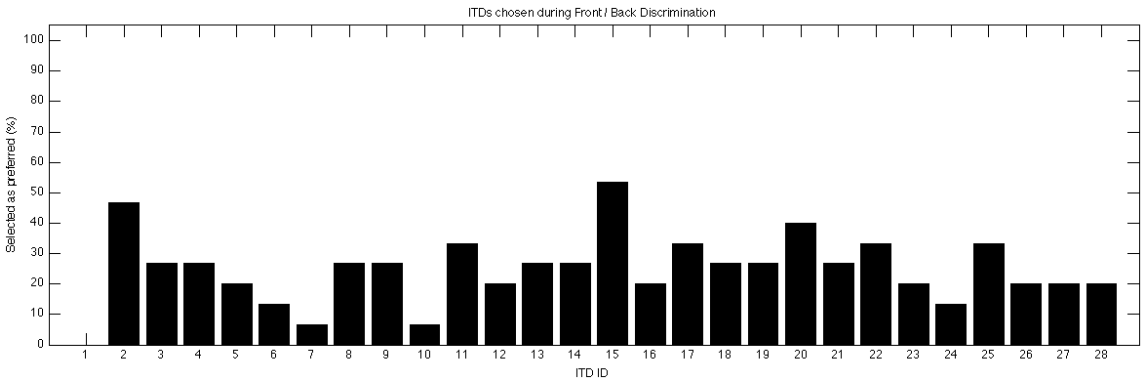


Figure 3.16: Percentage of subjects that selected each HRTF during the front/back discrimination stage of ITD selection. Along the abscissa are the HRTF IDs and along the ordinate is the percentage of participants that perceived each HRTF's externalization cues.

3.5 Results

3.5.1 ITD selections

The results of ITD selection differ considerably from those shown for selection based on spectral coloration. Figures 3.14-3.16 show the percentage of subjects that chose each of the 27 HRTF sets for the externalization, elevation and front/back stages of the selection process, respectively. In contrast to the choice of spectral coloration, there is relatively little clustering in selection around particular collections of HRTF sets. Some selectivity is achieved, and this selectivity improves with each subsequent phase. The average selection rate for externalization was 57%. This decreased to 42% after the elevation phase and was further reduced to 26% upon completion of the third phase. Thus, listeners are able to reject options at each stage, but their individual choices do not agree with those of the entire group.

When broken out by individual subject, the same differences are noted between the ITD selections and the spectral coloration selections. As shown in Figure 3.17, the 'x' marker represents items that passed externalization before being rejected, the '*' marker represents items that passed both externalization and elevation before being

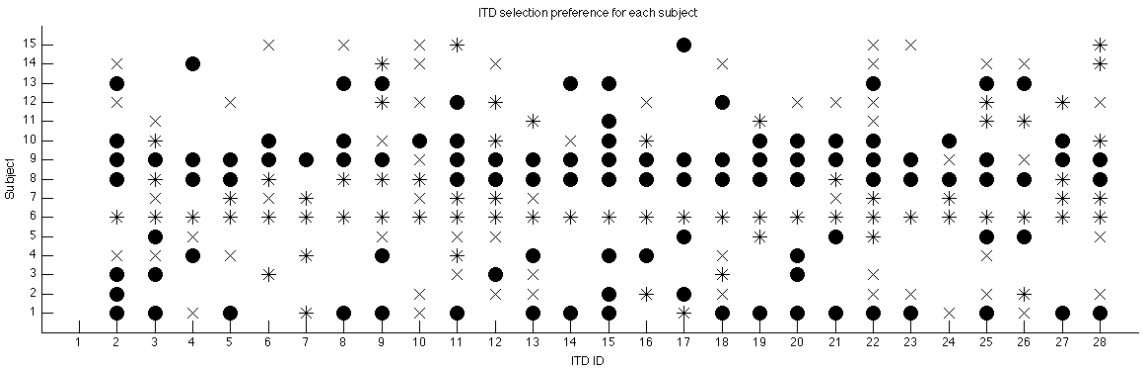


Figure 3.17: ITD preference for each subject, for the 3 judgment tasks: externalization (x), elevation discrimination (*), and front/back discrimination (filled circles). Along the abscissa are the HRTF IDs and along the ordinate are the Subject IDs.

rejected, and the filled circle represents items that passed all three criteria. Selection results are very similar across subjects. None of the subjects exhibited a strong preference for a specific ITD or set of ITDs from the HRTF database. There is no single ITD that is rejected or accepted by most subjects. Furthermore, most subjects find a large number of ITDs to be acceptable under these criteria. The IRCAM ITDs were selected as often as the CIPIC ITDs.

3.5.2 Customized HRTFs

The HRTFs created in the two experiments are shown in Table 3.1. As in the SC selection procedure, the listener's preferred ITD was identified. Of the 15 customized HRTFs that were created, 12 were unique. One participant (#14) selected a spectral coloration and ITD from the same HRTF in the database.

3.6 Discussion

The results of the present work replicate the earlier findings of *Roginska et al.* (2010a) without the need for an individually-measured ITD. Repeating the selection procedure for spectral coloration suggests that listeners tend to reject previously

Table 3.1: HRTF IDs of the spectral coloration and ITD used to create each customized HRTF

Participant	Coloration	ITD
1	22	15
2	25	2
3	22	2
4	22	15
5	24	17
6	24	15
7	24	27
8	24	15
9	25	15
10	22	20
11	27	15
12	22	11
13	22	15
14	24	24
15	22	17

selected options rather than accept ones they had previously rejected.

Additionally, the results affirm that there are discriminable spectral cues that most listeners prefer over another, even when listening using standardized ITDs. Similar to the observations in *Roginska et al.* (2010a), there was a group of listeners that preferred the spectral colorations of the HRTFs from the CIPIC and KEMAR databases. None of the listeners preferred the IRCAM HRTFs. Furthermore, we were also able to identify common HRTFs that did not provide elevation and front/back distinction to any listener in the present study and *Roginska et al.* (2010a). These conclusions are important with respect to the use of spatial audio in the field. By establishing that the same pre-measured HRTFs are selected using a pre-measured ITD, we have shown that it is not necessary to individually measure the ITD for the listener in a practical customization procedure. Furthermore, we have shown that listeners do prefer some ITD sets over others, and that they can refine their preferences through the same three stages of evaluation.

The results indicated that each ITD had about an equal likelihood of being preferred in the discrimination tasks. Listeners, as a group, did not show preference for any particular subset of possible ITDs. It should be noted that participants informally reported that the ITD selection task was more challenging than the SC selection task, as there were smaller differences in the spatial cues.

CHAPTER IV

Virtual Auditory Search and Training

4.1 Introduction

Listeners need to be trained to search for sounds in a VAE. The ability to navigate in a VAE, while receiving real-time spatial audio cues, is a relatively new concept that may not be "natural" to listeners. Training is vital because spatial auditory displays depend critically on the user's ability to find the various sources of information (*Wenzel et al.* (1993)). In the same way that sight-impaired individuals require training to learn to navigate the physical world, listeners need training to search for sound sources in a virtual world.

The present chapter discusses the development of a training procedure to search for sounds in a VAE. Section 4.2 describes the search strategies that listeners develop independently, without receiving training. Successful independent search strategies were analyzed and used to design the training procedure described in section 4.3. The remainder of the chapter describes the experiment and evaluates the efficacy of the proposed training procedure.

4.1.1 Visually-impaired navigation training

To begin discussing non-sighted navigation of VAEs, we must consider a limiting case, orientation and mobility training for the visually-impaired. Orientation and

mobility (O&M) training aims to maintain travel independence by teaching visually-impaired adults to negotiate natural environments safely and independently (*Peterson et al.* (1998)). With training, visually-impaired adults gain a better understanding of their environment, which enables them to travel more comfortably, efficiently, and safely.

Sight-impaired individuals are often trained to use navigation tools called mobility aids or electronic travel aids (ETAs), which primarily provide near-field detection of immediate objects. Examples of such devices are the sonar-cane (*Kay* (2011)), laser cane (*Raycal* (2006)), and infrared signage (*Crandall et al.* (1999)). Most ETAs help to identify close or near-field objects. Blind travelers could potentially benefit from receiving a larger representation of their space, such that all objects within a predefined radius are detectable.

From the O&M training, we see that visually-impaired persons require training to learn to navigate the environment. From here, we may infer that sighted individuals will also need training to learn to navigate a VAE. Training to search for auditory objects has not been thoroughly investigated in the literature. Our approach examines the search strategies of successful navigators and characterizes their behavior. From the characterization, a training method was developed and analyzed.

4.2 Experiment: Self-trained search

When listeners are allowed to search for sounds in a VAE without training, they exhibit a variety of behaviors. In this section, we characterize the behaviors observed in two search experiments that were performed in collaboration with New York University (NYU). The first study examines head rotation search strategies. The second study examines changing-position search strategies within the VAE.

4.2.1 Search by head rotation

From a fixed position, head movement improves sound localization accuracy (*Wallach* (1940); *Pollack and Rose* (1967); *Wightman and Kistler* (1999); *Zahorik et al.* (2006)). For example, *Pollack and Rose* (1967) investigated the role of head motion in the localization of sounds presented on the horizontal plane. Localization accuracy improved in conditions where head motion was permitted, as compared to conditions in which the listener's head remained stationary. The average localization error was 10-15% less when head motion was allowed. Thus, listener movement improves localization accuracy; however, the strategy used during the localization process is unknown.

In a previous study (*Roginska et al.* (2011)), our research group assessed listener strategies when localizing a static sound source from a fixed position. In the study, 21 participants completed 3 interaction conditions:

1. *Static Interaction*: In the static condition, the listener did not interact with the VAE. The sound source was presented at a fixed location. The listener judged the location of the source without moving.

2. *Avatar Interaction*: In the avatar mediation, the listener used a mouse interface to change their orientation in the environment by orienting the nose of the avatar. As the orientation of the avatar was changed, the audio cues updated as well to reflect the relative change in orientation.

3. *Natural Interaction*: In the natural mediation, the listener used a head-tracker to interact with the environment. As the listener turned their head, the spatial audio cues updated to convey the location of the sound source.

Participants were told to locate a sound source on the horizontal plane and mark its location in the graphical user interface (GUI) (Figure 4.1). In a balanced de-

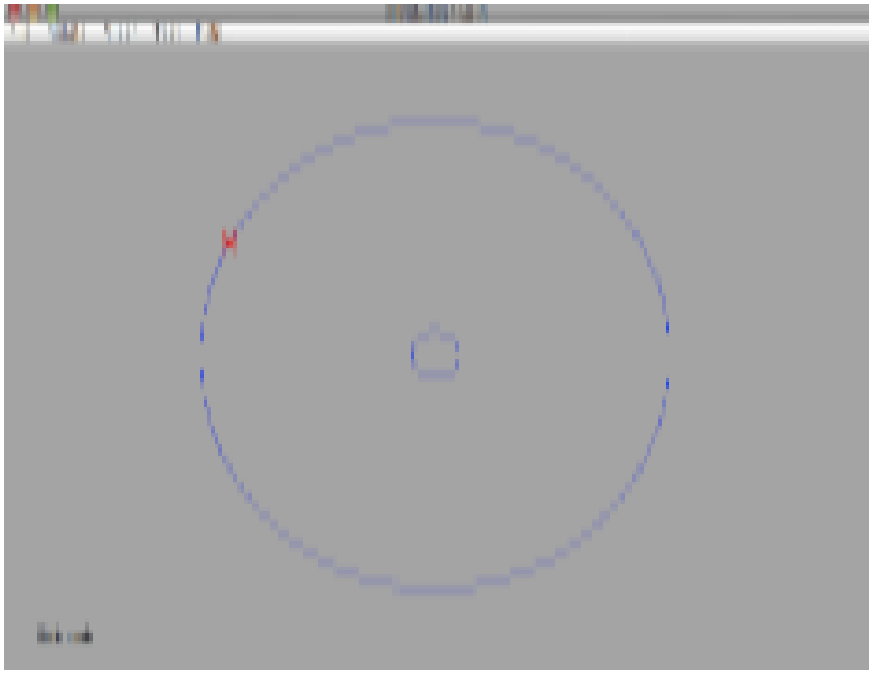


Figure 4.1: GUI used in fixed-position search to mark the location of the sound source (from *Roginska et al.* (2011)). The red 'X' indicates the marked location of the sound source. The listener's orientation is represented by the direction of the blue listener's "nose" in the center.

sign, each participant completed the three experimental conditions. Each condition consisted of a training phase followed by a testing phase.

In the orientation strategy analysis, each listener's three fastest and slowest trials were examined in each interaction condition. The static condition is omitted from this discussion because the listener's orientation remained constant. Each trial was categorized according to the strategies used to localize the sound source. For example, in 40% of the fastest avatar condition trials, listeners used the equal-level strategy (Figure 4.2). In the equal-level strategy, the listener rotates in the direction of the sound source and stops when the sound's intensity reaches an equal level in both ears.

In another 14% of the fastest avatar trials, listeners were observed overshooting the sound source by rotating in the direction of the target source, letting the sound source pass through the leading ear, then to the opposite ear. Then the listener corrected their position by moving so that the sound source was directly in front of

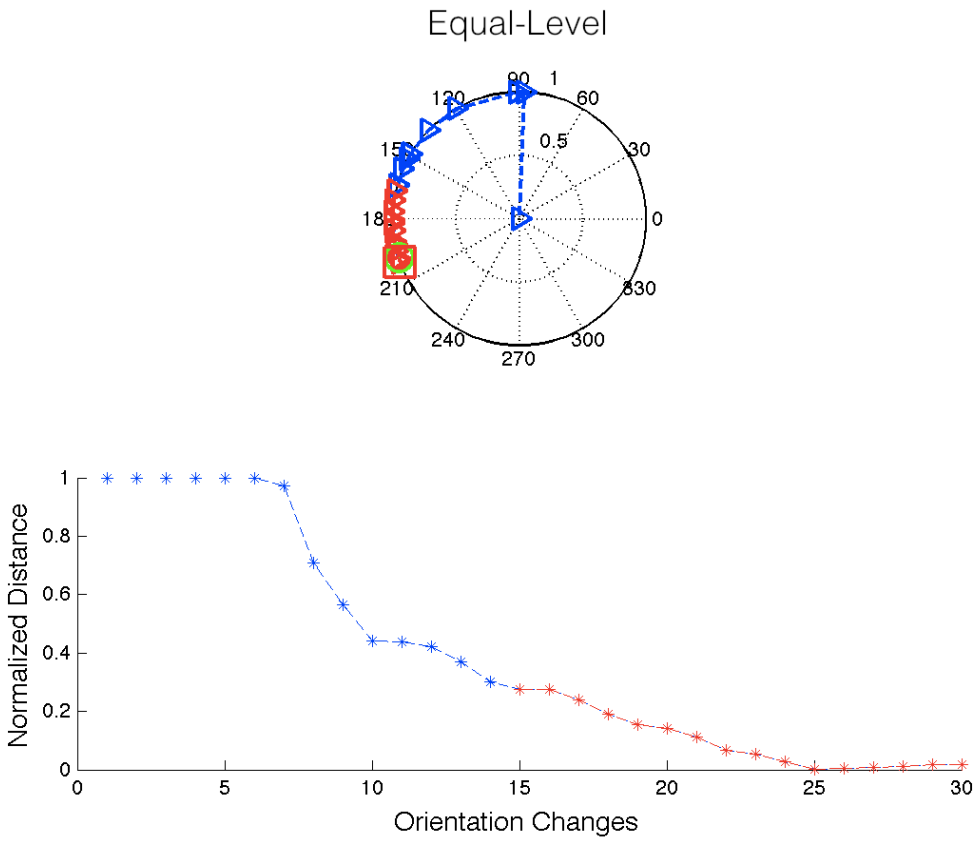


Figure 4.2: Equal-Level rotation strategy.
Top Panel: The green circle represents the actual location of the target sound. The red square represents the location of the target sound, as indicated by the listener. The dashed lines represent the order in which the listener faced each location. The first half of the rotation path is shown in blue and the last half of the rotation path is shown in red.
*Bottom Panel:*The listener's angular distance from the sound source. Blue corresponds to the first half of the rotation path and red corresponds to the second half of the rotation path. Along the abscissa is each change in rotation, and along the ordinate is the angular distance from the sound source (normalized by the shortest path length).
In the equal-level strategy, the listener rotates in the direction of the sound source and stops when the sound's intensity reaches an equal level in both ears.

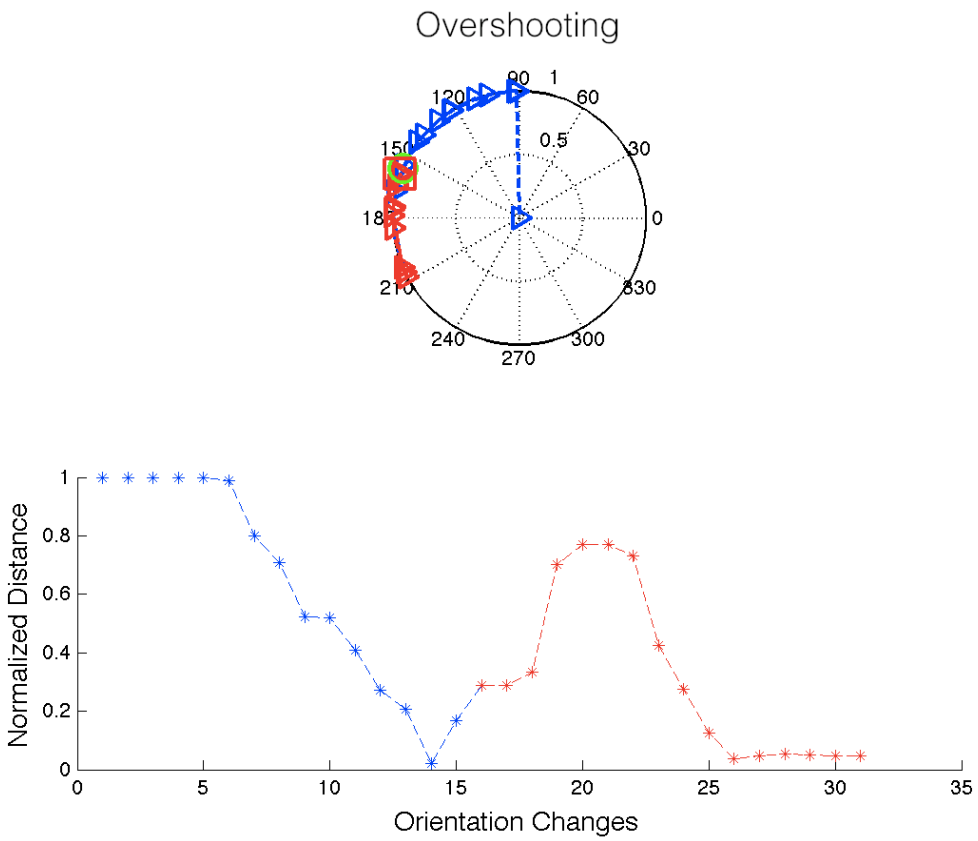


Figure 4.3: Overshooting rotation strategy.
Top Panel: The green circle represents the actual location of the target sound. The red square represents the location of the target sound, as indicated by the listener. The dashed lines represent the order in which the listener faced each location. The first half of the rotation path is shown in blue and the last half of the rotation path is shown in red.
*Bottom Panel:*The listener's angular distance from the sound source. Blue corresponds to the first half of the rotation path and red corresponds to the second half of the rotation path. Along the abscissa is each change in rotation, and along the ordinate is the angular distance from the sound source (normalized by the shortest path length).
In the overshooting strategy, the listener rotates in the direction of the sound source and lets it pass through both ears. Then the listener reverses their rotation direction until the sound's intensity reaches an equal level in both ears.

them (Figure 4.3).

In 51% of the slowest avatar condition trials, listeners were observed making sudden left/right jumps to localize the sound. In 44% of the slowest trials, listeners made front/back jumps while localizing the sound (Figure 4.4). It was also observed that in 48% of the slowest avatar trials, when listeners began searching they rotated in the direction opposite the sound source (Figure 4.5).

In the avatar condition, the fastest search trials involved achieving equal sound volume levels in both ears (using the equal-level or overshooting strategies). Changing position to search for a sound provides more effective distance cues than rotation alone. The next section investigates search strategies of listeners that can change their orientation and position during search.

4.2.2 Search by change of position

A moving listener needs to be able to mentally update the location of a stationary sound source. *Loomis et al.* (2002) coined this skill as "spatial updating". Studies have demonstrated that humans are capable of spatial updating when perceiving virtual sound sources (*Ashmead et al.* (1995); *Loomis et al.* (1998); *Loomis et al.* (1993)).

Another study (*Roginska et al.* (2010b)), was performed by our research group in which listeners were not explicitly trained to hone skills to search for sound sources. Listeners were asked to move as efficiently as possible to the location of the sounds. Their search trajectories were characterized and classified. Similar to *Roginska et al.* (2011), listener performance using the avatar and natural mediations were evaluated. In the avatar condition, the listener used a mouse and keyboard to interact with the VAE. The mouse controlled the position of the listener (on the horizontal plane), and the keyboard controlled orientation. In the natural condition, a head-tracker was used to detect the listener's position and orientation.

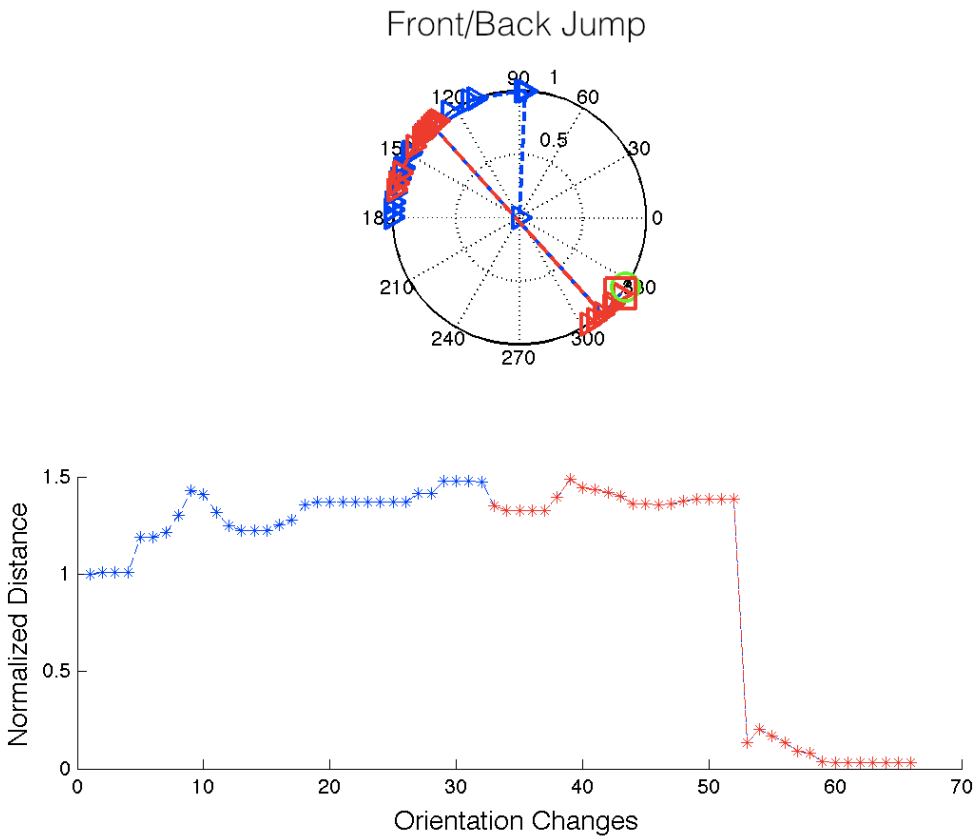


Figure 4.4: Front/Back jump.
Top Panel: The green circle represents the actual location of the target sound. The red square represents the location of the target sound, as indicated by the listener. The dashed lines represent the order in which the listener faced each location. The first half of the rotation path is shown in blue and the last half of the rotation path is shown in red.
*Bottom Panel:*The listener's angular distance from the sound source. Blue corresponds to the first half of the rotation path and red corresponds to the second half of the rotation path. Along the abscissa is each change in rotation, and along the ordinate is the angular distance from the sound source (normalized by the shortest path length).
In front/back jumping, the listener rotates to localize the sound then turns around suddenly to localize the sound source.

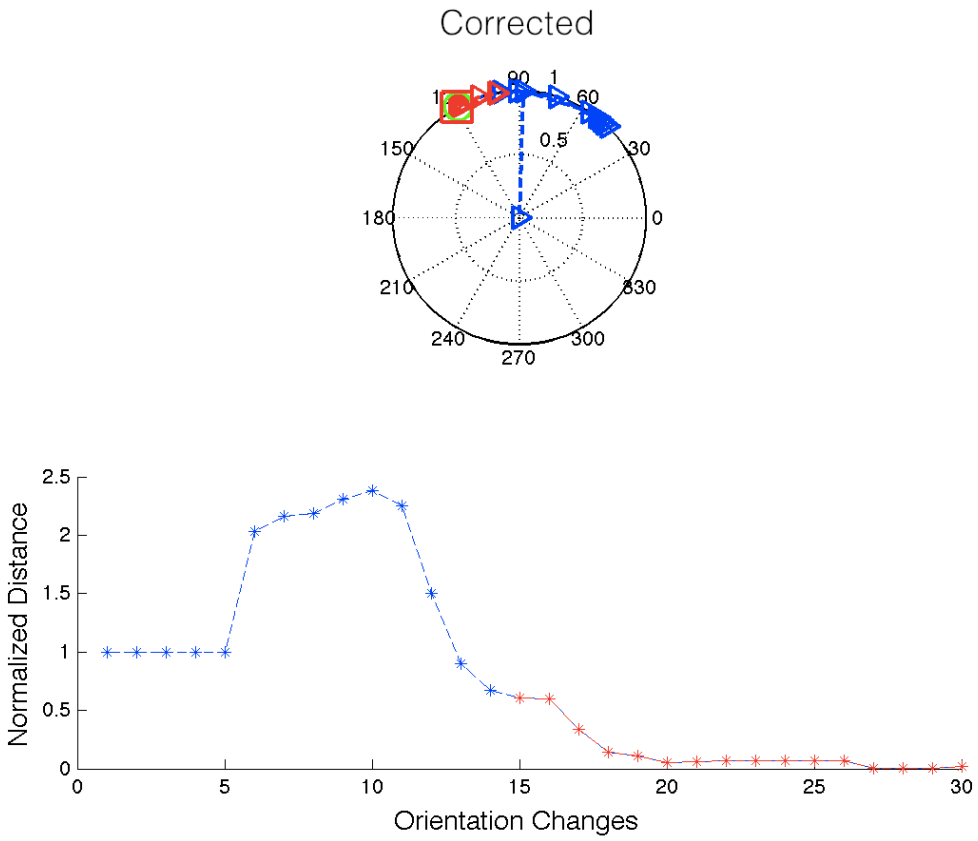


Figure 4.5: Initial Rotation Correction

Top Panel: The green circle represents the actual location of the target sound. The red square represents the location of the target sound, as indicated by the listener. The dashed lines represent the order in which the listener faced each location. The first half of the rotation path is shown in blue and the last half of the rotation path is shown in red.

*Bottom Panel:*The listener's angular distance from the sound source. Blue corresponds to the first half of the rotation path and red corresponds to the second half of the rotation path. Along the abscissa is each change in rotation, and along the ordinate is the angular distance from the sound source (normalized by the shortest path length).

In initial rotation correction, the listener begins the trial by rotating in the direction opposite the sound source. Next, they correct their behavior by reversing their rotation direction to localize the sound source.

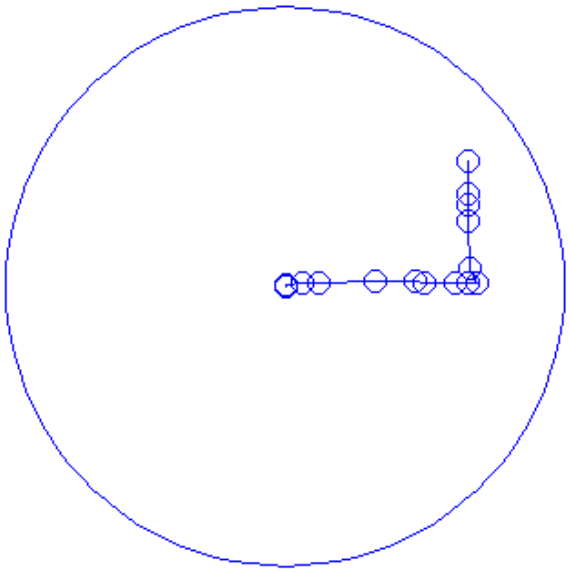


Figure 4.6: Equal-level search strategy. The listener began each trial in the center of the circle. The listener began moving laterally until achieving equal sound levels in both ears. Next the listener moved longitudinally to locate the target sound. Circles represent the search trail.

In a balanced-design experiment, 18 participants used the avatar and natural mediations to move and localize sounds in the VAE. Every trial began with a source (or sources) placed randomly along a fixed circle around the user. The listener was placed in the center of the auditory environment and instructed to find the source(s). The listener successfully localized the sound, by coming within a fixed radius of the target. First, the listener practiced finding one sound source. Then, the listener completed 20 testing trials of localizing one source in a one-source context. Finally, the listener completed 20 testing trials of finding four sources in a four-source context. After the participant finished training and testing in one modality, they repeated the procedures in the alternate modality.

Search strategies were analyzed while finding one source in the one-source and four-source contexts. *Tellevik* (1992) found that a listener's search strategy changes over time as a result of learning. Many virtual environment and spatial cognition researchers (*Buechner et al.* (2009); *Hill et al.* (1993); *Thinus-Blanc and Gaunet*

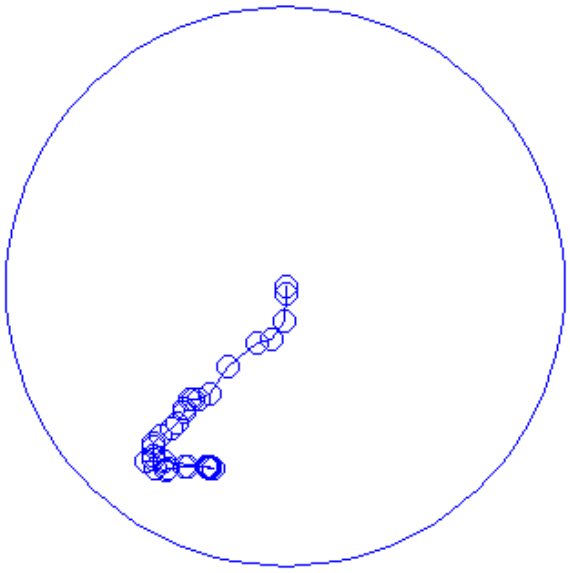


Figure 4.7: Salient search strategy. Listeners began each trial in the center of the circle. The listener began moving in the direction of the sound's most salient auditory cue (in this case, backwards). Circles represent the search trail.

(1997)) have classified spatial search patterns into *novice* and *experienced* search performance. In novice search, the listener typically traveled a longer, indirect path to locate the sound source. Experienced search performance used a shorter, more direct path to the source. These classification schemes were used to categorize each participant's search pattern.

Equal-level and salient were classified as experienced search strategies. Listeners who used the equal-level strategy (Figure 4.6) moved laterally in the direction of the sound source. Once achieving equal sound levels in both ears, the listener moved longitudinally to locate the sound. In the salient strategy (Figure 4.7), the listener primarily moved in the direction of the most salient difference between their position and the target sound's position.

Circling was classified as a novice search strategy. In the circling strategy, the listener was observed circling the perimeter of the auditory environment to find the target sound. Although listeners were successful in finding the target sound, use

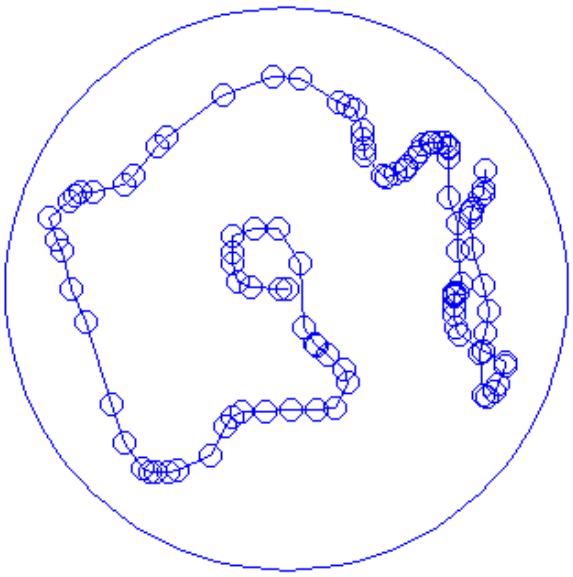


Figure 4.8: Circling search strategy. Listeners began each trial in the center of the circle. The listener circled the auditory environment to localize the target sound. Circles represent the search trail.

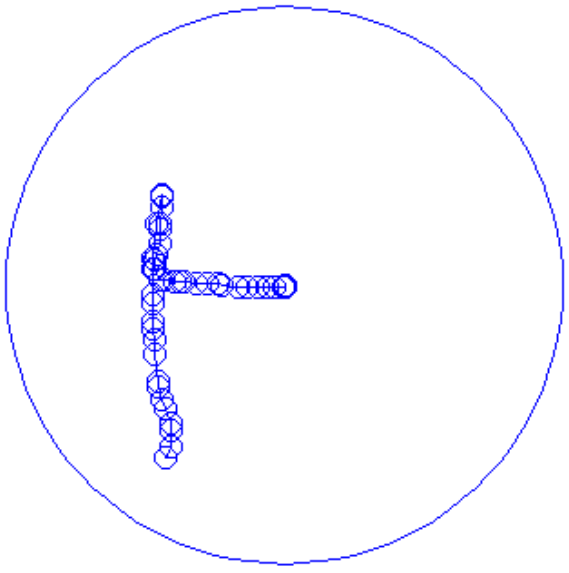


Figure 4.9: Front Back Confusion. The listener begins by moving laterally until achieving equal sound levels in both ears. Next, the listener exhibits front/back confusion by first moving forward to locate the target sound located behind them. Circles represent the search trail. Listeners began each trial in the center of the circle.

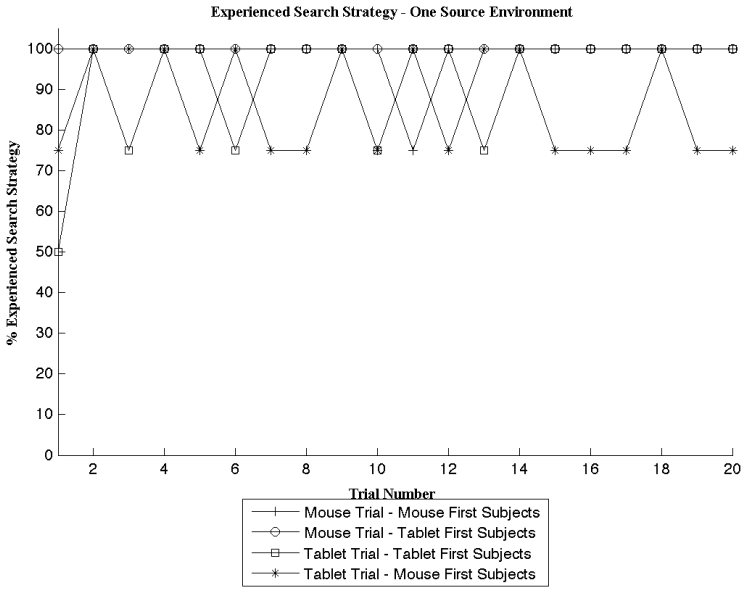


Figure 4.10: Experienced search strategy usage in the one-source VAE (reprinted with permission from *Roginska et al.* (2010b)). Along the abscissa is the trial number and along the ordinate is the percentage of trials in which an experienced search strategy was used.

of this strategy does not provide the most direct path to locate the target sound and is not indicative of learning (Figure 4.8). Listeners also experienced front/back confusion during search (Figure 4.9).

Figures 4.10 and 4.11 show the percentage of trials in which experienced strategies were used when searching for one source in the one-source and four-source contexts. In a majority of the trials, listeners used experienced search strategies.

These observations suggest that lateral movement to achieve an equal sound level in both ears and moving longitudinally to locate the sound is a very common search strategy. Equal-level strategy was observed in VAEs where the listener could only rotate from a fixed position and in contexts where they could rotate and change position. Additionally, longitudinal (front/back) movement was often used to localize sound sources.

Listener rotation usage was also examined. In the VAE, position and orientation

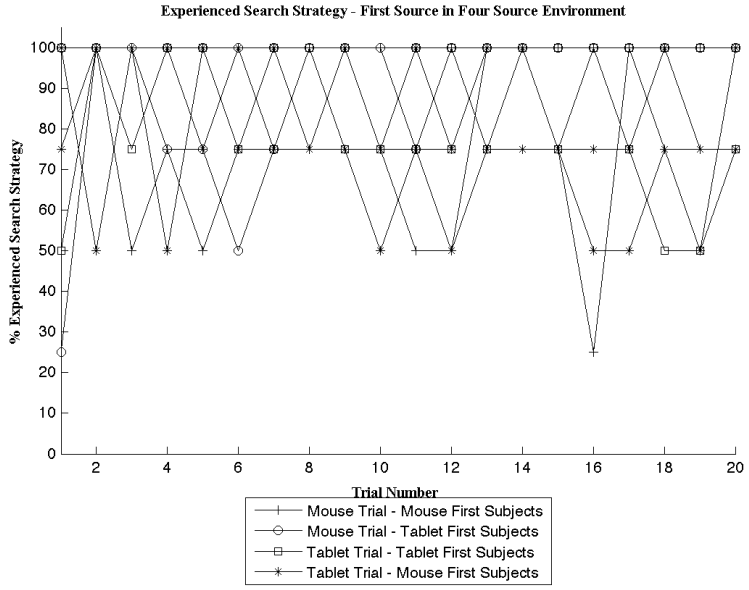


Figure 4.11: Experienced search strategy usage in the four-source VAE (reprinted with permission from *Roginska et al.* (2010b)). Along the abscissa is the trial number and along the ordinate is the percentage of trials in which an experienced search strategy was used.

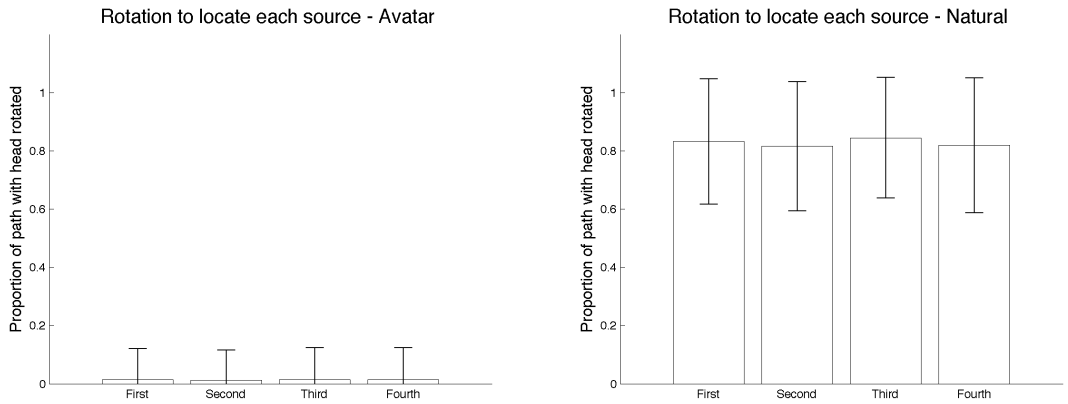


Figure 4.12: Rotation while searching for each source in the avatar mediation (left panel) and natural mediation (right panel). Along the abscissa is the localized source. Along the ordinate is the proportion of path steps in which the listener rotated their head while searching for the target source. In both panels, the error bars indicate the standard deviation.

were sampled at a rate of 10 Hz. Each sample was regarded as one step in a path to localize the intended sound source. In the avatar mediation condition, listeners moved while rotating their heads in 2% of their path steps. In the natural mediation condition, listeners moved while rotating their heads in 83% of their path steps (Figure 4.12).

The usage of rotation may have been affected by mediation constraints. The head-tracker interface used in the natural mediation supported a more natural head rotation, whereas, the avatar condition used a less natural rotation method (pressing the left and right keyboard buttons). This observation suggests that listeners may not be inclined to rotate in the avatar condition due to a more unnatural interface interaction. On the other hand, the Polhemus head-tracker system was used in a very small area to compensate for tracking range. This may have encouraged listeners to rotate more, to avoid running into walls while changing position. One might also infer that the change in level is a more effective spatial updating cue than change in orientation and this encouraged users to use positional change to find sounds, rather than orientation change.

Equal-level balancing was also observed in a comparable virtual sound search experiment. For example, in *Loomis et al.* (1990), listeners stood in the center of a gymnasium and walked to find virtual sound sources that were located at randomly determined azimuths. The stimulus was a pulse train created from a 5 Hz square-wave signal that was modified by a high pass filter, resulting in 10 pulses per second. Walking trajectories of five participants are shown in Figure 4.13. In the participants' walking trajectories, listeners often moved their heads and/or paths from left to right in a zigzag pattern, perhaps in an effort to maintain equal sound levels in both ears. This suggests that participants of *Loomis et al.* (1990) also used the equal-level strategy to localize sounds.

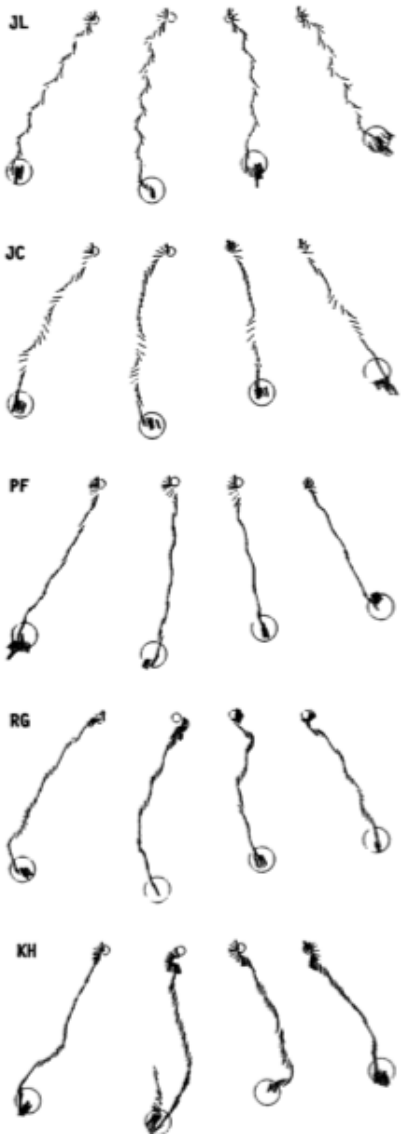


Figure 4.13: Walking trajectories of 5 participants in localizing virtual sound sources (from *Loomis et al.* (1990)). The line segments represent head orientation. Listeners began the trial at the smaller circle and ended at the center of the larger circle.

4.2.3 Training procedure

The present study uses our findings obtained from *Roginska et al.* (2010b) and *Roginska et al.* (2011) to design a sound search training procedure. We observed that listeners used the equal-level strategy and longitudinal movements to localize sound sources. Level changes seem to aid spatial updating more than angular orientation. Additionally, front/back confusion remains as a major challenge when localizing spatial sound sources. The proposed training procedure focuses on honing equal-level balance and front/back resolution skills.

The proposed training procedure consists of two consecutive stages: axial training and random-source training. In axial training, front/back resolution and equal sound level balancing skills are practiced. Listeners begin each trial in the middle of a VAE. One sound source is placed at a randomly determined range directly in front, behind, or at either side of the listener. Listeners practice equal-level balancing skills when localizing the lateral sounds. Listeners practice front/back distinction when localizing the sounds in front of or behind them.

The second stage of the training procedure is random-source training. Listeners begin each trial in the middle of a VAE. A sound source is placed at a randomly determined range and a randomly determined azimuth. Listeners use their training from the axial task to search for the randomly placed sound sources.

In both stages, feedback was given after the listener marked the position of the source. The feedback displayed the actual location of the sound, alongside the listener's marked location. Feedback is very critical in training. For example, *Zahorik et al.* (2006) found that giving feedback following sound-source localization could facilitate rapid accuracy improvements, or rapid perceptual recalibration. In some cases, the auditory recalibration can take place within minutes (*Lewald* (2002); *Recanzone* (1998)).

Additionally, while training, the listener indicated the position of the sound by

explicitly marking their location. This differed from our work in *Roginska et al.* (2010b), in which search concluded when the listener arrived within a certain radius of the sound source. Explicitly marking the sound's position allows the user to represent their positional response more precisely than trial termination when the source is discovered. Additionally, explicit marking allows the listeners the opportunity to train to distinguish subtle auditory cues, since they are able to make small movements very close to the source while attending to the spatial cue changes.

The training procedure was assessed to determine how it affected the listener's performance in a search task. Efficacy was determined by comparing pretest and posttest data. The pretest served as a baseline measurement of listener performance. In the pretest, listeners use the avatar interface to walk around and mark five randomly-distributed sound sources. Next, the listener completed the proposed training procedure. Finally, the listener completed the posttest, which followed an identical procedure as the pretest with different randomly-determined sound locations.

To assess the efficacy of the training procedure, we measured the effects of training on search accuracy and time in the posttest as compared to the pretest. Assessment included measuring the change in positioning accuracy, search time, and front/back confusion rates during training.

The present experiment also observed search behavior in the pretest and posttest. Specifically, we determined whether stimuli spectral cues affected search accuracy. Each trial was also examined to detect the presence of learning effects during the trial.

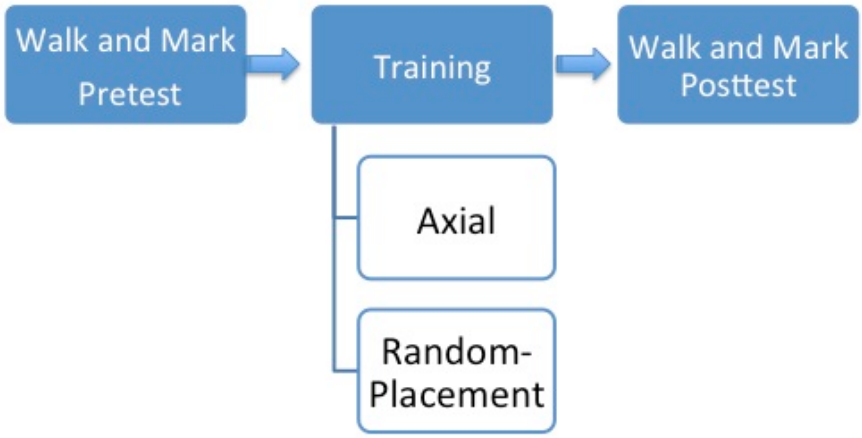


Figure 4.14: Experimental tasks used to assess training procedure. First the listener completed the *"Walk and Mark"* pretest. Next, the listener completed a two stage training procedure. After completing the axial training and random-placement training, each listener repeated the *"Walk and Mark"* posttest.

4.3 Experiment: Effects of training ("Walk and Mark")

4.3.1 Methods

The experiment consisted of a pretest, training, and posttest (Figure 4.14). In the pretest, each participant completed the *"Walk and Mark"* task. In the task, the listener navigated through the interface and marked the positions of five sound sources. The listener completed this task 20 times. After completing the pretest, the listener completed axial training. In axial training, the listener learned to localize a sound source located to their immediate left or right or behind or in front of them. Next, in random-placement training the listener learned to localize a sound source that was placed anywhere in the interface. Finally, in the posttest, each listener repeated the "Walk and Mark" experimental task.

4.3.2 Participants

Five of the fifteen listeners from the previous experiment participated in the current experiment. The experiment required about 2.2 hours of listening and was com-

pleted in one session. Participants were paid $10 per hour. Before taking part in the study, each participant gave his or her consent by reading and signing a consent form (Appendix A).

4.3.3 Stimuli

The experimental stimuli consisted of five distinct environmental sounds (see Table 4.1). The sounds were chosen from *BBC* (1991) based upon their ability to be perceptually segregated as five individual sounds. These specific sounds were chosen because of their distinct spectro-temporal patterns. Each signal also contained broadband energy distribution and transients, which aid sound search (*Blauert* (1983); *Begault* (2000)). Furthermore, these signals are mutually inconsistent (unlikely to occur together in a natural acoustic environment). Each sound was between 23 and 80 seconds long, and repeated continuously. Stimuli were presented at an audible level, as adjusted by the participant.

Figures 4.15 through 4.19 present the time-frequency analysis of the five stimuli. The distribution of energy is described in time, frequency, and intensity. The red portions indicate more intensity and the blue portions indicate less intensity. The time-frequency analysis was performed using the `specgram()` function of the MATLAB Signal Processing Toolbox. `Specgram()` splits the signal into overlapping sections and applies the specified window to each section. Then, it computes the discrete-time Fourier transform of each section to produce an estimate of the short-term frequency content of the signal.

4.3.4 Apparatus

Experiments were performed using the same system that was used in experiments of the previous chapter. As in *Roginska et al.* (2010b), a real-time spatial auditory system programmed in MATLAB was used (Figure 4.20). The system spatialized

Table 4.1: Environmental sounds and their labels

Sound	Labels
Drumsticks striking a drum at regular intervals	Drums
Computer generated electronic noises	Electronic
A river flowing rapidly	River
Crickets chirping	Crickets
Typewriter keys being pressed	Typewriter

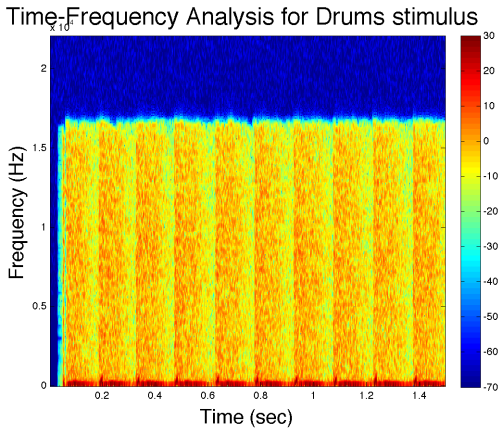


Figure 4.15: Time-Frequency analysis of drumsticks striking a drum at regular intervals (Drums)

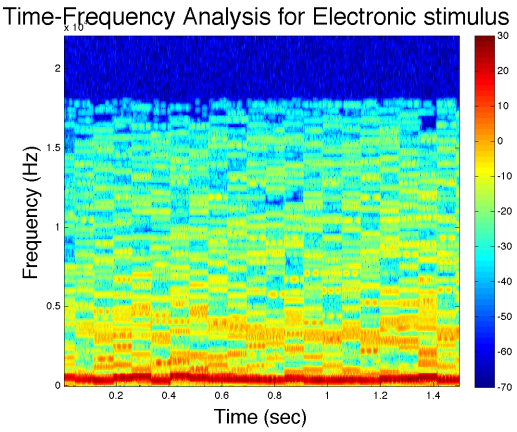


Figure 4.16: Time-Frequency analysis of computer generated electronic noises (Electronic)

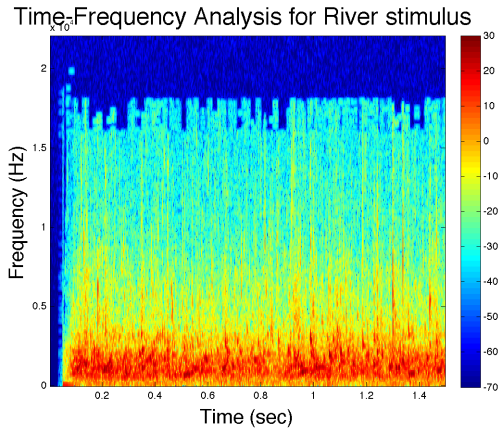


Figure 4.17: Time-Frequency analysis of a river flowing rapidly (River)

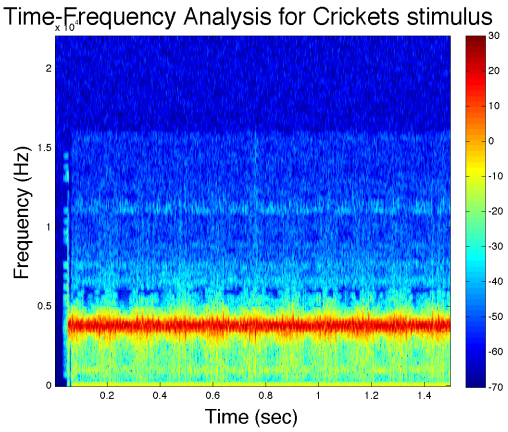


Figure 4.18: Time-Frequency analysis of crickets chirping (Crickets)

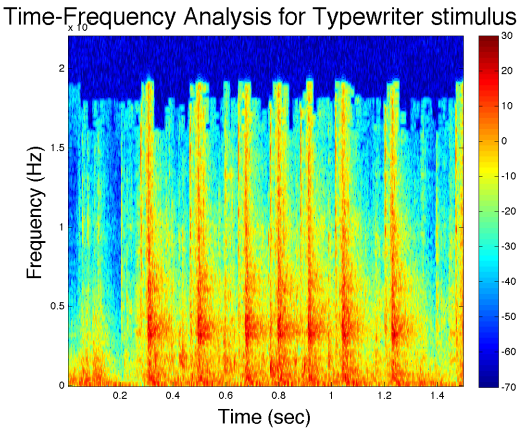


Figure 4.19: Time-Frequency analysis of typewriter keys being pressed (Typewriter)

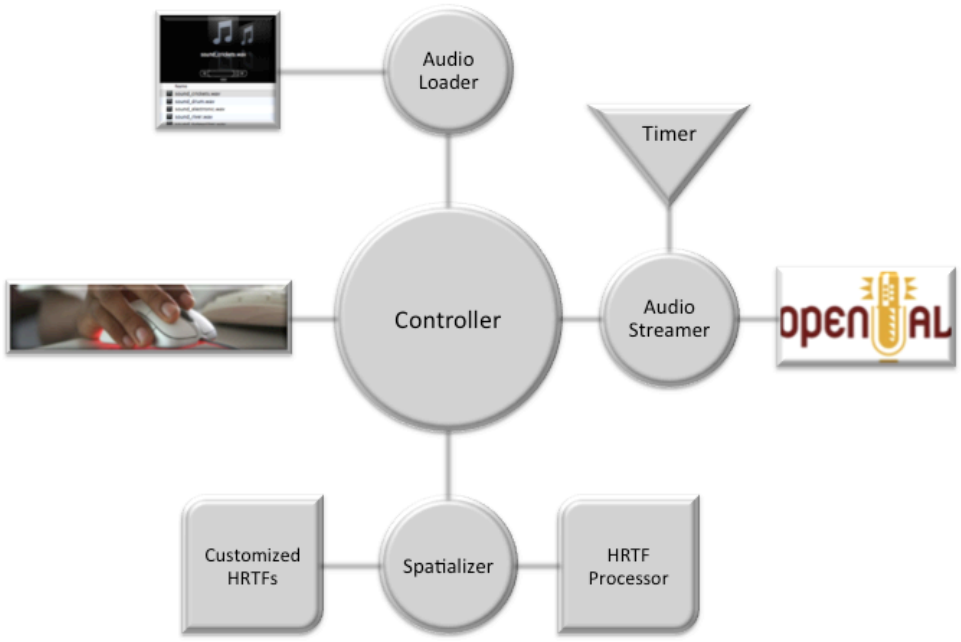


Figure 4.20: Structure of spatial audio system implemented in MATLAB.

the auditory stimuli using each participant's customized HRTF (created from the experiment in Chapter III). The controller sampled the participant's position and orientation at a rate of 10 Hz. To facilitate a real-time double buffering audio scheme, a timer was used to generate a new frame of audio corresponding to the participant's position within the VAE. The timer called routines to read the sound file from disk and convolve the audio input using the orientation-adjusted interpolated HRTFs. Audio was controlled using the PsychToolbox extension of OpenAL by double buffering. The timer also queried OpenAL sound source to determine if one of the buffers had finished playing, so that the next frame of audio could be loaded into the buffer, to be played after the current buffer. Interpolation of the HRTFs was implemented by constructing the minimum phase impulse response of a system whose magnitude spectrum is determined from a log mixture of the adjacent measured HRTFs (sampled every 10 degrees) and convolving the result with an all-phase system using a fractional-

delay method. An inverse-square law was used to attenuate the sounds as the listener moved through the VAE. The VAE was scaled so that there was a 13 dB drop in gain for a sound located at the farthest possible distance from the listener. Sounds were attenuated such that any sound could be detected from any point within the interface.

Trials were conducted in the same location as the main experiment of Chapter III. Interface mediation was controlled by a mouse and keyboard interface. Standard mouse to screen cursor mapping was used. Clicking or dragging the mouse in a new location moved the position of the onscreen avatar. The keyboard's right and left arrows were used to control the yaw of the avatar's head by rotating their orientation in steps of two degrees.

4.3.5 Procedure

The pretest and posttest were identical tasks. In each, the listener was instructed to use the mouse and keyboard to explore the environment and discover the locations of the five sound sources. The listener was told to mark the positions at which they would be standing if he/she were in the same location as the sound source.

The users began the pretest and posttest facing forward in the middle of the silent VAE. After pressing a button, the five stimuli played simultaneously at randomly chosen locations within a unit circle around the listener (Figure 4.21). Each sound was about 10 seconds long with repeated components, played in a loop throughout the trial. The locations of each sound were chosen pseudo-randomly on each trial so that each sound was spaced at least 30° apart and separated by a distance that equaled one-third of the diameter of the auditory space.

The listener's position and orientation were recorded as they moved around the environment. When the participant was standing in the same location as the sound source, they pressed the space bar, which placed a blue marker on the screen at the

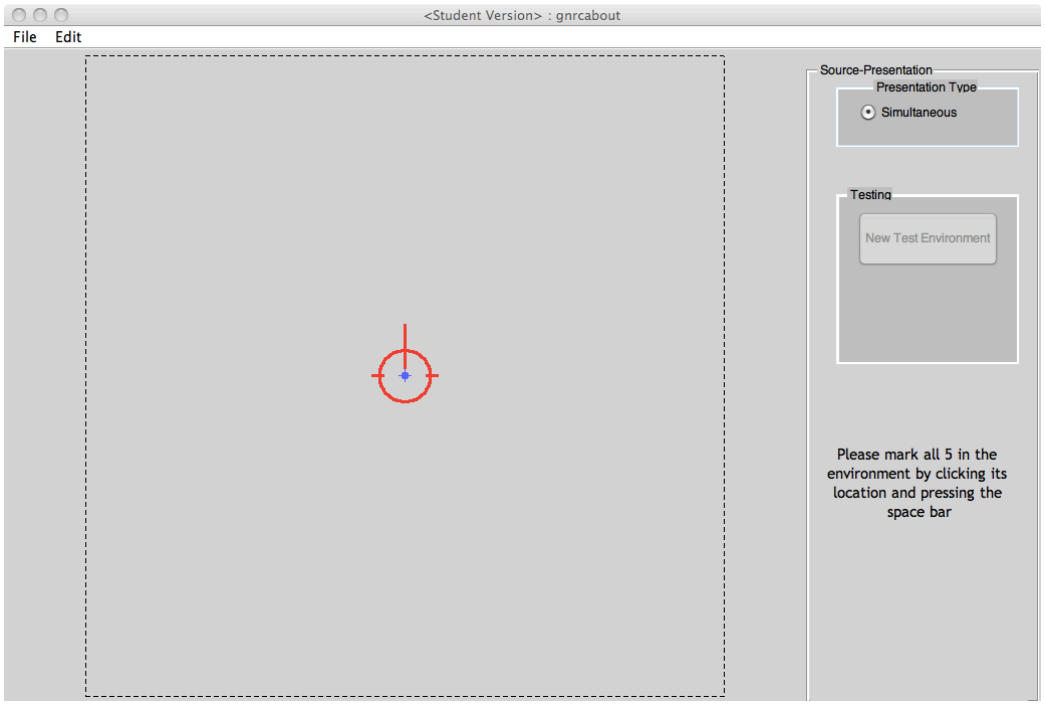


Figure 4.21: Auditory interface shown at the beginning of each trial. The red circle represents the listener's position and the long red line intersecting the circle represents the heading. The listener pressed "New Test Environment" to begin each trial.

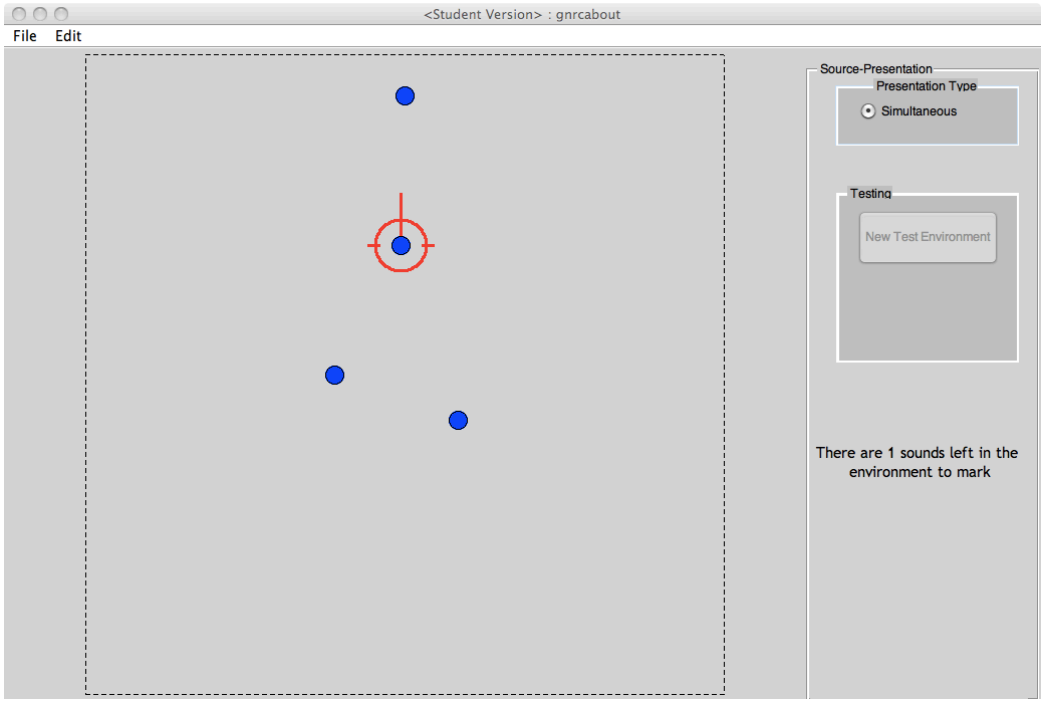


Figure 4.22: The participant's position is represented by the large red circle. In this example, the listener has already marked the locations of four sound sources, represented in blue.

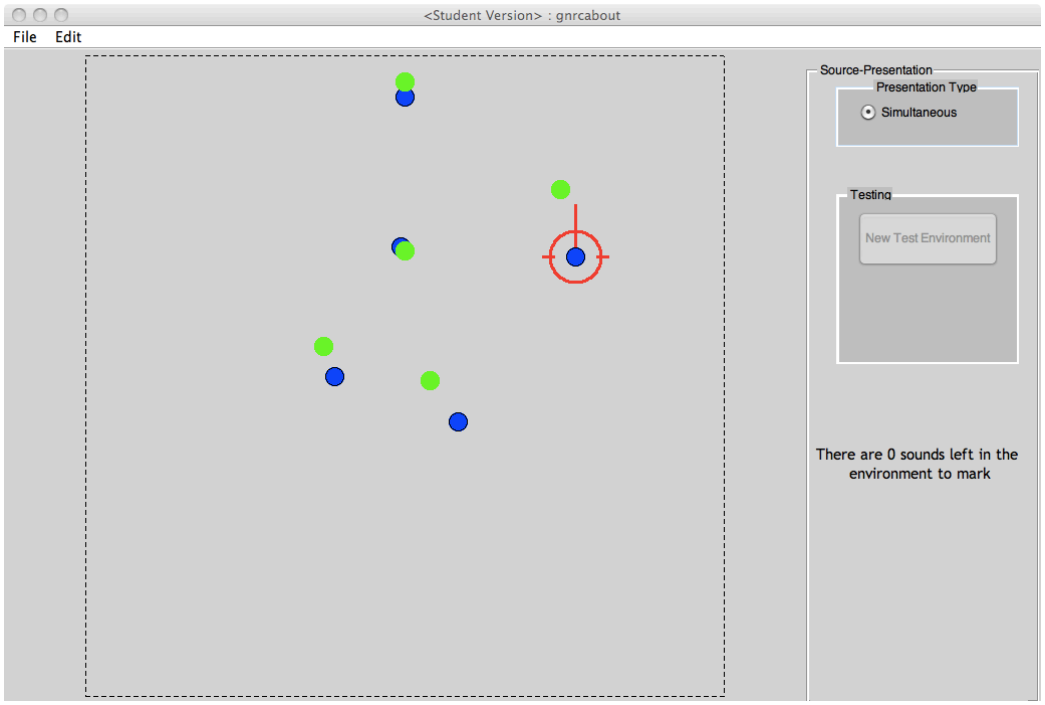


Figure 4.23: The participant has marked the perceived locations of the five sound sources (in blue). The true locations of the sound sources were shown in green for five seconds before the user begins the next test environment.

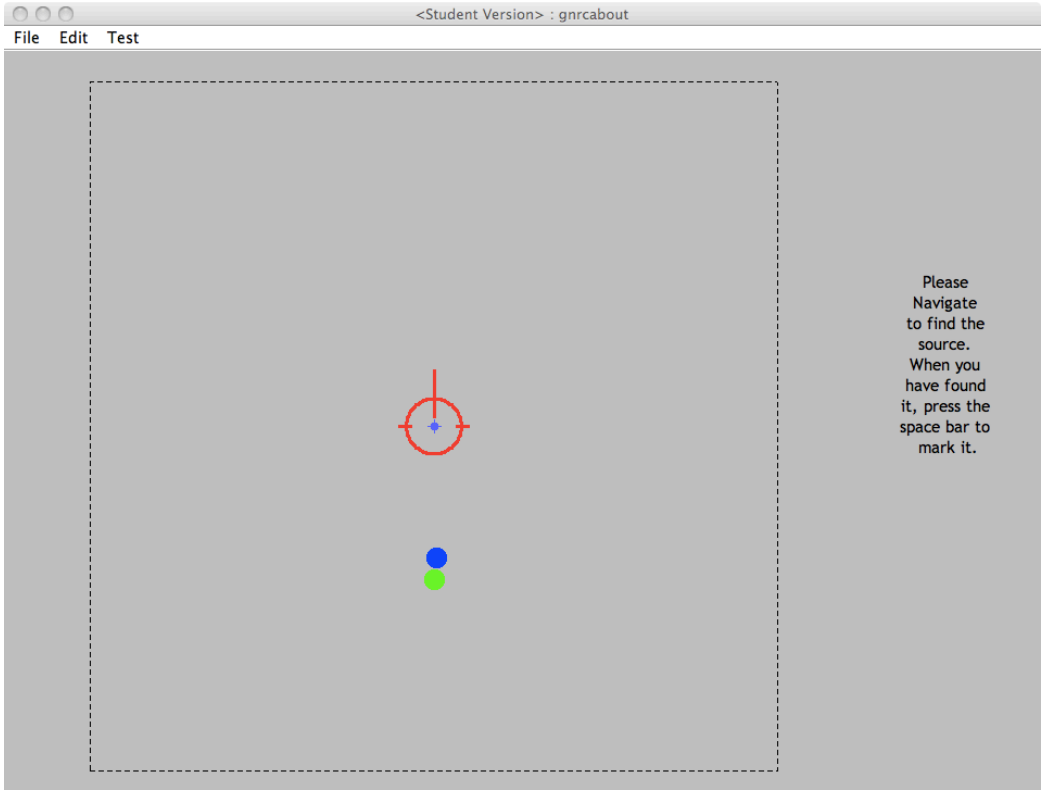


Figure 4.24: The participant's position is represented by the large red circle. In this example, the user has already marked the location of the sound source directly behind them. Feedback was given in green before the next trial began.

avatar's current position. Each participant repeated this procedure to mark the locations of the remaining sound sources. An example of this procedure can be seen in Figure 4.22. Once the participant marked all of the sources, the sounds in the environment stopped playing. The true locations of the sound sources and the participant's marks were shown simultaneously for five seconds (Figure 4.23). This procedure constituted one experimental trial. Each listener completed 20 experimental trials. Participants began each trial in the center of a new configuration of the five sound sources.

Following the pretest, each listener completed a two-phase training procedure. During training, the participant learned to localize one spatialized sound source. In axial training, the participant began facing forward in the middle of a silent auditory

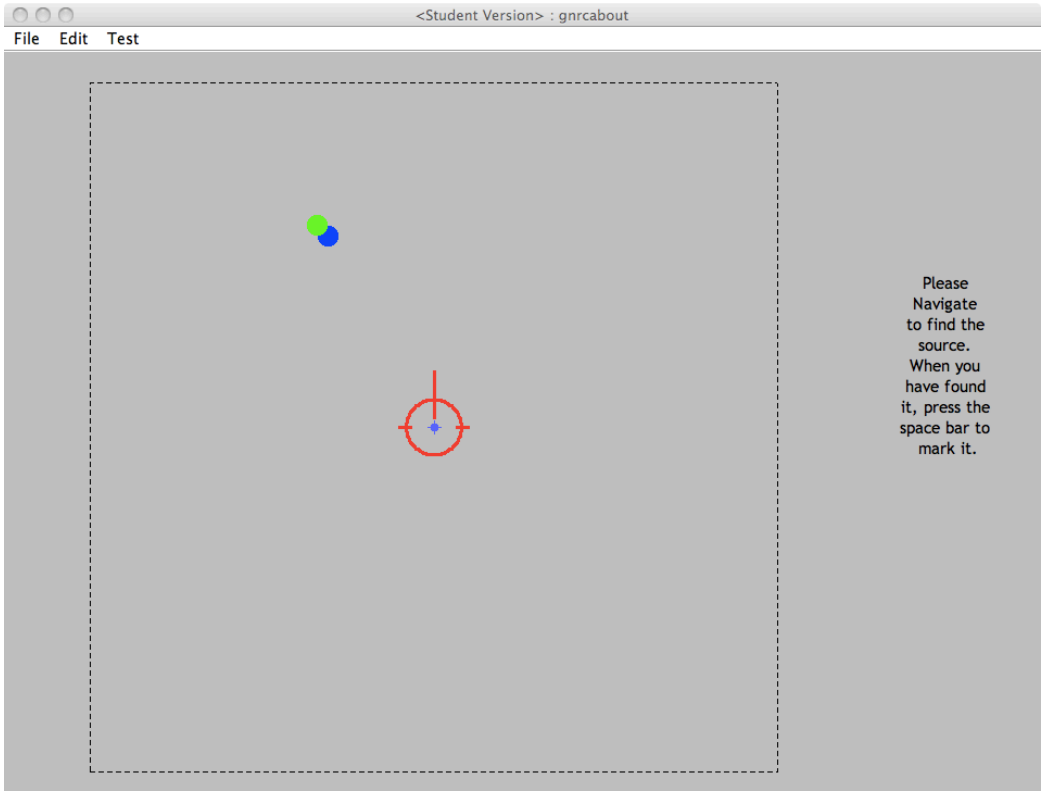


Figure 4.25: The participant's position is represented by the large red circle. In this example, the user has already marked the location of the sound source as ahead and to the left. Feedback was given in green before the next trial began.

environment. The listener localized one of five randomly chosen stimuli in each trial. Participants were told that the target stimulus would be located directly in front, behind, to the left, or to the right of them. After pressing a button, the source was placed in the environment. Participants used the mouse and keyboard to explore the environment. When the listener was standing in the same location as the sound source, they indicated its location by pressing the spacebar, which placed a blue marker on the screen. An example of this can be seen in Figure 4.24. Axial training continued until the participant reached a predetermined accuracy criteria for five consecutive trials. After successful completion of axial training, the participant moved on to random-placement training.

As in axial training, during random-placement training, the listener began each trial in the center of a silent auditory environment. In each trial, one of the five stimuli was randomly chosen to be localized. Participants were told that the target stimulus could be located anywhere around them. After pressing a button, the source was placed in the environment. The participant used the mouse and keyboard to explore and change their position in the environment. When each listener was standing in the same location as the sound source, they marked its location using the spacebar. An example of this can be seen in Figure 4.25. Training continued until the participant reached a predetermined accuracy criteria for five consecutive trials. After the listener completed the training procedure, they completed the posttest.

4.3.6 Statistical Treatment

During the training phases, pretest and posttest, the MATLAB based GUI recorded the listener's responses, exploration times, and paths. Both the pretest and posttest contained 20 trials in which 5 data points were taken, resulting in 100 data points per subject, per test. The results of the next section were assessed with an ANOVA. An a priori α level of 0.05 was applied in all inferential statistical analyses. Error bars in

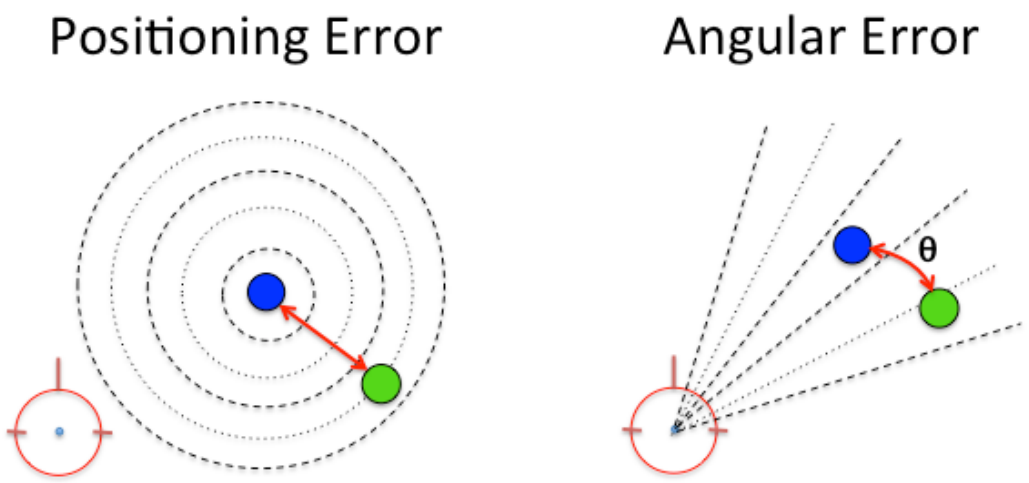


Figure 4.26: Positioning error (left) and angular error (right). In both figures, the blue circle represents the true location of the sound source and the green circle represents the location of the source as marked by the user.

all figures show 95 % confidence intervals.

4.4 Results

The difference between the actual configuration and the user's marked configuration is taken as the performance measure. Two evaluation metrics were used to assess this metric: positioning error and angular error (Figure 4.26). Positioning error was defined as the straight line distance between the true and marked sound location. Position error was calculated in term of R_{Max}, which represents the maximum distance from which a sound source could be located with respect to the listener's position in the center of the interface. Angular error was defined as the unsigned angular difference between the true and perceived sound. The sound's angle was defined as the angle between the interaural axis and a line originating at the center of the interface that intersects the sound's location.

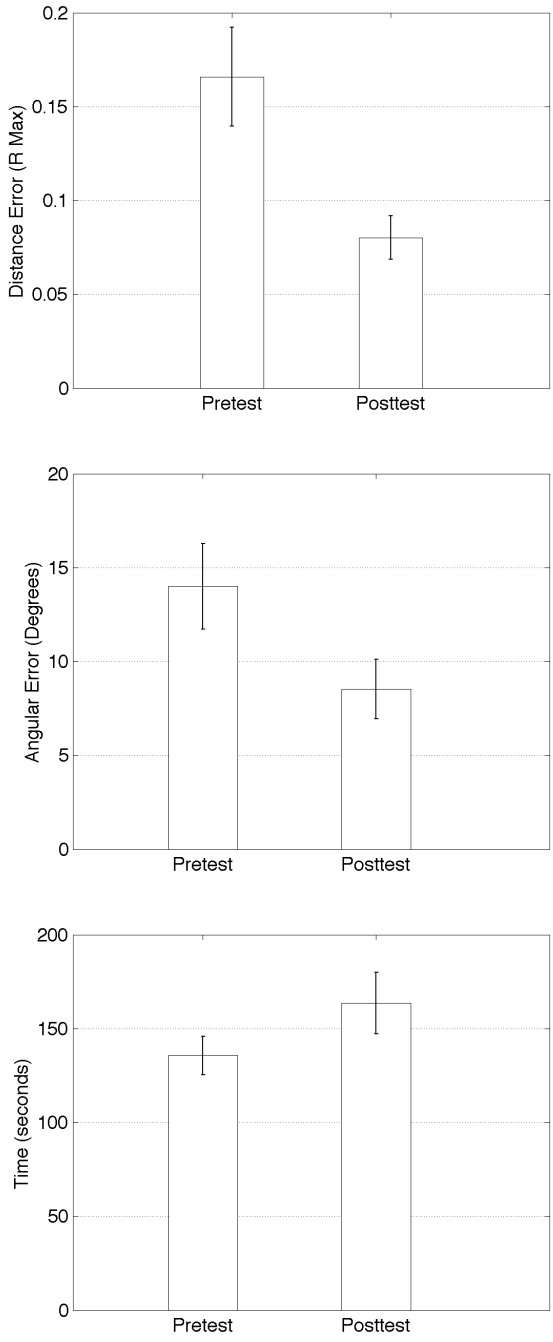


Figure 4.27: Effects of training on search accuracy and exploration time. The top panel compares positioning accuracy, the middle panel compares angular accuracy, and the bottom panel compares search times. Along the abscissa are the two factors: pretest and posttest and along the ordinate is the respective measure.

4.4.1 Effects of Training

A one-way ANOVA was used to test for performance differences as an effect of training across subjects. The average results, across subjects can be seen in Figure 4.27. In the top panel, we see that positioning error differed significantly between the pretest and posttest [$F_{1,998}$=34.09, $p<0.05$]. Similarly, in the middle panel, angular error differed significantly between the pretest and posttest [$F_{1,998}$=15.03, $p<0.05$]. In the bottom panel, time also significantly differed between the pretest and posttest [$F_{1,998}$=8.27, $p<0.05$]. In summary, training decreased search error and increased the exploration time needed.

4.4.2 Behavior during training

4.4.2.1 Positioning accuracy

We also investigated how search accuracy changed during training. The positioning error and search time were assessed at three portions of the training procedure: the first 5, middle 5 and last 5 training trials.

The top panel of Figure 4.28 compares listeners' positioning error at the beginning, middle, and end of the axial-training. It should be noted that the number of training trials differed across subjects. Each of the means is collapsed across the 5 subjects. A one-way ANOVA revealed a significant difference in performance over trials [$F_{2,72}$=6.8, $p<0.05$]. A Tukey LSD multiple comparison test shows that positioning error was significantly higher during the first 5 training trials than in the middle 5 and last 5. Similarly, the bottom panel of Figure 4.28 compares the mean positioning error during random-placement training. No difference was observed between the three sets of trials [$F_{2,72}$=0.34, $p=0.72$].

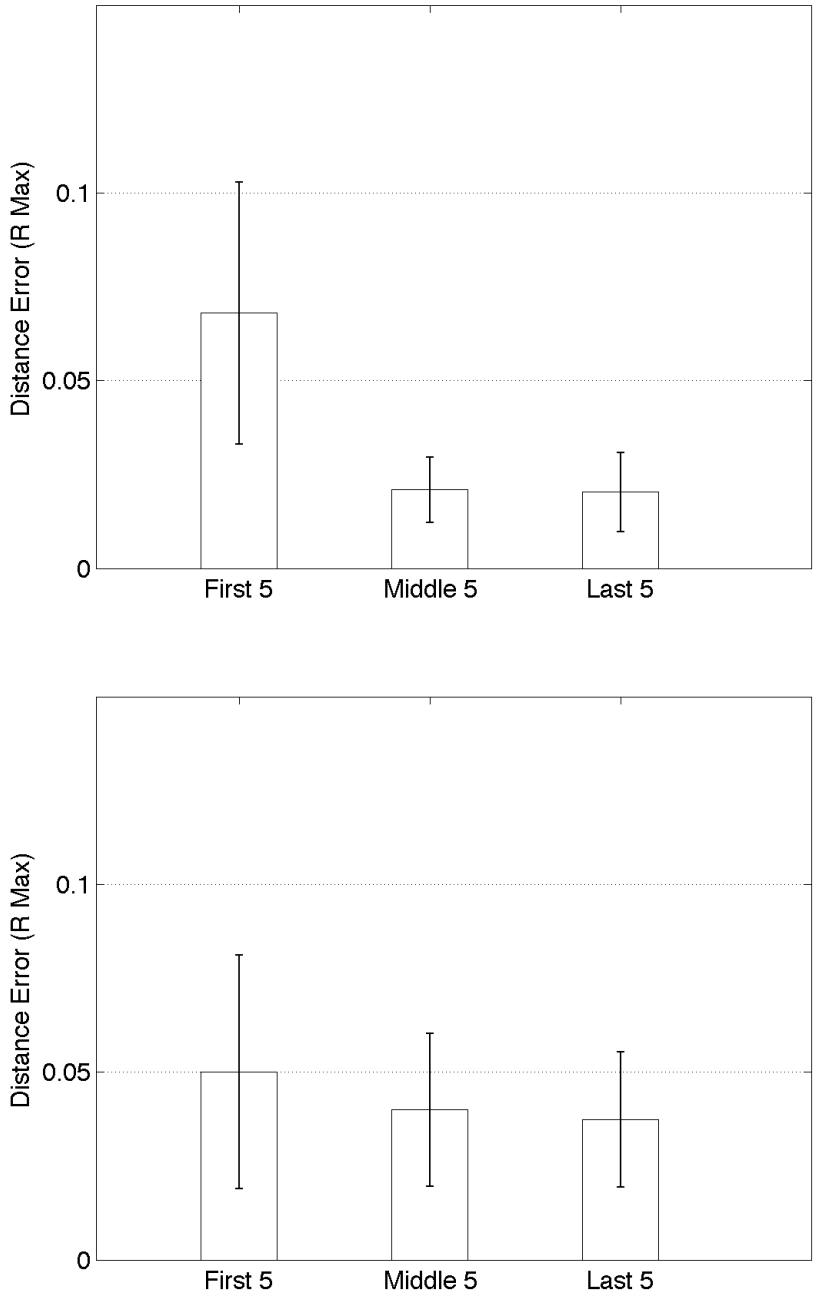


Figure 4.28: Positioning accuracy during axial training (top) and random-placement training (bottom) for selected trials. In both panels, along the abscissa are the selected training observations and along the ordinate is the positioning error.

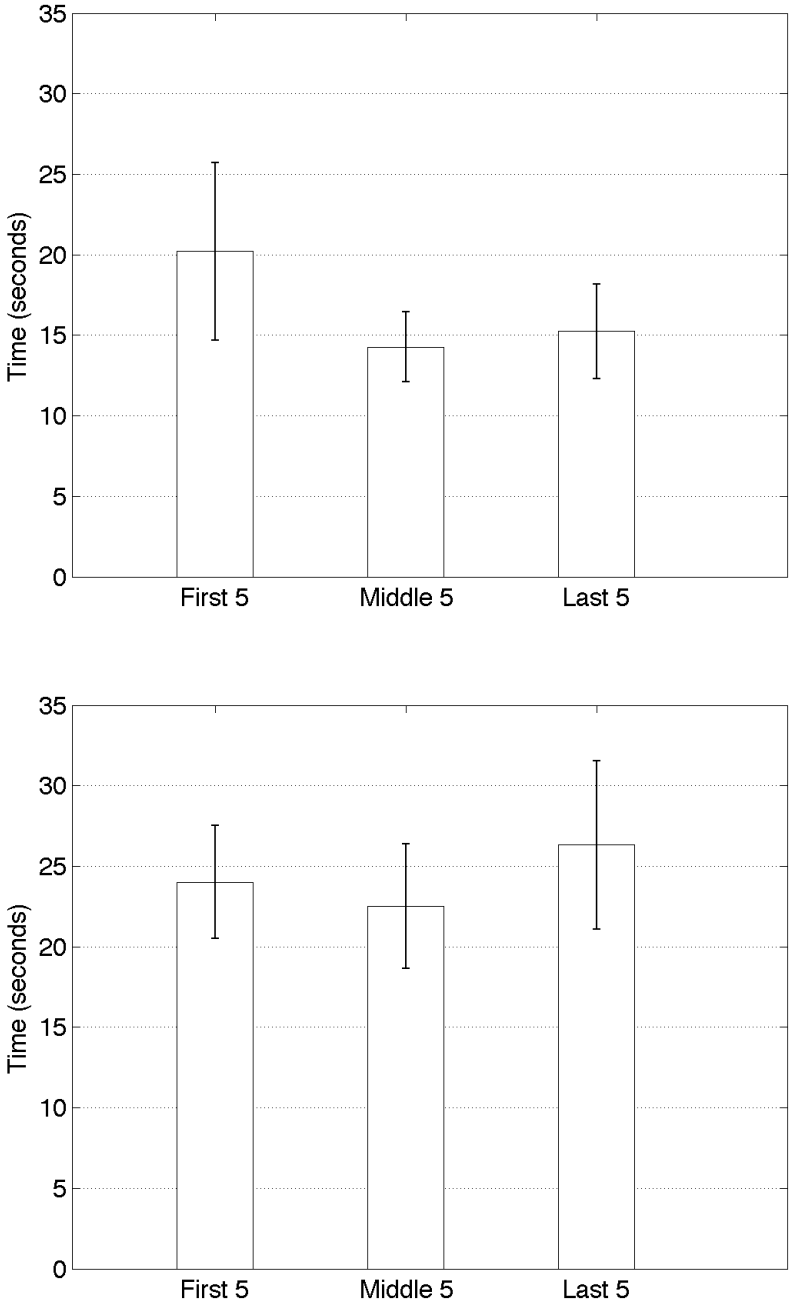


Figure 4.29: Search time during axial training(top) and random-placement training(bottom). In both figures, along the abscissa are the selected training observations and along the ordinate is the positioning error.

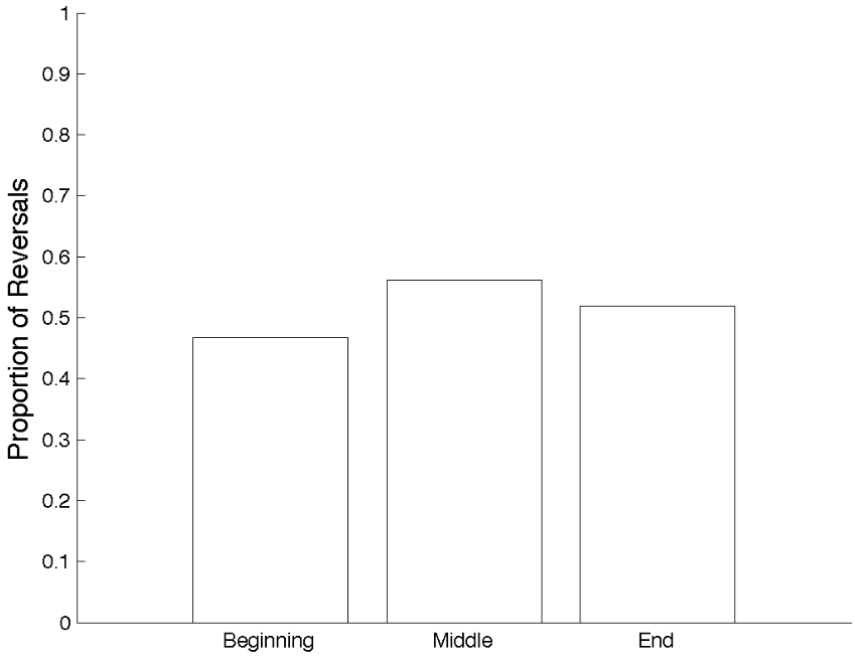


Figure 4.30: Frequency of front/back confusion during axial training. Along the abscissa is the portion of trials evaluated (one-third of overall total) and along the ordinate is the percentage of trials in which reversals were observed.

4.4.2.2 Search time

The amount of exploration time needed to locate each sound source during training was examined. Similar to the previous analysis, each participant's first, middle and last five training trials were analyzed. The results also followed a similar trend as the positioning error.

The top panel of Figure 4.29 compares search time during axial-training. Search time did not differ significantly in the three sets of trials [$F_{2,72}$=2.96, p=0.06]. Similarly, the bottom panel compares the search time during random-placement training. There difference between the sets of trials was insignificant [$F_{2,72}$=0.86, p=0.43].

4.4.2.3 Front/back confusion

To determine if training reduced front/back confusion, the training paths were analyzed. The analysis only included the axial training paths in which front/back

confusion was specifically trained. The path taken by each listener to localize the target sound source was analyzed to determine the frequency of front/back reversals. Each listener's search path was composed of a sequence of steps. Each step was examined to determine the direction of the listener's initial movement. If a listener initially moved away from the sound source and then corrected their path to move in the correct direction, this was regarded as a front/back reversal. Figure 4.30 shows the frequency of front/back confusion while navigating to a source that was in front of or behind them. Each participant's front/back training trials were divided into thirds. Averaged across the participants, the probability of reversal was 46.67 % in the first three, 56.14 % in the middle three and 51.92 % in the last three front/back trials.

4.4.3 Behavior during search

4.4.3.1 Accuracy by stimulus

We examined the positioning error, broken out by subject and stimulus type to determine if the type of stimulus used was affected by search accuracy. Although all of the stimuli contain transients, some sources (such as the Drum) have wide-band transients, which provide cues at high frequencies (dominated by IID) and low frequencies (dominated by ITD). On the other hand, the Crickets stimulus has transients mostly in the lower frequency range (less than 1 kHz). One would expect to observe higher search accuracy for rich stimuli with wide-band transients.

Positioning and angular error were analyzed for each stimulus in the pretest and posttest. The top panel of Figure 4.31 compares the mean positioning error in the pretest. There was no difference in positioning error between stimuli [$F_{4,495}$=0.28, p=0.89]. Similarly, the bottom panel of Figure 4.31 compares the mean angular error of each stimulus in the pretest. There was no difference in angular error between stimuli [$F_{4,495}$=1.36, p=0.25].

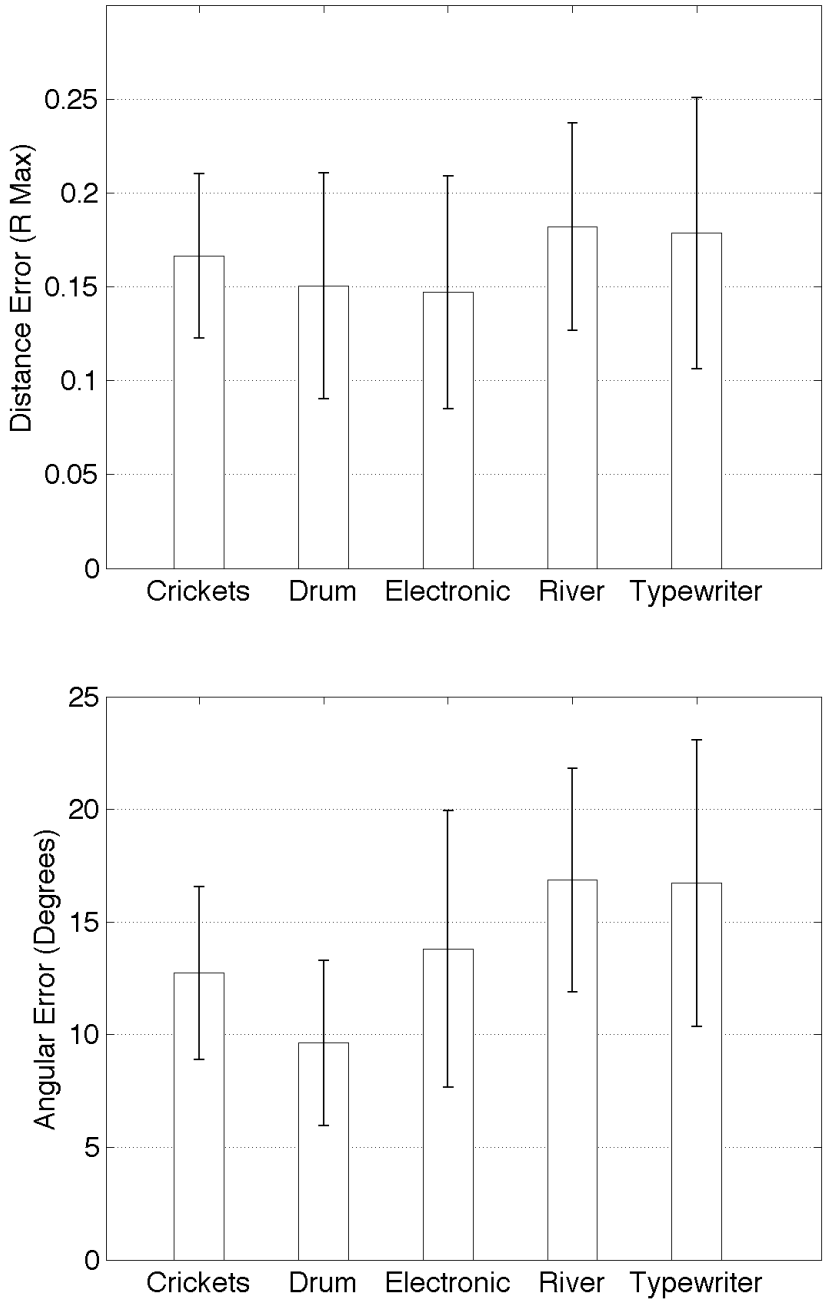


Figure 4.31: Effects of stimulus on search accuracy in the pretest. Positioning error is in the top panel. Along the abscissa is the stimulus and along the ordinate is the positioning error. Angular error is shown in the bottom panel. Along the abscissa is the stimulus and along the ordinate is the angular error.

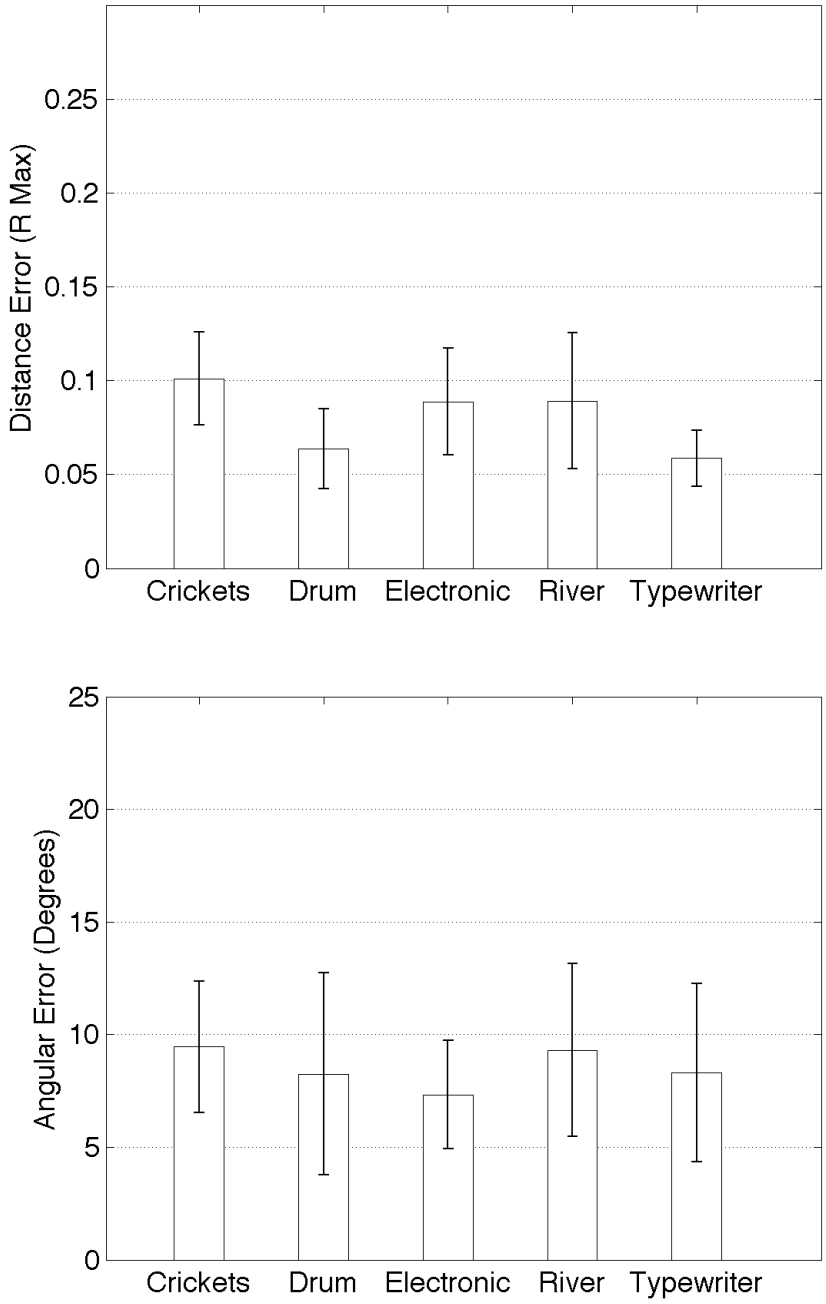


Figure 4.32: Effects of stimulus on search accuracy in the posttest. Positioning error is in the top panel. Along the abscissa is the stimulus and along the ordinate is the positioning error. Angular error is shown in the bottom panel. Along the abscissa is the stimulus and along the ordinate is the angular error.

The top panel of Figure 4.32 shows the mean positioning error in the posttest. Similar to that observed in the pretest, there was no difference in positioning error between stimuli [$F_{4,495}$=1.91, p=0.11] and as shown in the bottom panel, there was no difference in angular error between stimuli [$F_{4,495}$=0.23, p=0.92].

4.4.3.2 Sequence effects

Our study also examined if listeners learned the positions of other sounds in the environment, while localizing target sounds. The time spent localizing each sound was compared. If listeners learned the locations of other sounds during search for a different target sound, there should be significantly less time spent locating latter sounds.

The top panel of Figure 4.33 depicts the amount of time listeners spent finding each source (in its sequential discovery order) in the pretest. Analysis revealed that sound source discovery sequence did not effect search time [$F_{4,495}$=2.23, p<0.06].

Posttest results are shown in the bottom panel of Figure 4.33. There was a significant difference in source discovery order [$F_{4,495}$=4.2, p<0.05]. A Tukey LSD multiple comparison test showed that listeners spent significantly more time searching for the fifth sound than the first, second, and third. Listeners also spent more time searching for the fourth sound than the second. All other comparisons were insignificant.

4.5 Discussion

The present experiment assessed a training procedure created from search strategy observations. Adequate training is necessary because searching for sound sources in a VAE is not an inherent skill. Training is required to help listeners adapt and develop these virtual sound search skills.

The training procedure improved listeners' positioning and angular accuracy (Figure 4.27). Results suggest that the increase in search accuracy following the training

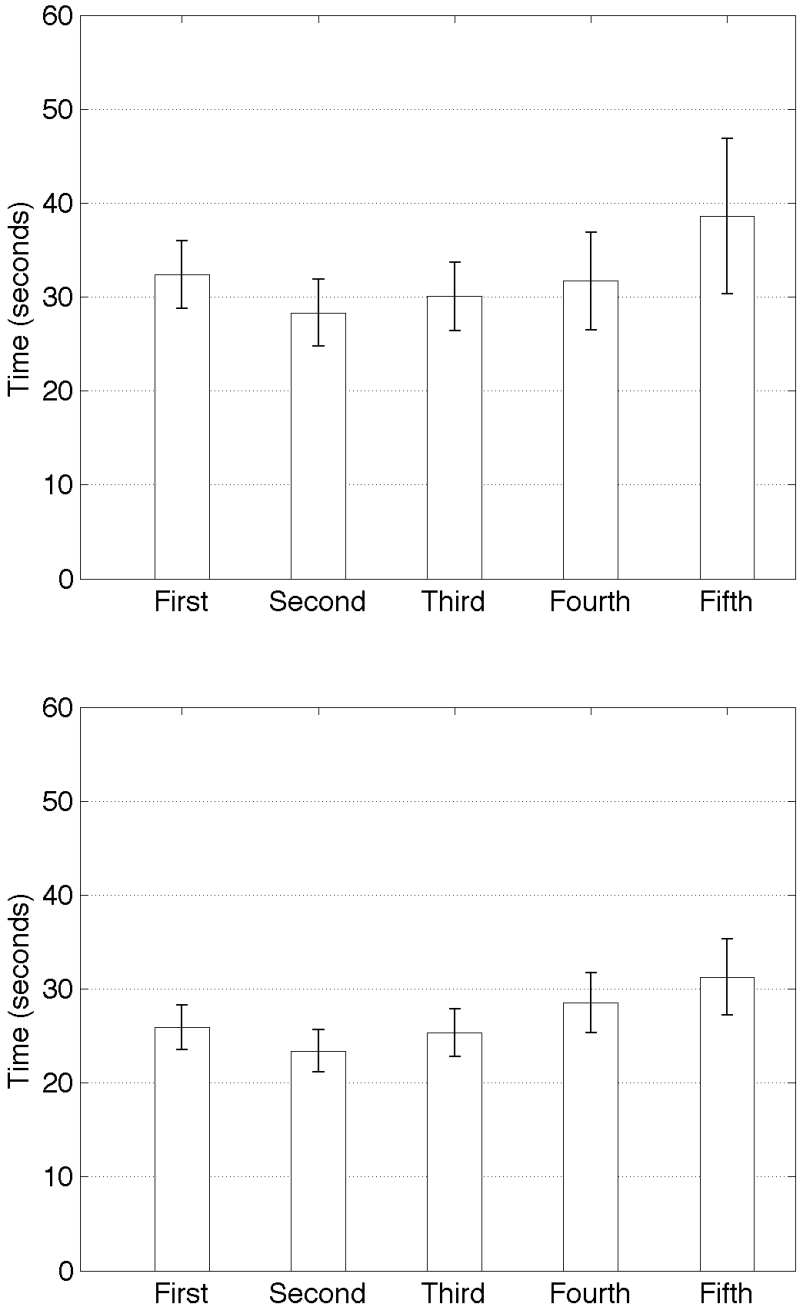


Figure 4.33: Effects of sequence on search time in the pretest (top) and posttest (bottom). In both panels, along the abscissa is the sequence order and along the ordinate is the number of seconds used to find each source.

procedure is a result of improved processing of spectral cues. On the other hand, one might argue that the listeners' search performance improved because of practice effects. Because this is a perceptual task, listeners are learning to distinguish auditory cues. Mere exposure to the environment would not have facilitated such performance improvements. However, it would be useful to compare the present results with an untrained control group to determine the magnitude of any practice effect.

Observations indicate that listeners spent more time localizing sound sources during the posttest. A possible explanation is that during training, listeners learned new skills to localize sound sources. s a result, listeners needed more time to incorporate the additional skills into their search procedure during the posttest.

Search accuracy improved significantly after the first half of axial training (Figure 4.28). Each listener's positioning error and search time significantly decreased between the first 5 and the middle 5 trials of the axial training stage. The time needed to localize each source was also significantly higher in the first five trials of the axial training. This observation is similar to the rapid perceptual recalibration observed in *Zahorik et al.* (2006) in which listeners (when given feedback) were able to learn the spatial auditory cues needed for localization. After completing half of the training, the listeners' responses were significantly more accurate and the listeners' required significantly less time to localize the sound sources. In the present study, recalibration was achieved after about one hour. *Hofman et al.* (1999) found that listeners can recalibrate to spectral cues over a period of weeks. Participants of *Zahorik et al.* (2006) showed performance improvements in two 30 minute sessions spread 4 days apart. The present training procedure may be more desirable in time-limited situations, since significant improvements were observed in a one-hour training session.

Despite axial training, listeners exhibited considerable front/back reversals (Figure 4.30). This was evidenced by the near chance performance of listeners initially walking in the correct direction of the sound source (forward or backwards). This finding

demonstrates the well-known challenge in spatial audio perception involving cone of confusion difficulties experienced when perceiving a sound while remaining stationary. Once listeners moved in the incorrect direction, they reversed their paths to move towards the sound source. The frequency of front/back reversal rate is similar to that observed in previous studies in which listeners used standardized HRTFs (*Gardner and Martin* (1995); *Wenzel et al.* (1993); *Brungart and Simpson* (2001)). It is possible that the high reversal rates are an artifact of the evaluation metric, which measures the listener's judgment of the direction of the sound before they moved. This type of static localization is known to decrease localization accuracy as compared to dynamic (moving) localization.

There was no difference in the search accuracy for each stimulus in the posttest and posttest conditions. One could expect higher search accuracy for the richer stimuli. However, *Banks and Green* (1973) and *Yost et al.* (1971) found that the discrimination of a sound's angle largely depends on the low-frequency content of the transient. The Crickets stimulus had transients in the lower frequency range(under 1 kHz). Perhaps this is the reason that search accuracy for the Crickets stimulus did not significantly differ from the search accuracy of the richer stimuli.

After receiving training, listeners needed more time to search for latter sources. We measured the amount of time spent searching for each source to determine if listeners learned the positions of other sounds as they walked through the environment. Generally, search times were higher when listeners searched for latter (fourth and fifth) sources, as seen in Figure 4.33. One might expect listeners to spend less time searching for latter sounds in each trial, because their positions were detected during the search for earlier sounds. Although the listeners were aware of the other sound sources in the environment while a target source was being sought, this knowledge did not speed the search for the other sounds within the trial. This finding could imply that there is no information aggregation during search, meaning that the listener

begins the search for each sound as though they had no prior environment knowledge. The increase in time needed to find the last two sources could suggest that the task became harder as other sound sources were located. Nevertheless, the present study created and validated an effective search training procedure. An improvement in search accuracy was observed as an effect of training.

Based on our results, we propose that designers of auditory interfaces should include a self paced training procedure similar to that described in the present chapter, to teach listeners to search for each sound. Listeners should complete one training session, lasting between thirty minutes to an hour to learn the cues needed to find sounds in a VAE. One who uses this training procedure should expect to see improvement similar to the results presented.

CHAPTER V

Auditory Spatial Memory

5.1 Introduction

Since it has been established that listeners can locate sounds in a VAE, the present chapter investigates listeners' memory of sound sources. Given a VAE of randomly positioned sounds, we would like to determine the accuracy at which listeners can remember the environment. Sound location memory is especially critical in VAEs where operators keep track of the positions of multiple sound sources (*Martin et al.* (2011)).

Many cognitive models of spatial memory and navigation attempt to explain the representation of the spatial structure of an environment in memory (*McNamara et al.* (2008)). To assess the extent of a listener's auditory spatial memory, we examine the components of the models deemed important in spatial memory. *Postma and De Haan* (1996) propose that spatial memory is made of three components: object recognition, positional encoding, and object-location binding. Object recognition is the ability to perceive each object. Positional encoding is the ability to recall the position of the objects in the environment. Object-location binding is the ability to recall the specific object that was located at each position. These criteria were used to score and analyze spatial memory in the present chapter. The terms positional encoding and object-location binding are referred to as positioning and labeling accuracy, respectively.

Several studies have characterized auditory spatial memory (*Parmentier and Jones* (2000); *Klatzky et al.* (2002); *Klatzky et al.* (2003); *Martin et al.* (2011)). Typically, in these types of experiments distinct stimuli are presented serially. As a result, pronounced primacy and recency effects are often observed when listeners were asked to recall multiple serial sound sources.

On the other hand, few studies have characterized listeners' memory of multiple concurrent spatial sound sources. Concurrent sound source presentation may not be subject to the primacy and recency effects observed in previous studies.

The principal aim of the present study is to characterize a listener's ability to recall the positions of multiple concurrent sound sources. We examined the effects of various recall methods on accuracy. Three conditions were created in which the amount of time the information was stored, and the recall method were varied. In the first condition, listeners immediately marked positions and after a delay labels were freely recalled. In the second condition, positions were freely recalled after a delay imposed by exploration and labels were freely recalled after positions were recalled. In the third condition, after exploring the environment, listeners were acoustically cued in an ordered recall task to indicate the position of a predetermined sound.

5.2 Experiment: Recall of Auditory Spatial Objects

5.2.1 Methods

The present experiment measured a listener's recall of the configuration of a VAE in three experimental conditions. Listeners were required to memorize the locations and labels of five sound sources in a VAE.

For the first condition, *Immediate Positioning then Delayed Labeling*, the participant walked around the auditory interface, serially marking the locations of each of the five sound sources as they were encountered (similar to the experimental task

presented in the previous chapter). After all of the sound sources were marked, the participant was asked to label their identities. The locations of the sounds were marked as they were encountered and the labeling was recalled one step after the environment was encountered.

In the second condition, *Delayed Positioning then Delayed Labeling*, the participant walked around the auditory environment while memorizing the locations of the five sound sources. After the participant indicated that they had learned the spatial configuration of the environment, they marked the locations of the sound sources. Afterwards, the participant labeled the identities of the sound sources they had previously marked. In this condition, the locations were marked one step after they were encountered and the labels were marked two steps after the environment had been encountered.

In the third condition, *Delayed Positioning and Delayed Labeling*, the participant walked around the auditory environment and memorized the locations of the five sound sources. After the participant indicated that they had learned the spatial configuration of the environment, they were aurally cued with a sound from the interface. After hearing the sound, the participant clicked the position in the interface at which they remembered hearing the sound. In this condition, the locations and the labels were simultaneously retrieved one step after they were encountered.

Before each session, participants completed a training procedure to familiarize themselves with the auditory environment. The training procedure was adopted from Chapter IV. During the two-phase training, the listener first completed a minimum of 20 axial training tasks. Training continued until the participant reached a predetermined accuracy criteria for five consecutive trials. Random-placement training followed axial training, and in a similar fashion, training continued until the participant reached a predetermined accuracy criteria for five consecutive trials. After both of the training phases were completed, the participant completed the desired

experimental condition.

5.2.2 Participants, Stimuli, and Apparatus

The five observers from the previous experiment participated in the current experiments. The three conditions required about 6.5 hours of listening and were completed in three sessions. Each of the three conditions was conducted in separate sessions occurring between two and seven days apart. The present study used the same real-time MATLAB-based spatial auditory system and the same stimuli from Chapter IV.

5.2.3 Procedure

5.2.3.1 Immediate Positioning then Delayed Labeling

After training, all participants completed Immediate Positioning then Delayed Labeling (IP + DL) in one session in an average of 2 hours and 12 minutes. The Immediate Positioning then Delayed Labeling (IP + DL) condition is very similar to the experimental task of Chapter IV. In IP + DL, the listener immediately marked the positions of sound sources as they were encountered, then labeled the identities of the sound sources.

The listener began each trial facing forward in the middle of a silent VAE (Figure 5.1). The five stimuli were presented after the participant indicated that they were ready to begin the trial. The five stimuli played simultaneously at randomly chosen locations within a circle around the listener. Each sound was about 10 seconds long with repeated components that was played in a loop for the duration the trial. The locations of each sound were chosen pseudo-randomly on each trial so that each sound was spaced at least 30° apart and separated by a distance that equaled one-third of the diameter of the auditory space.

The listener's position and orientation were recorded as they moved around the environment to listen and locate the five sound sources. When the listener thought they

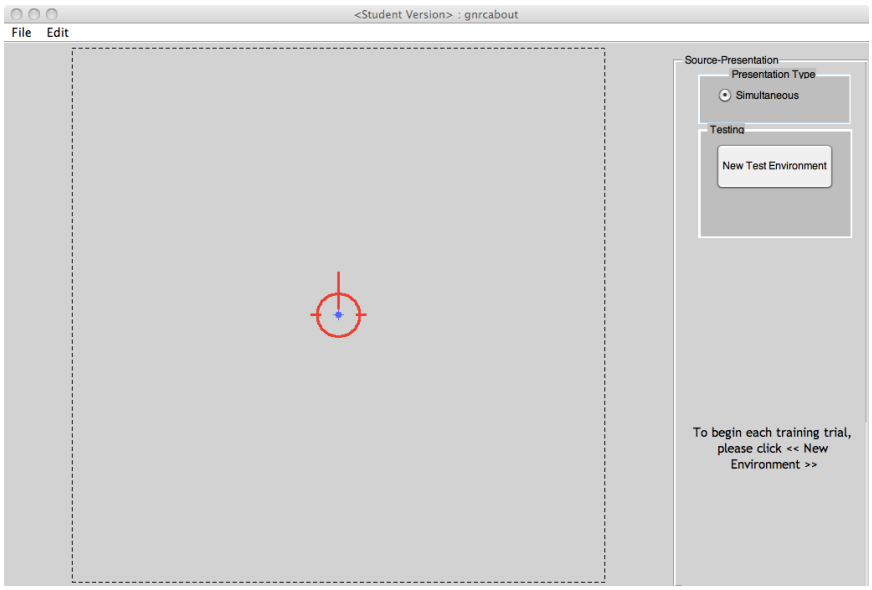


Figure 5.1: Initial interface for the IP + DL condition. Auditory interface shown at the beginning of each trial. The red circle represents the listeners position and the long red line intersecting the circle represents the heading. The listener pressed "New Test Environment" to begin each trial.

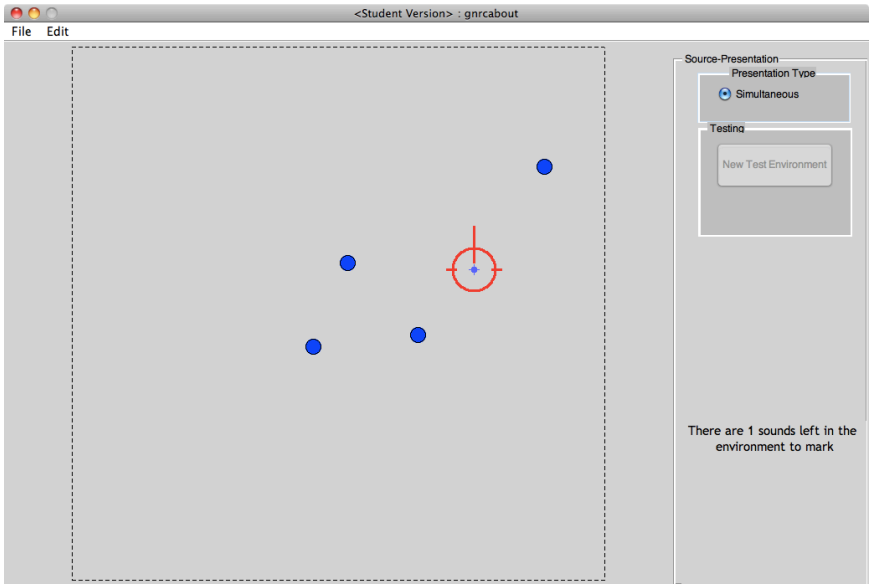


Figure 5.2: Marking sound sources in the IP + DL condition. In this example, the listener has already marked the locations of four of the five sound sources, represented in blue.

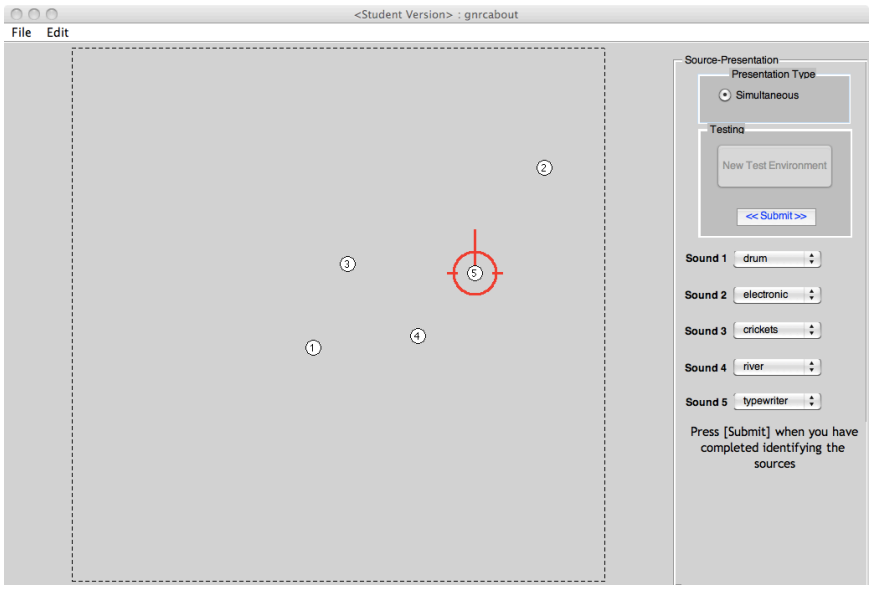


Figure 5.3: In this example, the listener used the five drop-sound lists to label the identities of the corresponding sound sources. After all sounds were labeled, the listener pressed the "Submit" button to continue.

were standing in the same location as the sound source, they indicated its location by pressing the spacebar, which placed a blue marker on the screen. Each participant repeated this process, marking the locations of the remaining sound sources in the environment. A depiction of this procedure can be seen in Figure 5.2.

Once the participant marked all of the sources, all sounds were removed from the environment. At the locations where the listener had marked each sound source, a numeral from 1 to 5 appeared (Figure 5.3). The numbers appeared in the same order as they were marked in the interface with 1 denoting the first-marked sound and 5 denoting the last-marked sound. On the left of the screen, five drop-sound boxes appeared. In the drop-sound box, the listener selected the label of the sound source that corresponded to each sound that had been numbered in the environment. When the listener selected the name of the sound source, it played monaurally for one second. The listener then confirmed and submitted the sound labels.

After the labels were confirmed, the true locations of the sound sources and the participant's responses were shown simultaneously for seven seconds (Figure 5.4). The

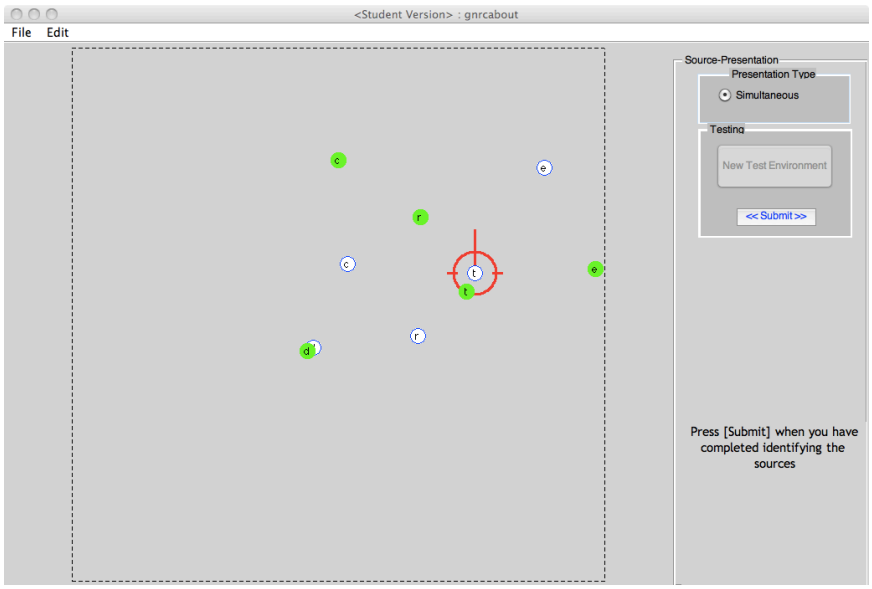


Figure 5.4: Localization feedback in IP + DL condition. The user-marked (white) and true locations (green) of the sound sources were shown for seven seconds. On each marker, the first letter of the sound source label was also displayed.

entire procedure constituted one trial. Each listener completed 20 trials. Participants began each trial in the center of a new configuration of the five sound sources. To avoid auditory fatigue, the participant was allowed to take 1-5 minute breaks between trials, as needed.

5.2.3.2 Delayed Positioning then Delayed Labeling

Next, the listener completed the Delayed Positioning then Delayed Labeling condition (DP + DL). Before the condition, the listener completed the same training procedure that preceded IP + DL. All participants completed DP + DL in one session, in an average of two hours. In this condition, the listener explored the environment, while memorizing the locations and labels of each sound. After learning the environment, the listener marked the locations of the sound sources, then labeled the identities of the marked sound sources.

As in the previous condition, the listener began each trial facing forward in the

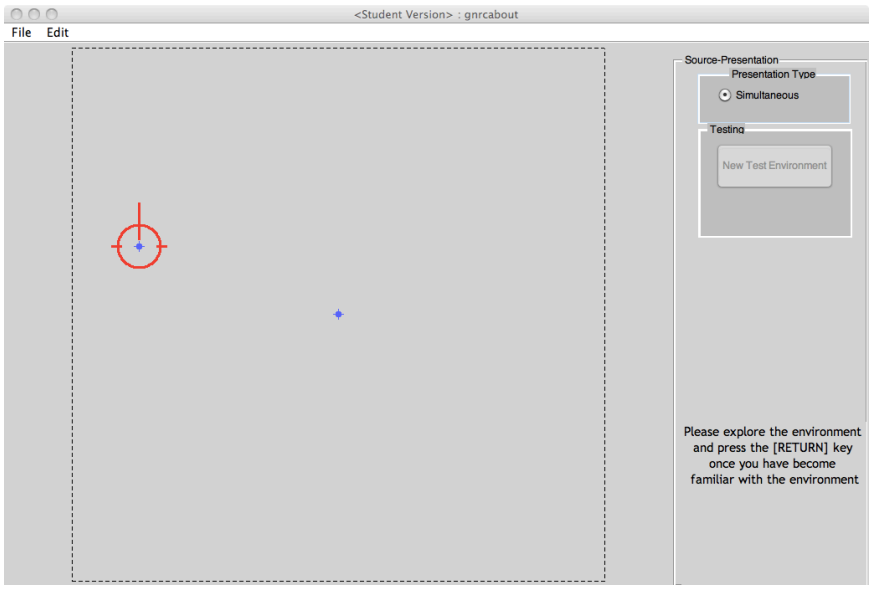


Figure 5.5: Exploration during the DP + DL condition. The red circle represents the listeners position and the long red line intersecting the circle represents the heading. This was the interface seen as the listener moved in the environment, while localizing each sound source.

middle of a silent VAE. The five stimuli were presented after the participant indicated that they were ready to begin the trial. The five stimuli played simultaneously at randomly chosen locations within a circle around the listener. Each sound was about 10 seconds long with repeated components, played in a loop for the duration the trial. The location of each sound was chosen pseudo-randomly on each trial so that each sound was spaced at least 30° apart and separated by a distance that equaled one-third of the diameter of the auditory space. The listener's position and orientation were recorded as they moved around the environment to learn the locations of the five sound sources (Figure 5.5). Once the locations and identities of the sound sources had been memorized, the listener began the recall task in which they were instructed to use the mouse to click the five locations in the interface where the sound sources were positioned. The order in which the sounds were marked did not matter (Figure 5.6).

At the places where the listener had marked each sound source, a numeral from

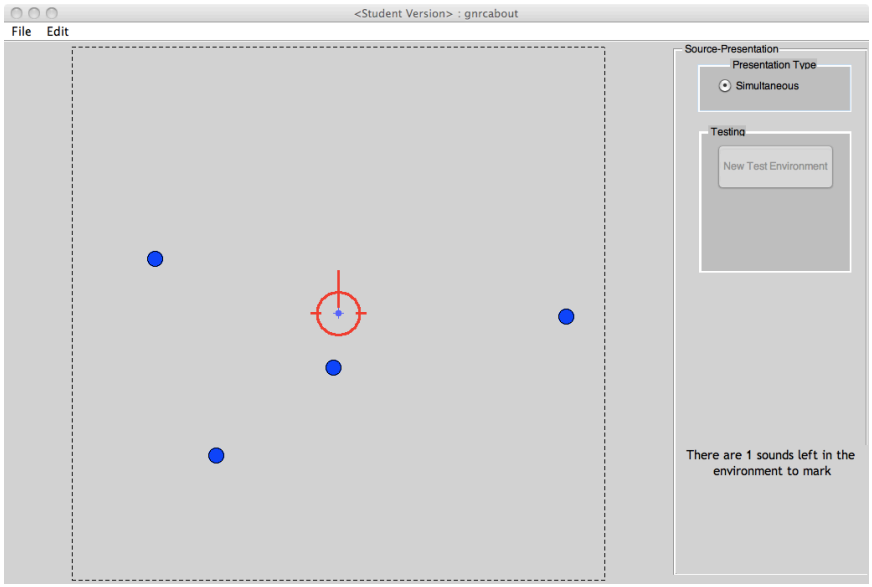


Figure 5.6: Marking sound sources in the DP + DL condition. In this example, the listener has already marked the locations of four of the five sound sources, represented in blue.

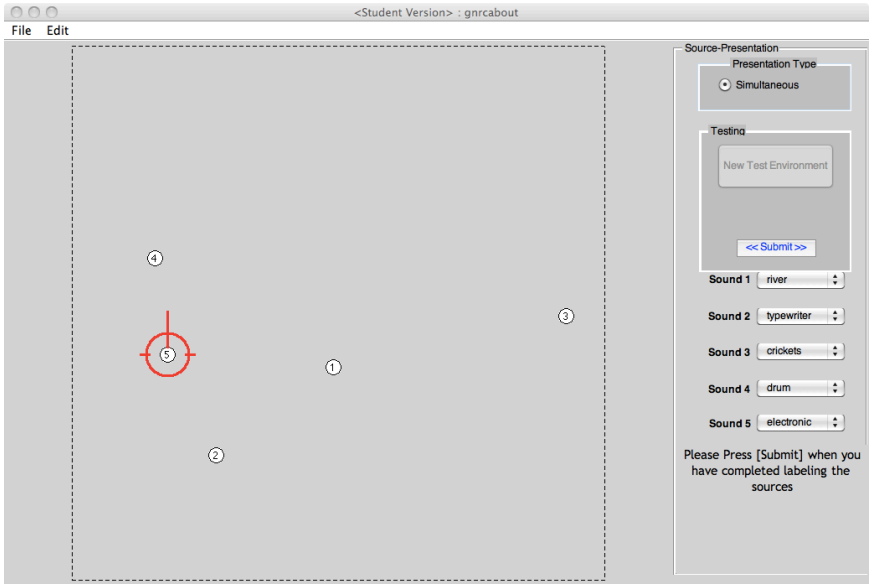


Figure 5.7: In this example, the listener used the five drop-sound lists to label the identities of the corresponding sound sources. After all sounds were labeled, the listener pressed the "Submit" button to continue.

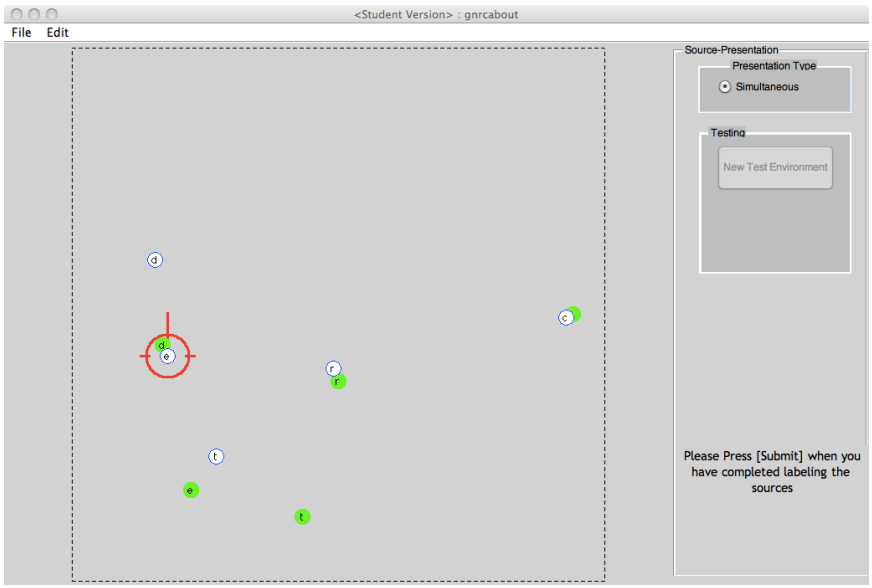


Figure 5.8: Localization feedback in DP + DL condition. The user-marked (white) and true locations (green) of the sound sources were shown for seven seconds. On each marker, the first letter of the sound source label was displayed.

1 to 5 appeared (Figure 5.7). The numerals appeared in the same order as they were marked in the previous step. On the left of the screen, five drop-sound boxes appeared. The listener selected the label of the sound source that corresponded to each sound that had been numbered in the environment. When the listener selected the name of the sound source, the sound source played monaurally for one second. Once the listener confirmed the labels, the true locations of the sound sources and the participant' responses were shown simultaneously for seven seconds(Figure 5.8). This entire procedure constituted one trial. Each listener completed 20 trials. The participant began each trial in the center of a new configuration of the five sound sources. To avoid auditory fatigue, the participant was allowed to take 1-5 minute breaks between trials as needed.

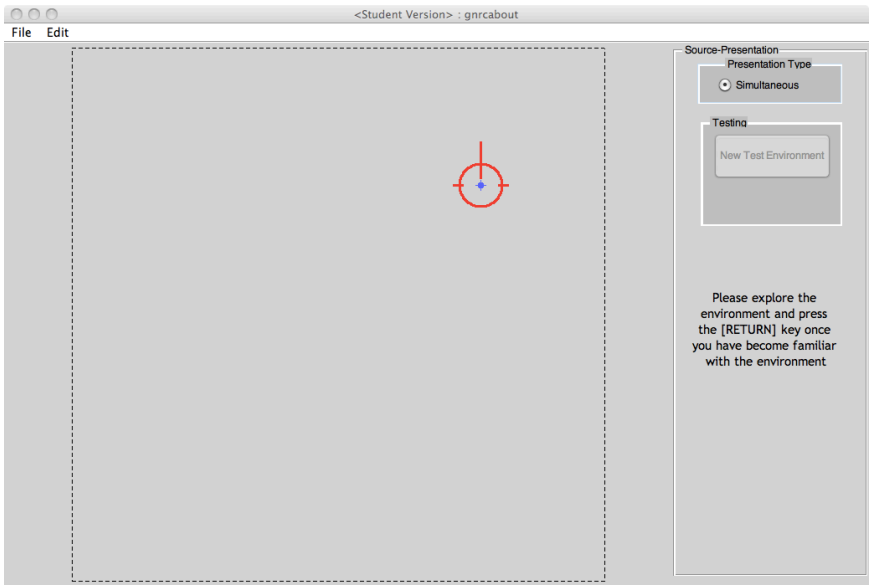


Figure 5.9: Exploration during the DPDL condition. The red circle represents the listeners position and the long red line intersecting the circle represents the heading. This was the interface seen as the listener moved in the environment, while localizing each sound source.

5.2.3.3 Delayed Positioning and Delayed Labeling

Finally, the listener completed the Delayed Positioning and Delayed Labeling (DPDL) condition. First, the listener completed the same training procedure that preceded conditions 1 and 2. All participants completed DPDL in one session, in an average of 2 hours and 18 minutes. In this condition, the listener explored the environment while memorizing the locations and labels of each sound. Next, the listener was randomly monaurally cued with each sound and was asked to click the location corresponding to the sound that played. In this condition, positioning and labeling recall occurred simultaneously.

As in the previous condition, the listener began each trial facing forward in the middle of a silent VAE. The five stimuli were presented after the participant indicated that they were ready to begin the trial. The five stimuli played simultaneously at randomly-chosen locations within a circle around the listener. Each sound was about 10 seconds long with repeated components, played in a loop for the duration the trial.

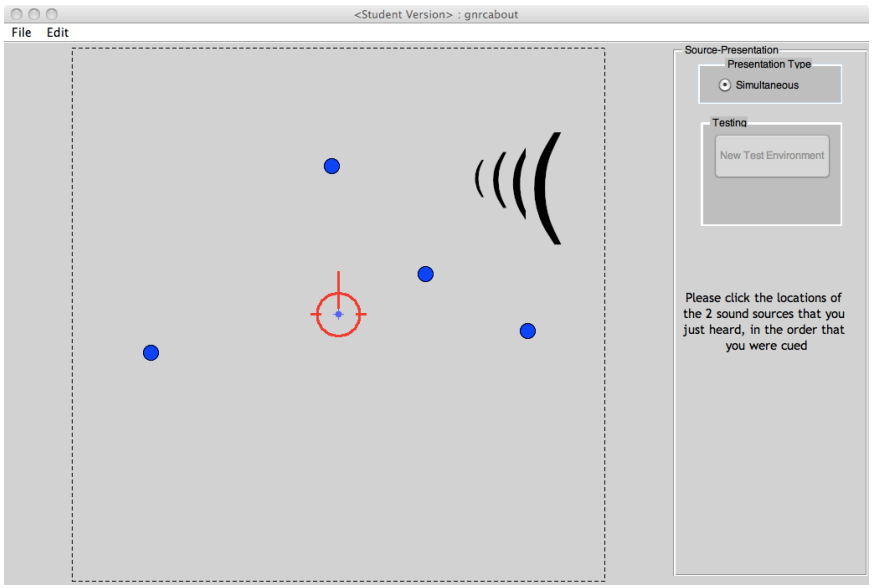


Figure 5.10: Acoustic cueing during the DPDL condition. In this example, the listener has heard and marked the locations of the first four sounds and is listening to the fifth sound (denoted in black).

The location of each sound was chosen pseudo-randomly on each trial so that each sound was spaced at least 30° apart and separated by a distance that equaled one-third of the diameter of the auditory space. The listener's position and orientation were recorded as they moved around the environment to learn the locations of the five sound sources (Figure 5.9). The listener indicated when the locations and identities of the sound sources had been memorized.

Next, the participant was monaurally cued with two randomly-selected sounds from the trial (Figure 5.10). After hearing the two sounds, the listener clicked the two locations at which they recalled hearing the two sounds, using the same order in which they were played. A blue circle was placed at the clicked locations. Next, in the same fashion, the listener heard two more sounds and indicated their positions. Finally, the listener indicated the position of the fifth sound.

As in the two previous conditions, after the five sounds had been marked, the true locations of the sound sources and the participant's responses were shown simultaneously for seven seconds (Figure 5.11). This entire procedure constituted one trial.

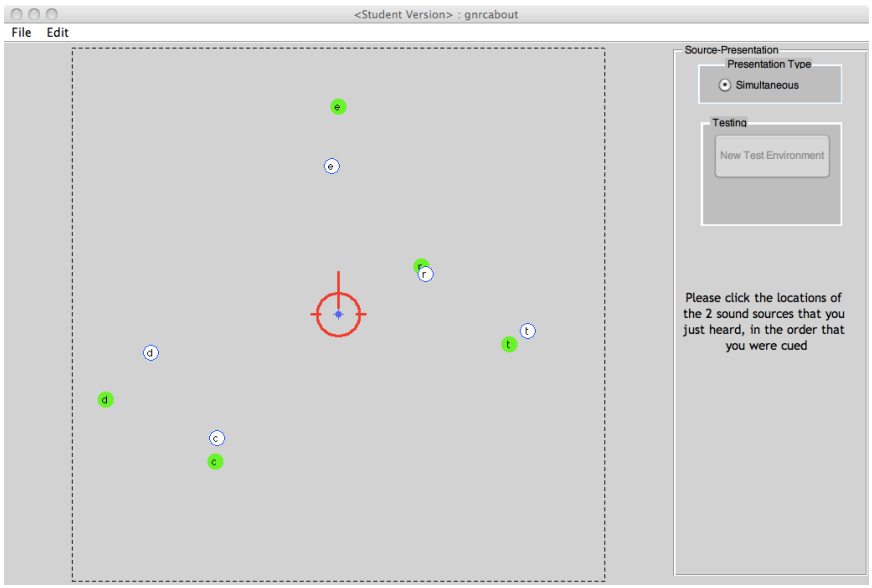


Figure 5.11: Localization feedback in DPDL condition. The user-marked (white) and true locations (green) of the sound sources were shown for seven seconds. On each marker, the first letter of the sound source label was displayed.

Each listener completed 20 trials. Participants began each trial in the center of a new configuration of the five sound sources. To avoid auditory fatigue, the participant was allowed to take 1-5 minute breaks between trials as needed.

5.3 Results

The data of the present experiment was evaluated using the same metrics and statistical treatment as the experiment of Chapter IV. In the conditions in which the memory of locations was involved (DP+DL and DPDL), multidimensional scaling was used to determine a linear transformation of the user's marked locations that best conformed to the real point configuration. Dissimilarity was then measured between the transformed configuration and the actual locations of the sound sources. This was done to account for any overall configuration scaling or shifting while marking the sources.

As previously mentioned, labeling error is an additional metric that was used to

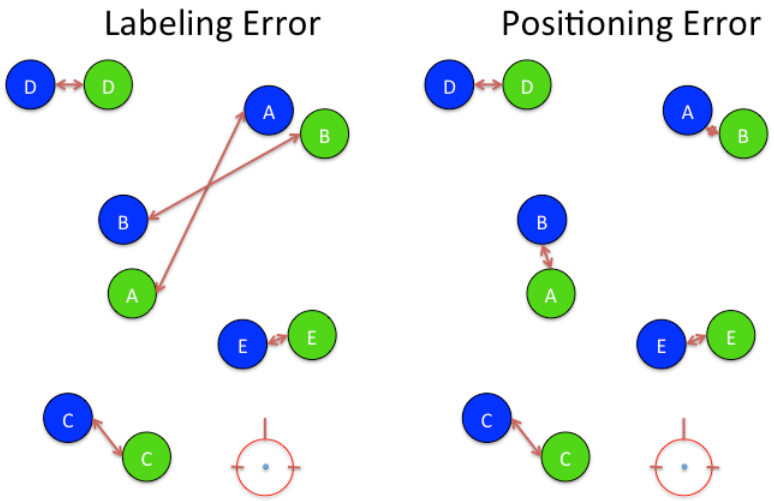


Figure 5.12: Labeling error (left) compared to positioning error (right). The blue circles represent the locations marked by the user and the green circles represent the actual locations of the sound sources. The red arrows represent the distance used to calculate the error.

analyze the data. It describes the difference between a sound's real location and its recalled location. This measure was introduced to distinguish the difference (if any) between recalling the configuration of the environment (positioning error) as compared to correctly recalling each source in its correct location. Figure 5.12 shows two panels with the same configuration of sound sources. The blue circles indicate the locations of the sound sources A through E, as marked by the user. The green circles indicate the true locations. The red arrows represent the distance used in the error calculation. In the left panel, labeling error is measured as the distance between each circle with the same label, while positioning error is measured as the minimum distance between each true source and the source marked by the listener, regardless of its label.

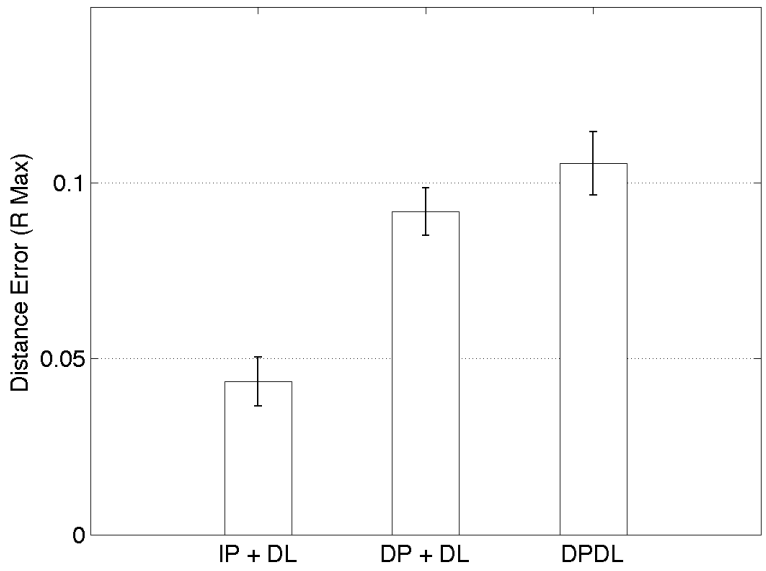


Figure 5.13: Effects of auditory spatial memory on positioning accuracy. Along the abscissa are the conditions and along the ordinate is the positioning error.

5.3.1 Effects of Recall Method

5.3.1.1 Positioning Accuracy

Figure **??** compares the mean positioning error in the IP+DL, DP+DL and DPDL conditions collapsed across trials. Positioning error differed significantly between the conditions [$F_{2,1497}$=70.26, $p<0.05$]. A Tukey LSD multiple comparison test shows that positioning error was significantly higher in the DPDL condition than in the DP+DL condition. Additionally, positioning error was higher in the DP+DL condition than the IP+DL condition.

5.3.1.2 Angular Accuracy

Figure 5.14 compares the mean angular error in the IP+DL, DP+DL and DPDL conditions. Angular error differed significantly between the conditions [$F_{2,1497}$=42.53, $p<0.05$]. A Tukey LSD multiple comparison test shows that angular error was signif-

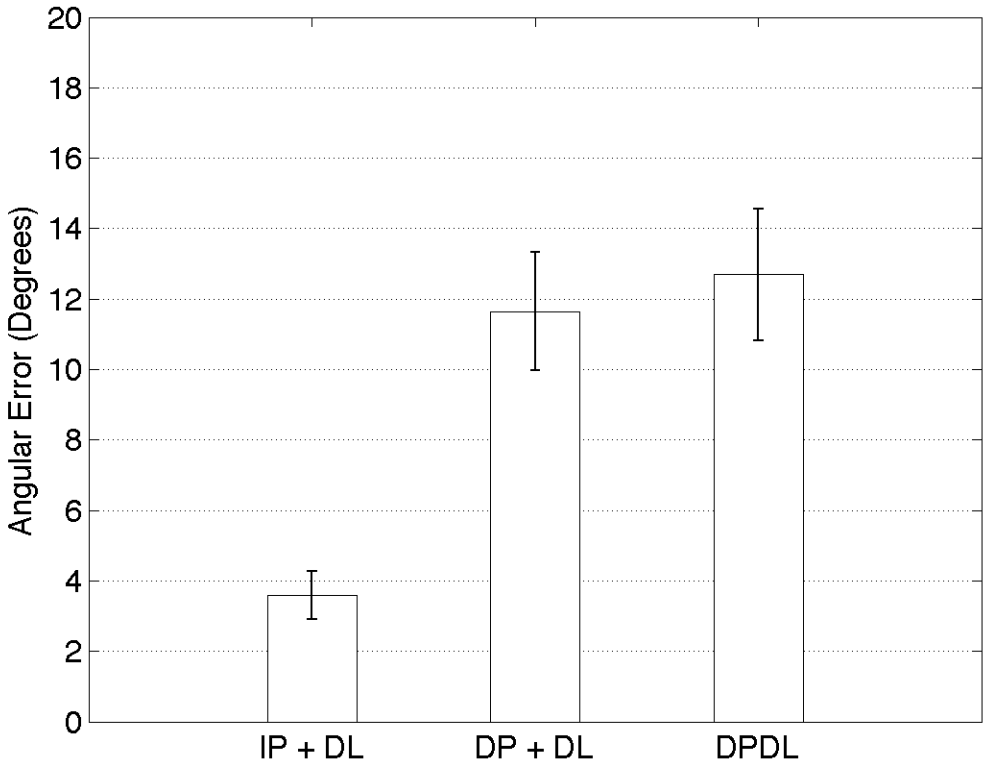


Figure 5.14: Effects of auditory spatial memory on angular accuracy. Along the abscissa are the conditions and along the ordinate is the angular error.

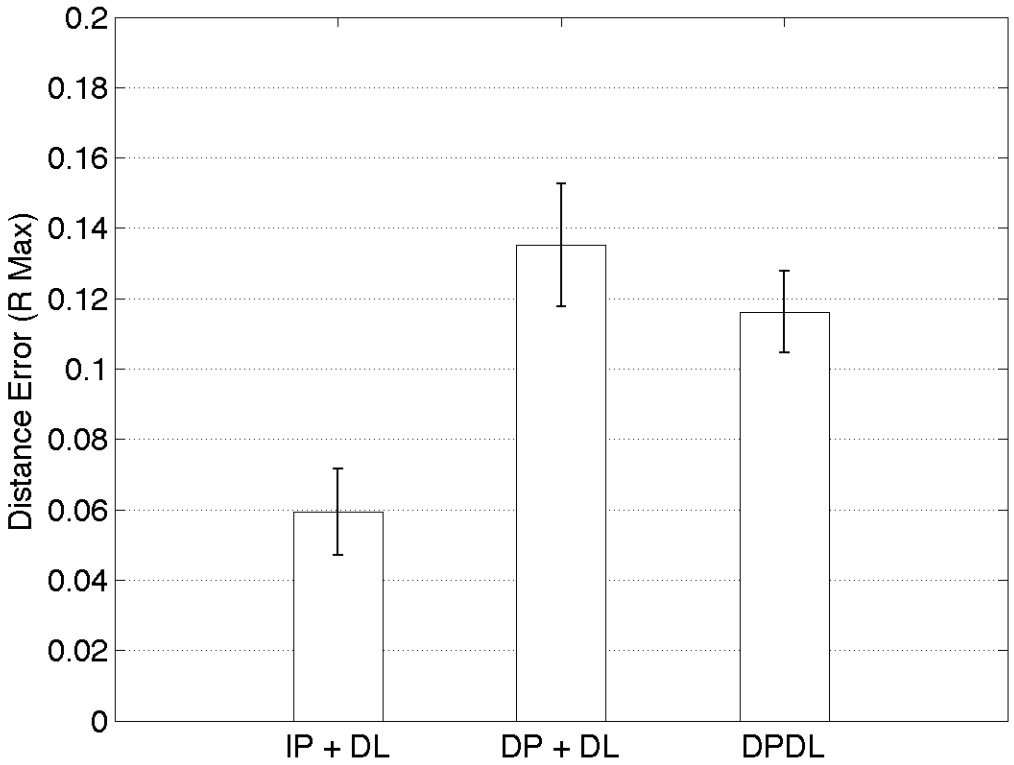


Figure 5.15: Effects of auditory spatial memory on labeling accuracy. Along the abscissa are the conditions and along the ordinate is the labeling error.

icantly lower in the IP+DL condition than in the DP+DL and DPDL conditions.

5.3.1.3 Labeling Accuracy

Similarly, Figure 5.15 compares the mean labeling error in the IP+DL, DP+DL and DPDL conditions. The results were similar to those of the angular accuracy data. Labeling error differed significantly between the conditions [$F_{2,1497}$=30.29, $p<0.05$]. A Tukey LSD multiple comparison test shows that angular error was significantly lower in the IP+DL condition than in the DP+DL and DPDL conditions.

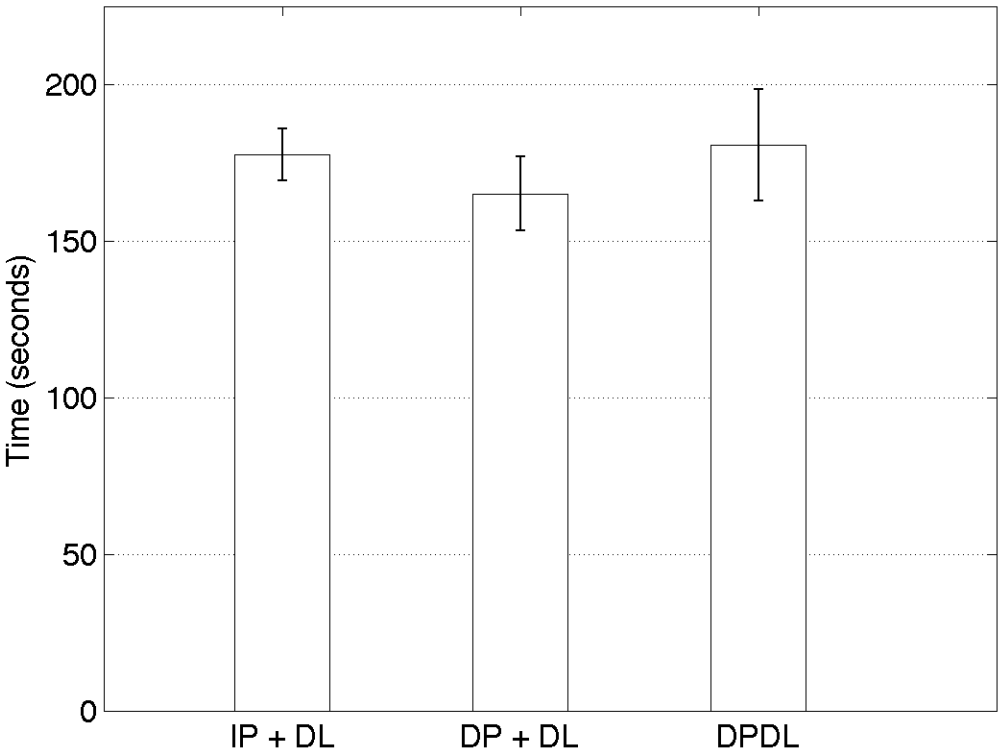


Figure 5.16: Effects of auditory spatial memory on exploration time. Along the abscissa are the conditions and along the ordinate is the exploration time.

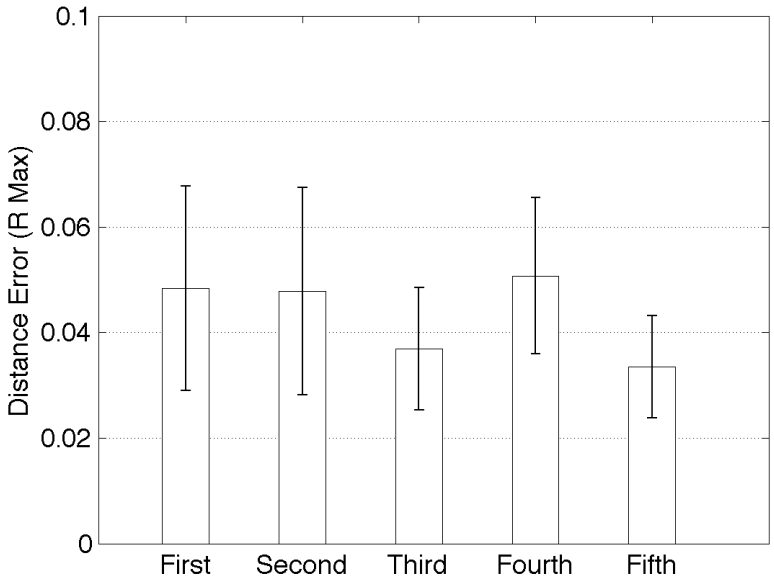


Figure 5.17: Effects of sequence on positioning accuracy in IP + DL condition. Along the abscissa is the sequence and along the ordinate is the positioning error.

5.3.1.4 Exploration Time

Figure 5.16 compares the mean exploration time needed to find each sound source in the IP+DL, DP+DL and DPDL conditions. No significant difference was observed between the conditions [$F_{2,397}$=1.54, p=0.22].

5.3.2 Sequence Effects

For each condition, the positioning, angular and labeling error was analyzed to determine the presence of sequence effects. One may expect error to increase with sequence order as sounds are marked and labeled.

5.3.2.1 IP + DL

Figure 5.17 shows the mean positioning error of the first through fifth sequentially marked sound sources in the IP+DL condition. No significant difference in marking order was observed [$F_{4,495}$=0.97, p=0.43].

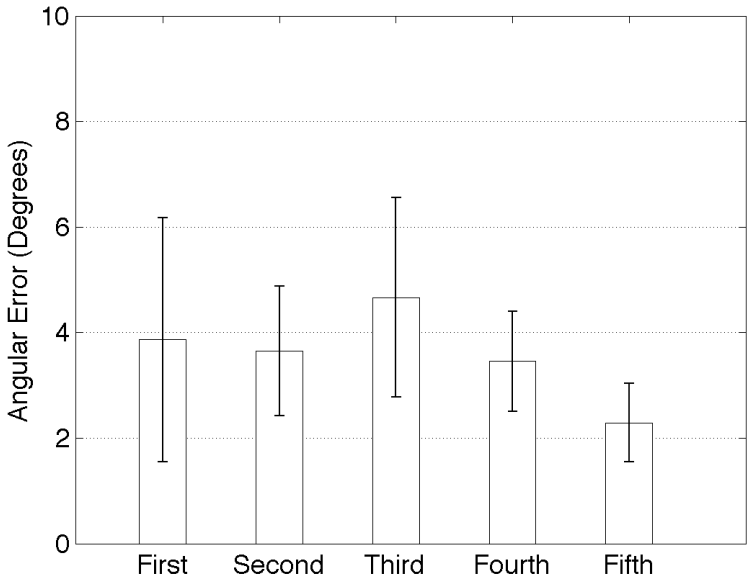


Figure 5.18: Effects of sequence on angular accuracy in IP + DL condition. Along the abscissa is the sequence and along the ordinate is the angular error.

Likewise, Figure 5.18 compares the mean angular error that listeners made while completing the IP+DL condition. Similarly, no difference in marking order was observed [$F_{4,495}$=1.22, p=0.30].

Figure 5.19 compares the mean labeling error that listeners made while completing the IP+DL condition. As in the previous two analyses, no difference in labeling accuracy was observed [$F_{4,495}$=0.21, p=0.93].

5.3.2.2 DP + DL

As in the IP+DL condition, the positioning, angular and labeling error for each source in the DP+DL condition was assessed. Figure 5.20 shows that there was there was no difference in positioning accuracy as an effect of sequence [$F_{4,495}$=0.58, p=0.68].

Figure 5.21 compares the mean angular error by marking order in the DP+DL condition. Interestingly, there was a significant difference observed [$F_{4,495}$=2.92, p<0.05].

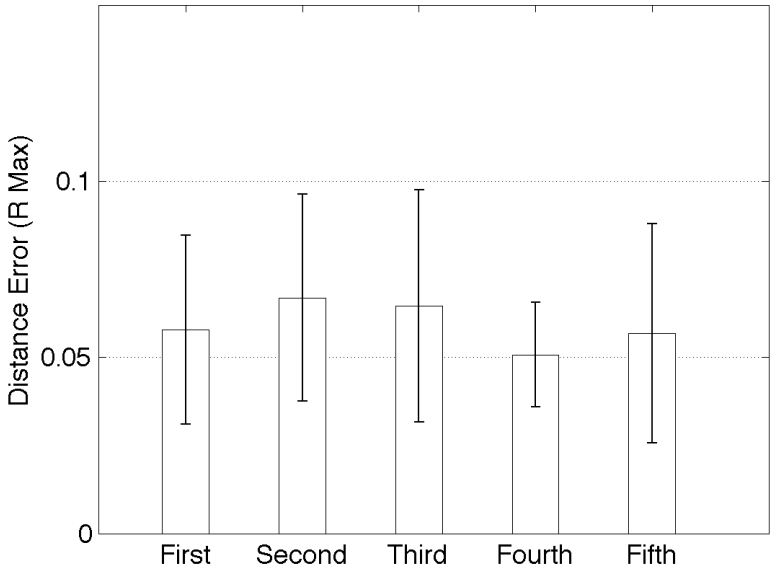


Figure 5.19: Effects of sequence on labeling accuracy in IP + DL condition. Along the abscissa is the sequence and along the ordinate is the labeling error.

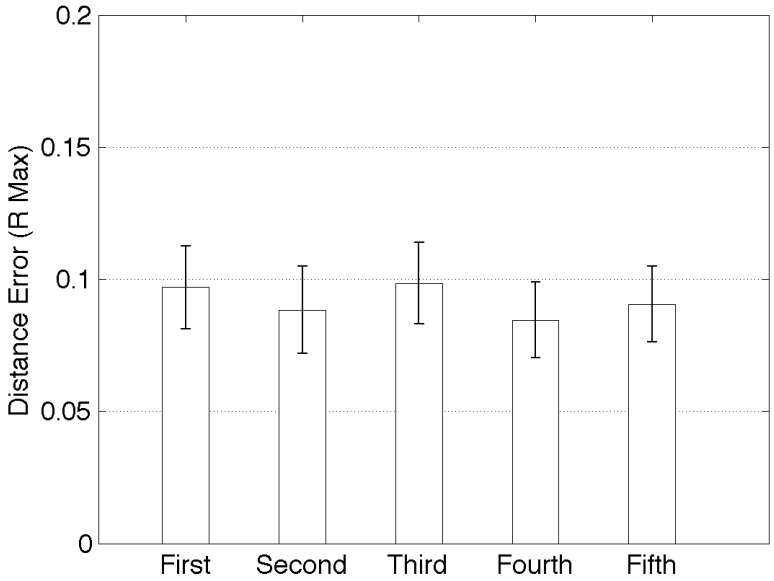


Figure 5.20: Effects of sequence on positioning accuracy in DP + DL condition. Along the abscissa is the sequence and along the ordinate is the positioning error.

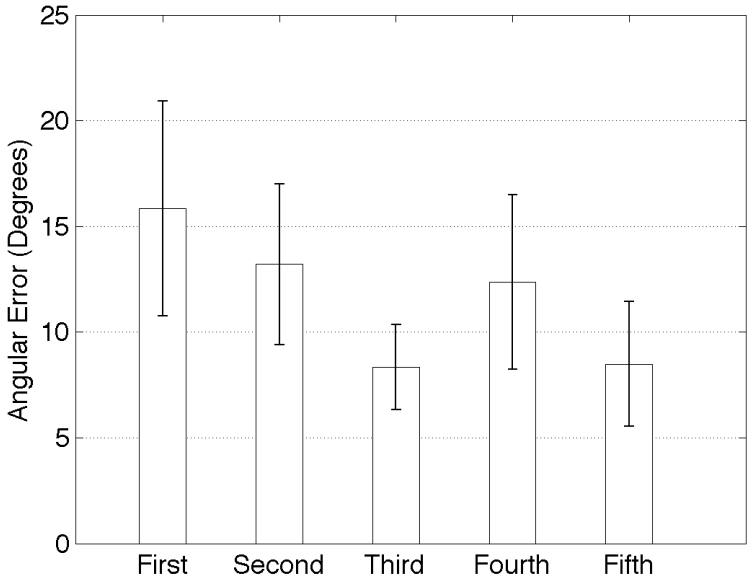


Figure 5.21: Effects of sequence on angular accuracy in DP + DL condition. Along the abscissa is the sequence and along the ordinate is the angular error.

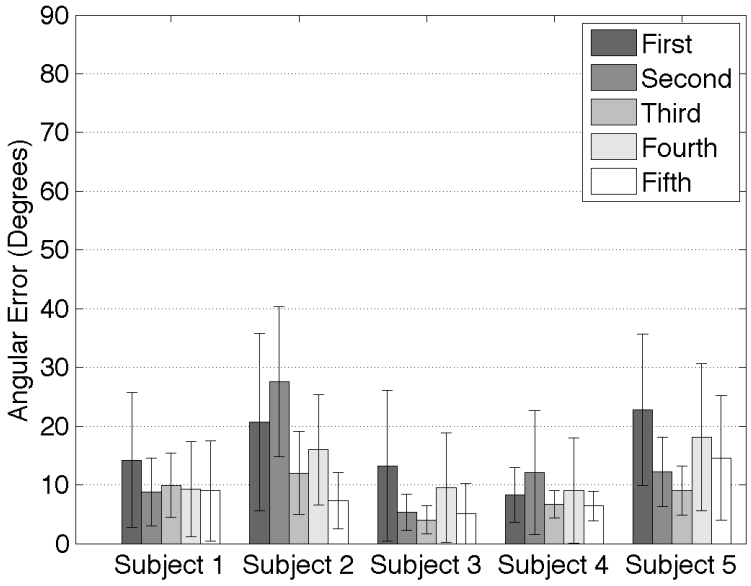


Figure 5.22: Order effects on angular accuracy in DP + DL condition, by subject. Along the abscissa are the subjects and along the ordinate is the angular error.

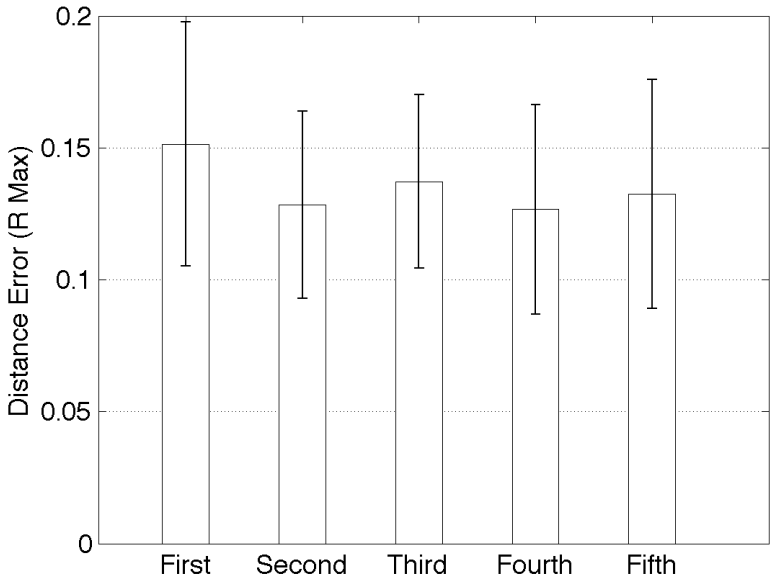


Figure 5.23: Effects of sequence on labeling accuracy in DP + DL condition. Along the abscissa is the sequence and along the ordinate is the labeling error.

The data was further broken down by subject (Figure 5.22) and an additional ANOVA was performed. For each subject, no significant difference in source sequence order was observed.

Figure 5.23 compares the mean labeling error that listeners made in the DP+DL condition. No difference in labeling accuracy as an effect of sequence was observed [$F_{4,495}$=0.25, p=0.91].

5.3.2.3 DPDL

In the same way as the previous two conditions, the positioning, angular and labeling error for each source was analyzed to measure any sequence effects in the DPDL condition. Figure 5.24 illustrates that no difference in positioning error was observed [$F_{4,495}$=0.23, p=0.92].

Likewise, Figure 5.25 compares the mean angular error by sequence order in the DPDL condition. There was there was no difference in angular error [$F_{4,495}$=0.26,

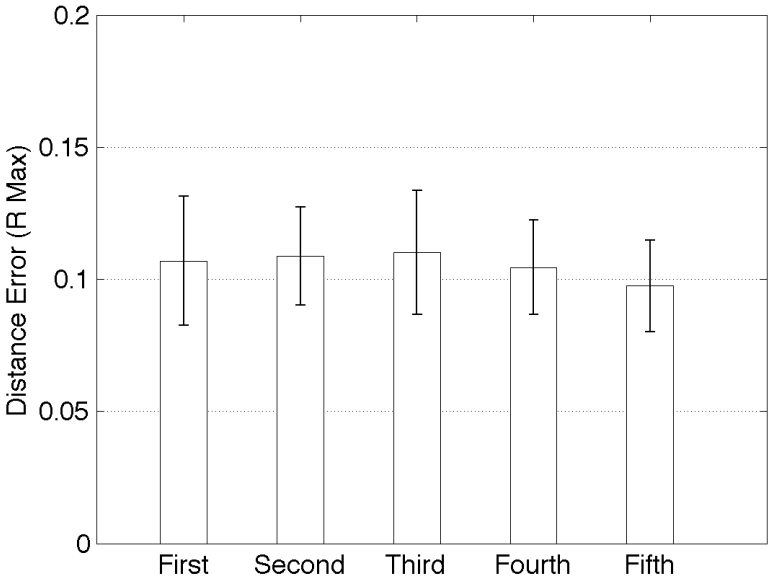


Figure 5.24: Effects of sequence on positioning accuracy in DPDL condition. Along the abscissa is the sequence and along the ordinate is the positioning error.

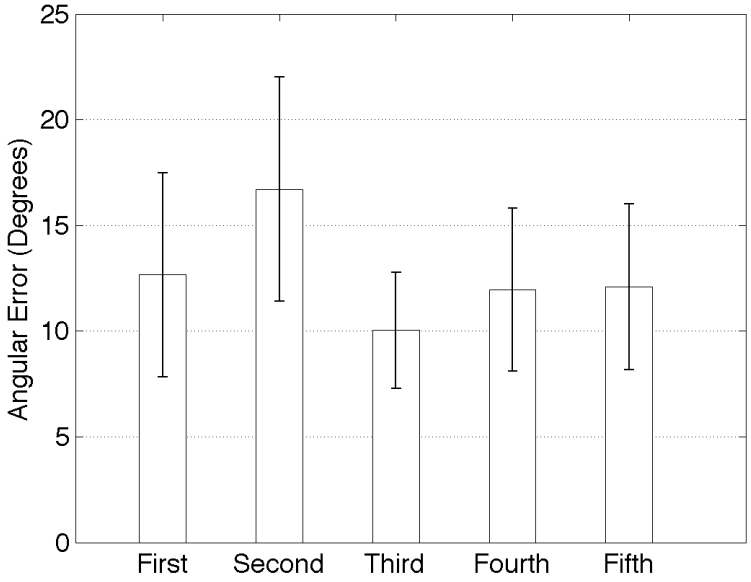


Figure 5.25: Effects of sequence on angular accuracy in DPDL condition. Along the abscissa is the sequence and along the ordinate is the angular error.

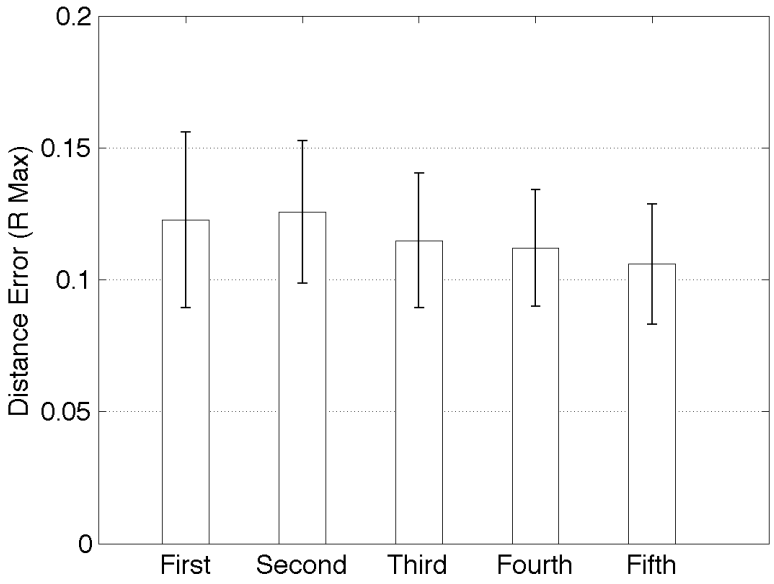


Figure 5.26: Order effects on labeling accuracy in DPDL condition. Along the abscissa is the sequence and along the ordinate is the labeling error.

p=0.26].

Figure 5.26 compares the mean labeling error in the DPDL condition. There was there was no difference in labeling accuracy as an effect of sequential order [$F_{4,495}$=0.36, p=0.84].

5.4 Discussion

The present chapter examined the effects of recall methods on auditory spatial memory of concurrent sound sources. The concept of auditory spatial memory was further explored by examining three conditions that imposed different memory requirements and recall methods. In the IP+DL condition, listeners were only required to remember the labels of marked sounds in a free recall task. In contrast, in the DP+DL condition, the listener was asked to recall location and label in a delayed free recall task. In the third condition, DPDL, the listener was asked to recall location and label in a delayed ordered recall task.

Temporal effects are typically observed in recall tasks (*Rundus* (1971)). Primacy and recency effects are also observed in the recollection of visual spatial objects. *Bonanni et al.* (2007) found that listeners more accurately recall early items in a sequence of visual spatial positions. Recency effect was only observed in the recall of longer sequences of spatial objects. Temporal effects were not observed in any of the conditions of this experiment. This observation may have been due to the fact that sounds were presented concurrently and the listener could revisit any sound while memorizing its location. However, it is also possible that listeners memorized and marked the sound positions in a specific order in each trial, therefore making the labeling task trivial. It is also possible that if listeners were asked to recall more sound sources, temporal effects may have been observed.

Participants exhibited significantly lower positioning and angular error when marking the sound sources in the IP+DL condition as compared to the DP+DL and DPDL conditions. This finding is not surprising given that the IP+DL condition did not require listeners to store the positions in memory. There was no difference in DP+DL and DPDL positioning accuracy. This suggests that listeners are able to remember the environment configuration after a delay, regardless of the recall method (free or ordered). In the DP+DL and DPDL conditions, location and labeling information memory were retained after stimulus presentation, possibly increasing the task demand of spatial information retrieval, resulting in degraded recall accuracy.

Participants exhibited the most labeling error in the ordered recall condition (DPDL). Additionally, when comparing the two free recall tasks, listeners showed more labeling error in the DP+DL condition than in the IP+DL condition. This finding suggests that the increase in elapsed time before labeling negatively affects the listener's ability to remember the correct location of each sound. This is not surprising given that after a delay, when mental processes are recalled, they tend to be over-generalized, over-summarized and over-rationalized and the stored repre-

sentation may not be an accurate depiction of the environment (*Nisbett and Wilson* (1977)).

Participants differed significantly in the amount of time needed to explore the auditory environment during the DPDL condition, as compared to the DP+DL condition. This is not surprising in light of the differing requirements. DPDL potentially imposes a greater memory demand than the other tasks, so listeners may have consciously used more time to memorize the environment, due to the challenging nature of the recall task. Another surprising finding is that one would expect listeners in the IP+DL condition to require significantly less exploration time, as the task imposed the least memory demands - yet the exploration time was not significantly different than the other two conditions.

The findings of the present chapter have important implications for the designers and users of auditory spatial systems. Results suggest that designers of auditory interfaces for sound search should create systems that minimize the amount of time that listeners must store location and label information in memory before recall. Also, when querying a VAE system operator (perhaps during a shift change), free recall should be utilized rather than cued recall.

CHAPTER VI

Three Issues in System Integration

6.1 Introduction

Prior to integrating spatial audio into real-time systems, there are system integration issues that must be explored. The experiments of the previous chapters affirmed that listeners could search for and recall sound sources under precise controlled conditions. These conditions may not be representative of the actual situations that a system operator may face. It is necessary to determine how much (if any) of the observed behavior applies to practical situations.

This chapter examines the implications of three practical circumstances. For example, we measured the effects of sound source uncertainty, wherein the listener must find and recall unknown sources. In a real-life system, it is likely that the operator will have visual cues that correspond to the auditory cues. The present chapter examines the interaction of the visual and audio cues and the effect on sound search and recall. Lastly, the experiments of the preceding chapters have only assessed listeners' behavior when interacting with five sounds. It is very likely that the system operator will need to keep track of more than five sounds. These extra sounds could clutter the interface with sound, making search more difficult. Drastic attenuation modeling may be needed to aid the perception of each sound. Differing attenuation models were investigated to determine their effects on memory and search.

6.2 Effects of Source Uncertainty

The present experiment investigates the effects of source uncertainty on source recall. Previous experiments required listeners to search for five known sources. In a real system, the operator may not have prior knowledge of the sound sources represented within the environment. It is important to investigate how listeners perform when finding and recalling unknown sounds.

The results of this study have critical implications for the design of VAEs that will be used in real systems. Performance could potentially degrade because of the additional classification required to identify the unknown sources. Finding and remembering unknown sound sources could potentially impose an extra cognitive load on the listener. This would force the system designer to make a tradeoff between subject performance degradation and sound source representation. On the other hand, in the more desirable case, listeners could possibly locate and remember unknown sounds with the same accuracy as known sounds. In this case, the operator of the system would be able to localize unknown sources and the system designer would not need to make a design tradeoff.

To test the effects of source certainty, listeners were asked to locate and label sound sources that were randomly drawn from a library of sources. Positioning accuracy, angular accuracy, labeling accuracy, and exploration time were measured and compared to the IP + DL condition of the experiment in Chapter V.

6.2.1 Methods

6.2.1.1 Participants and Apparatus

The five observers from the experiment of the previous chapter participated in the current experiment. The experiment required about 2.2 hours of listening and was completed in a single session apart from the experiments of the previous chapter.

The present study used the same real-time MATLAB-based spatial auditory system from Chapter IV and Chapter V.

6.2.1.2 Stimuli

25 distinct environmental sounds (see Table 6.2.1.2) were used in the present experiment. The sounds were chosen from *BBC* (1991) based upon their ability to be perceptually segregated and the presence of transients. The spectral content of each of the aforementioned sound sources is displayed in Appendix B.

Table 6.1: Larger collection of environmental sounds and their labels

Number	Sound	Labels
1	Alarm clock buzzer	Alarm
2	Audience Applause	Applause
3	Dog barking	Barking
4	Water bubbling	Bubbles
5	Cars accelerating	Cars
6	Crickets chirping	Crickets
7	Crowd yelling	Crowd
8	Drumsticks striking a drum	Drums
9	Electric shock	Electricity
10	Computer generated electronic noises	Electronic
11	Water sizzling in a hot pan	Frying
12	Horse galloping	Horse
13	A man laughing	Laughter
14	Cow Mooing	Moo
15	Gloomy low pitched warble	Ominous
16	Parade drums, whistles and marching	Parade
17	A river flowing rapidly	River
18	Rooster crowing	Rooster
19	Police car sirens	Sirens
20	Loud Snoring	Snoring
21	Electronic ascending whooshes	Space Gun
22	Clock Ticking	Tick-Tock
23	Typewriter keys being pressed	Typewriter
24	Slowly wavering pitch	Wow
25	Wrench tightening a bolt	Wrench

6.2.1.3 Procedure

First, the listener completed a sound recognition task to become familiar with the identities of the new sound sources. The participant serially listened to the 25 sound sources. As each sound was heard, the participant was shown the sound's label. After the serial presentation, the participant could replay any sound. After the participant indicated that all of the sounds had been learned, they completed a sound association test. In the test, a single sound was randomly played. The participant was asked to verbally state the label associated with the sound that had been played. The test continued until the participant was able to correctly label ten consecutive sounds.

Before beginning the experiment, the participant completed training to orient him or herself to the auditory environment. The training procedure used was identical to the training procedure described in Chapter V. 1 of the 25 sound sources was randomly selected to be localized in each training trial.

After training, the participant walked around the auditory interface, serially marking the locations of each of five randomly determined sound sources as they were encountered. After the positions of all of the sound sources were marked, the participant labeled each sound. Essentially, the experimental procedure was identical to the *Immediate Positioning then Delayed Labeling* (IP+DL) experiment of Chapter V.

6.2.2 Results

The results of the current experiment were compared with the results of the IP+DL condition of V, which represented an identical condition in which sound source identity was known. The statistical treatment is similar to that which was used in the previous chapter.

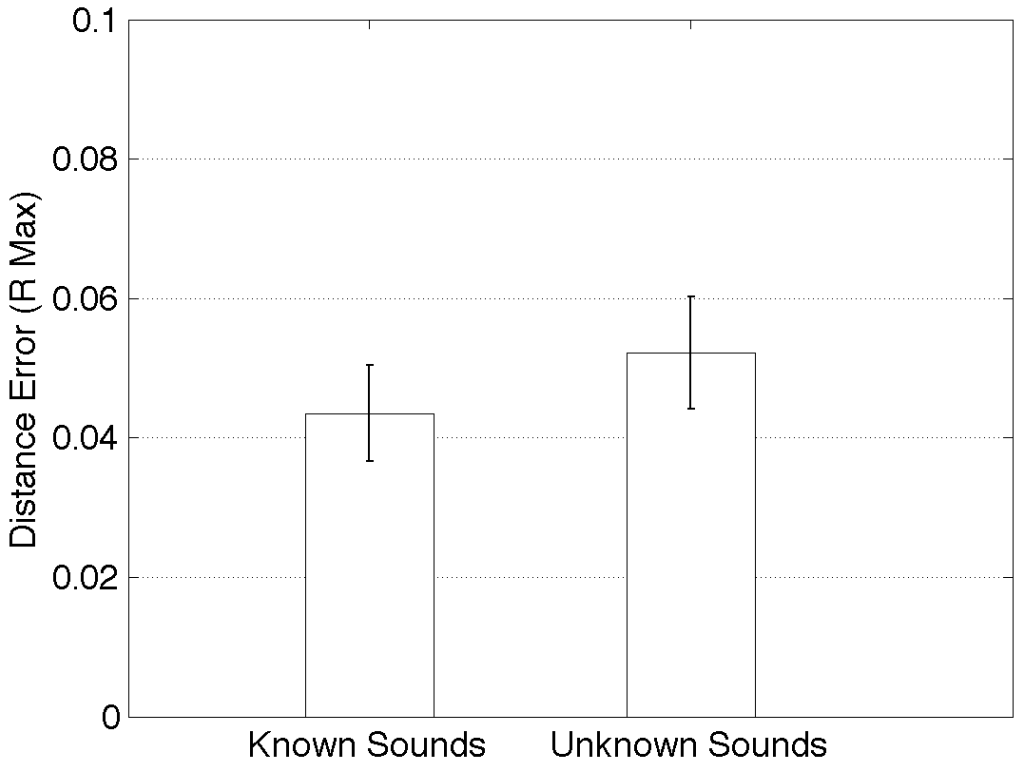


Figure 6.1: Effects of source certainty on positioning accuracy. Along the abscissa is the source certainty and along the ordinate is the positioning error.

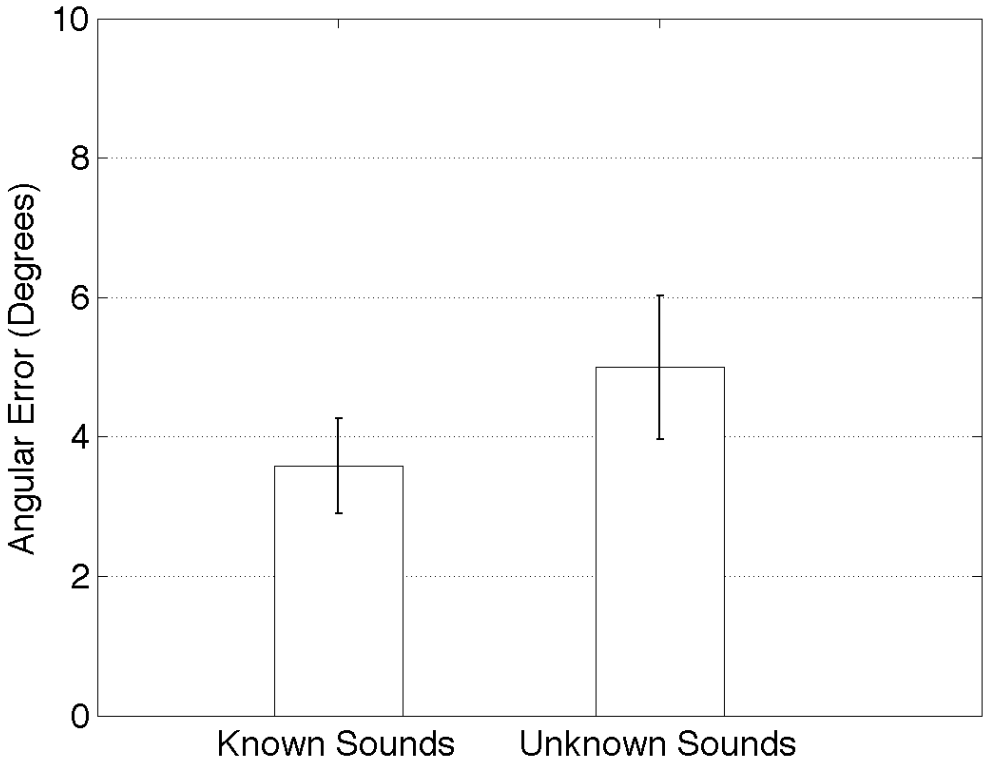


Figure 6.2: Effects of source certainty on angular accuracy, by subject. Along the abscissa is the source certainty and along the ordinate is the angular error.

6.2.2.1 Accuracy

Figure 6.1 shows the mean positioning error for known and unknown sound sources. No significant difference in positioning error was observed as an effect of source certainty [$F_{1,998}$=2.64, p=0.10].

Similarly, Figure 6.2 compares the angular error when searching for known or unknown sources. Positioning error differed significantly between the conditions [$F_{1,998}$=5.02, p<0.05]. The data was further broken down by subject (Figure 6.3) and an additional ANOVA was performed. Angular accuracy did not differ for three of the five subjects. Subject 3's performance significantly improved and Subject 4's performance degraded. No overall trend was observed.

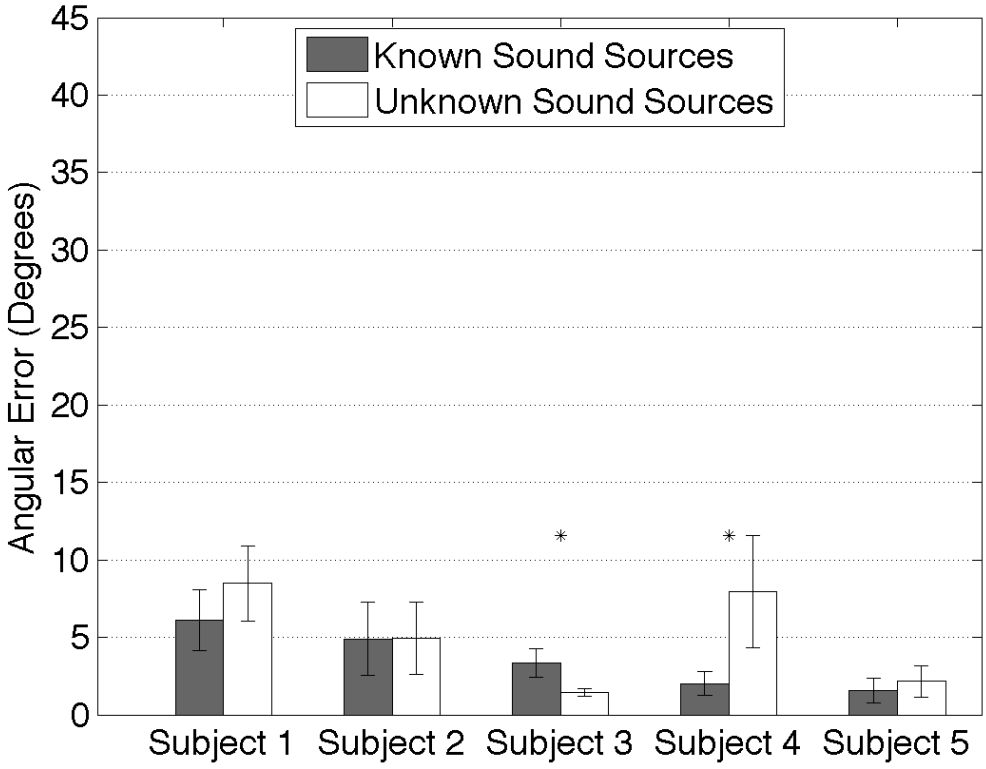


Figure 6.3: Effects of source certainty on angular accuracy, by subject. Along the abscissa is the source certainty and along the ordinate is the angular error.

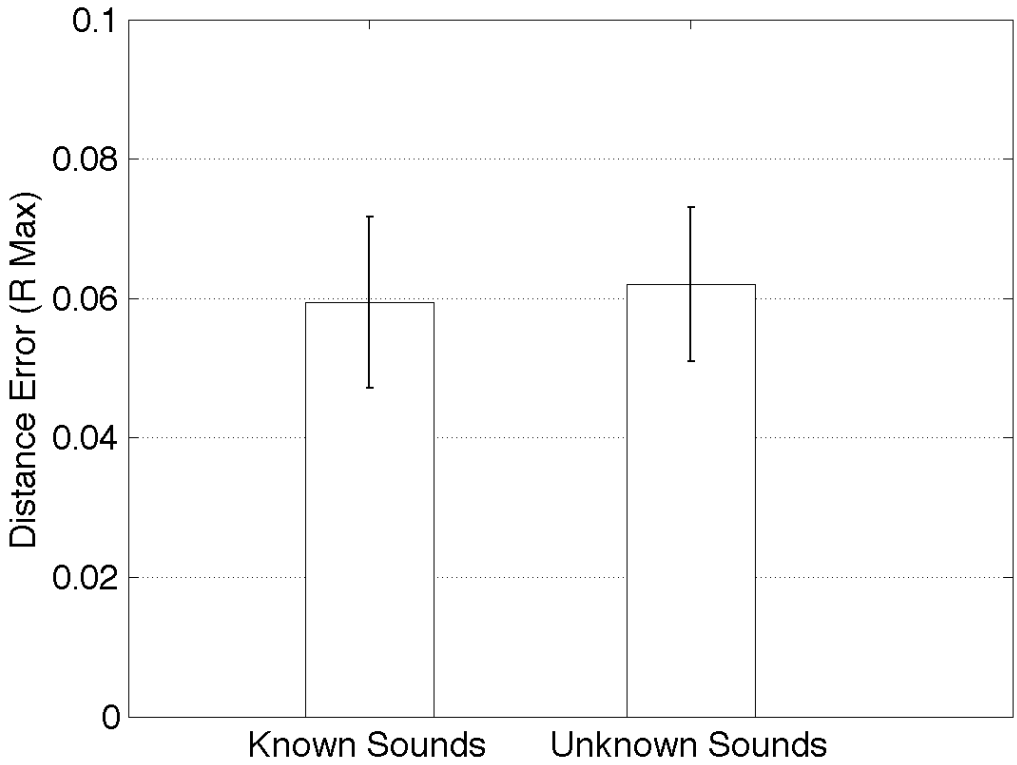


Figure 6.4: Effects of source certainty on labeling accuracy. Along the abscissa is the source certainty and along the ordinate is the labeling error.

Figure 6.4 compares the mean labeling error for known and unknown sources. No significant difference in labeling error was observed as an effect of source certainty [$F_{1,998}$=0.10, p=0.76].

6.2.2.2 Accuracy by Stimulus

Next, the effects of stimulus spectral content on search error were analyzed. Positioning and angular error were analyzed from the condition that occurred after training. The numbered stimuli correspond to the those identified in Table 6.2.1.2.

Figure 6.5 shows the mean positioning error by stimulus during search. A one-way ANOVA indicated that the main effect of stimulus type was insignificant [$F_{24,475}$=1.02, p=0.44].

Figure 6.6 shows the mean angular error by stimulus during search. A one-way ANOVA indicated that there was no significant difference in angular error [$F_{24,475}$=1.14, p=0.29].

Figure 6.7 shows the mean labeling error by stimulus during search. A one-way ANOVA indicated that there was no significant difference in labeling error [$F_{24,475}$=0.9, p=0.6].

6.2.2.3 Exploration Time

Source certainty effects were examined through a comparison of exploration time while searching for known and unknown sources. Figure 6.8 shows the mean exploration time during both conditions. Searching for unknown sources took significantly longer than the search for known sources [$F_{1,160}$=7.64, p<0.05].

6.2.3 Discussion

The results indicated that participants were able to locate and identify randomly determined sound sources with the same accuracy as static sound sources. This

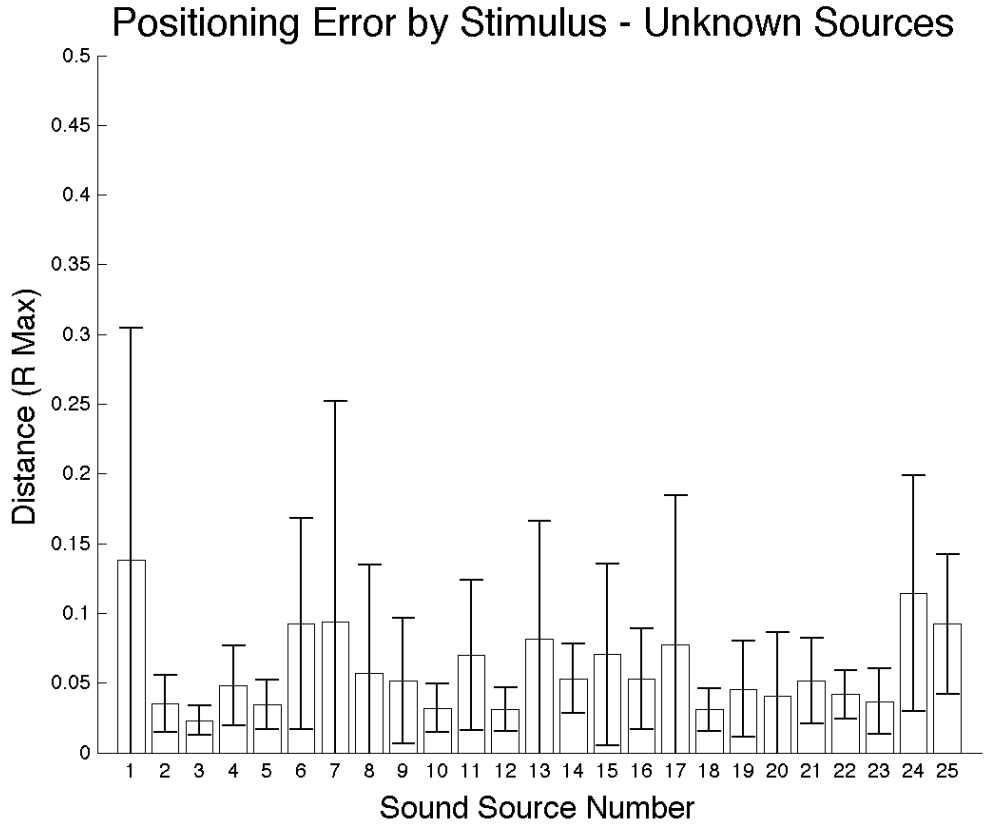


Figure 6.5: Positioning accuracy by stimulus. Along the abscissa is the stimulus number (from Table 6.2.1.2) and along the ordinate is the positioning error.

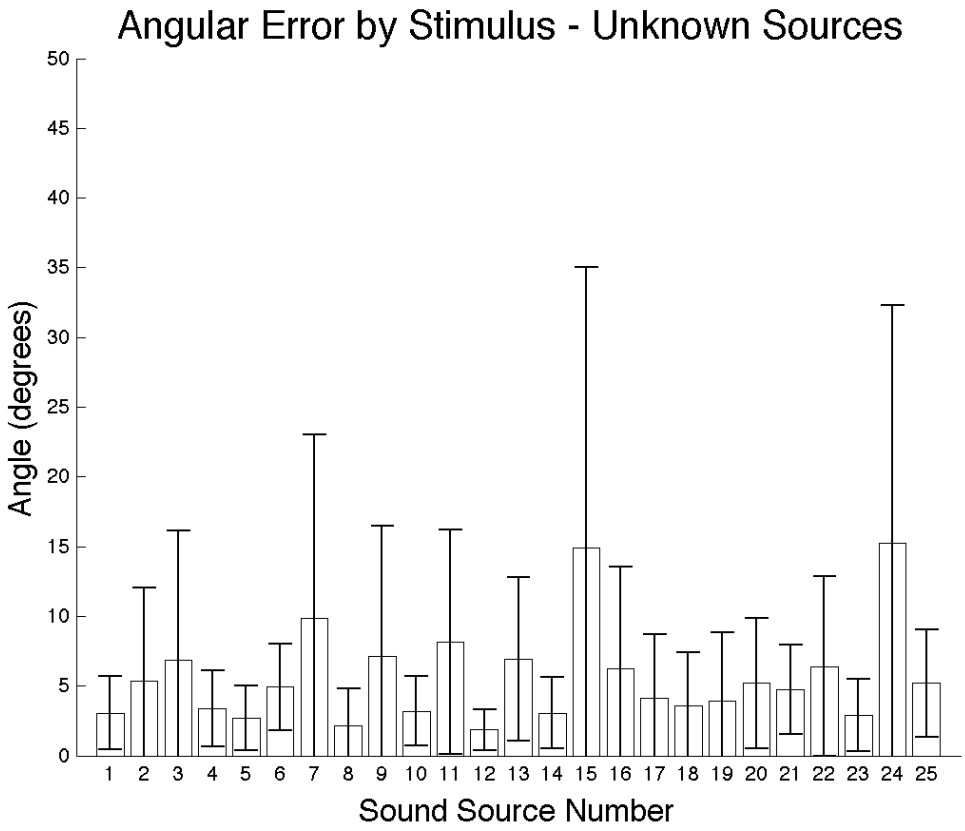


Figure 6.6: Angular accuracy by stimulus. Along the abscissa is the stimulus number (from Table 6.2.1.2) and along the ordinate is the angular error.

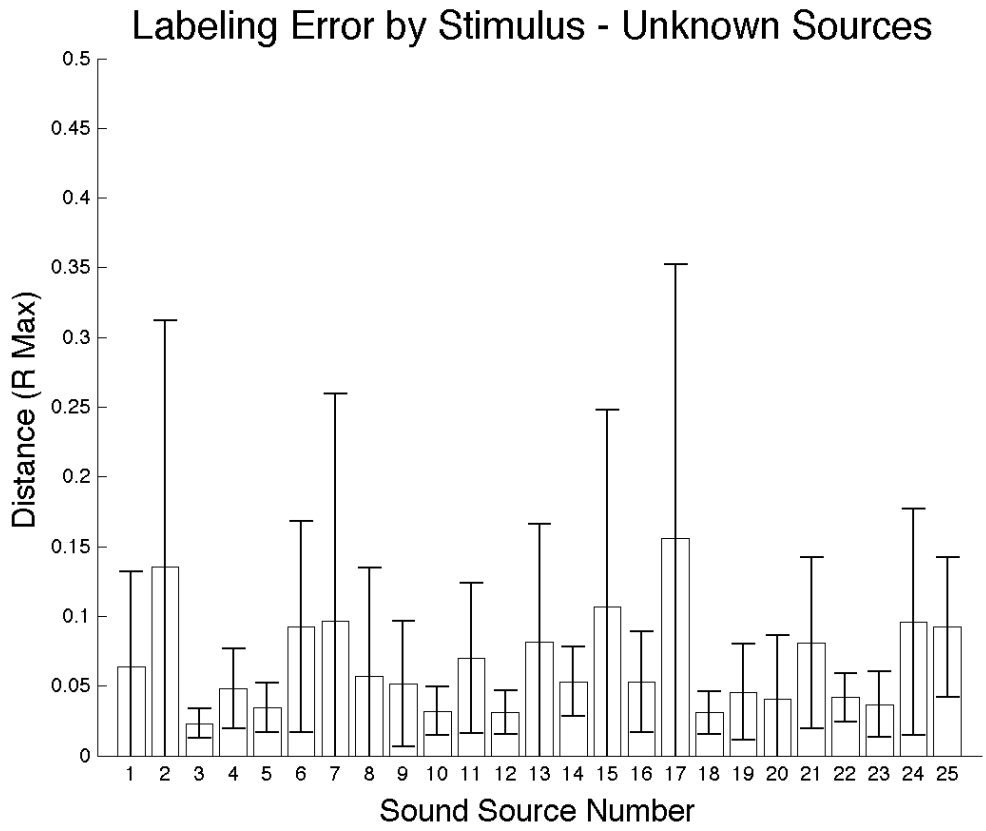


Figure 6.7: Labeling accuracy by stimulus. Along the abscissa is the stimulus number (from Table 6.2.1.2) and along the ordinate is the distance error.

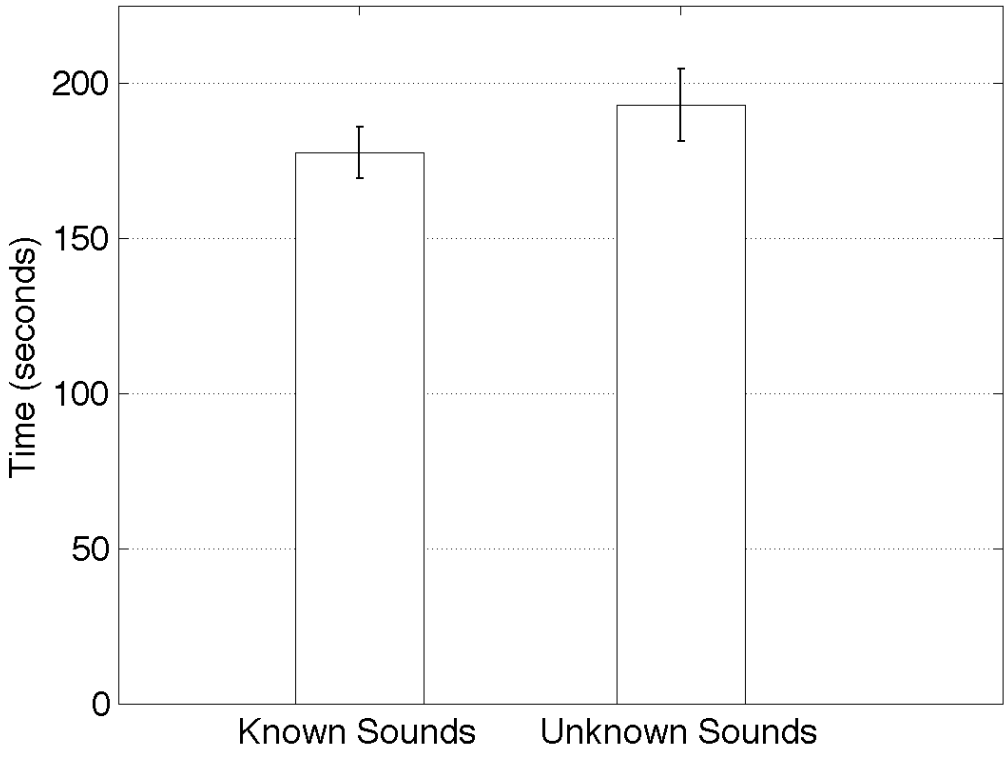


Figure 6.8: Effects of source certainty on exploration time. Along the abscissa is the source certainty and along the ordinate is the total time listeners explored the environment while marking the five sound sources.

suggests that operators of VAE systems will be able to detect unknown sounds that may be present in their listening environment, without experiencing an accuracy degradation.

However, although accuracy was not affected when unknown sources were used, listeners spent significantly more time exploring the environment to search for unknown sounds. This suggests that unknown sound sources take longer to find and/or remember. One possible explanation for the increase in time is that the listener has not had the advantage of rehearsing the identities of these sounds, and more time is needed to memorize the locations of the unknown sounds during exploration. Another possible explanation is that the listener needed extra time to accurately localize the unknown sound sources, as they had not been as thoroughly rehearsed as the known sources.

6.3 Effects of Visual Augmentation

In the experiments of the preceding chapter, listeners searched for and remembered spatial sounds, while using minimal visual cues. The only visual cues were shapes that indicated the listener's position and orientation, and markers that were placed by the listener. In a real-world system, it is likely that the operator will have positional visual cues along with the spatial audio cues during search and recall. Thus, it is necessary to examine how auditory and visual cues interact during the recall of auditory spatial objects. The present study seeks to determine the effects of using an advanced visual cue during exploration and recall.

It is possible that an enhanced visual augmentation, such as a coordinate system, would help the listener to recall the positions of the auditory objects. If so, system designers would be encouraged to consider incorporating a coordinate-based visual augmentation into the interface. On the other hand, the additional cognitive load of interacting with a visual reference frame while monitoring an auditory object may

degrade listener performance. In this case, a system designer of a highly visual interface may not choose to use spatial audio in that interface, since it degrades search recall and accuracy.

To test the effects of visual augmentation, listeners explored a five-source auditory environment that was augmented with a Cartesian or polar reference frame. Positioning accuracy, angular accuracy, labeling accuracy, and exploration time were measured and compared to the DP + DL condition of the experiment in Chapter V.

6.3.1 Methods

6.3.1.1 Participants, Stimuli and Apparatus

The five observers from the previous experiment participated in the current experiment. The experiment required about 3.8 hours of listening and was completed in a single session, apart from the previous experiment. The present study used the same real-time MATLAB-based spatial auditory system from Chapter IV and Chapter V. This experiment also used the same stimuli from Chapter IV

6.3.1.2 Procedure

Prior to beginning each experiment condition, participants completed training to become familiar with the auditory environment. The training procedure was identical to that which was outlined in Chapter V and augmented with a Cartesian or polar reference frame.

A balanced design experiment was conducted to investigate the influence of visual reference frames on auditory spatial memory within a VAE. Participants performed a modified version of the *Delayed Positioning then Delayed Labeling* (DP + DL) experimental condition that was outlined in the previous chapter. The participant walked around the auditory environment while they memorized the locations of the five spatialized sound sources. Next, the participant marked the locations of the

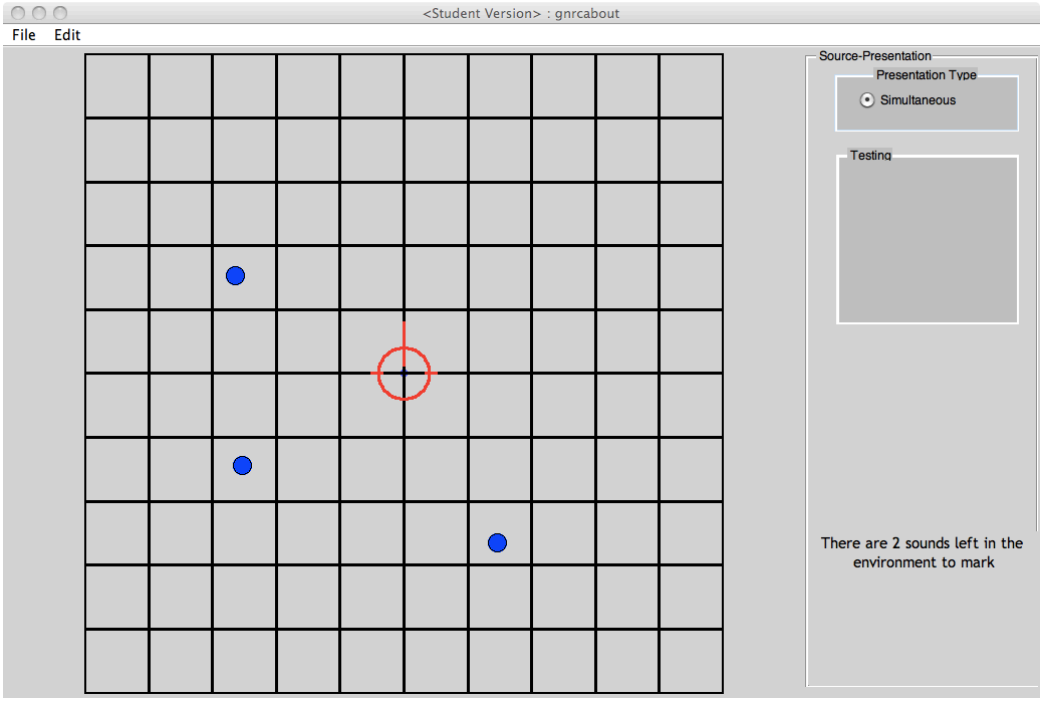


Figure 6.9: Cartesian reference frame experimental condition in which the participant has explored the environment and marked three of the five sound sources (blue circles).

sound sources. Finally, the participant labeled the identities of the sound sources that were marked in the previous step. Essentially, the participant performed the *DP + DL* condition while the screen was visually augmented with a Cartesian (Figure 6.9) or polar (Figure 6.10) reference frame during exploration and recall.

6.3.2 Results

Results of the current experiment were compared with the results of the DP+DL experiment from Chapter V, which represented an identical condition in which a reference frame was not used.

6.3.2.1 Accuracy

Effects were examined through a comparison of results of the previous experiment and the DP+DL experiment of Chapter V. Figure 6.11 shows the mean positioning

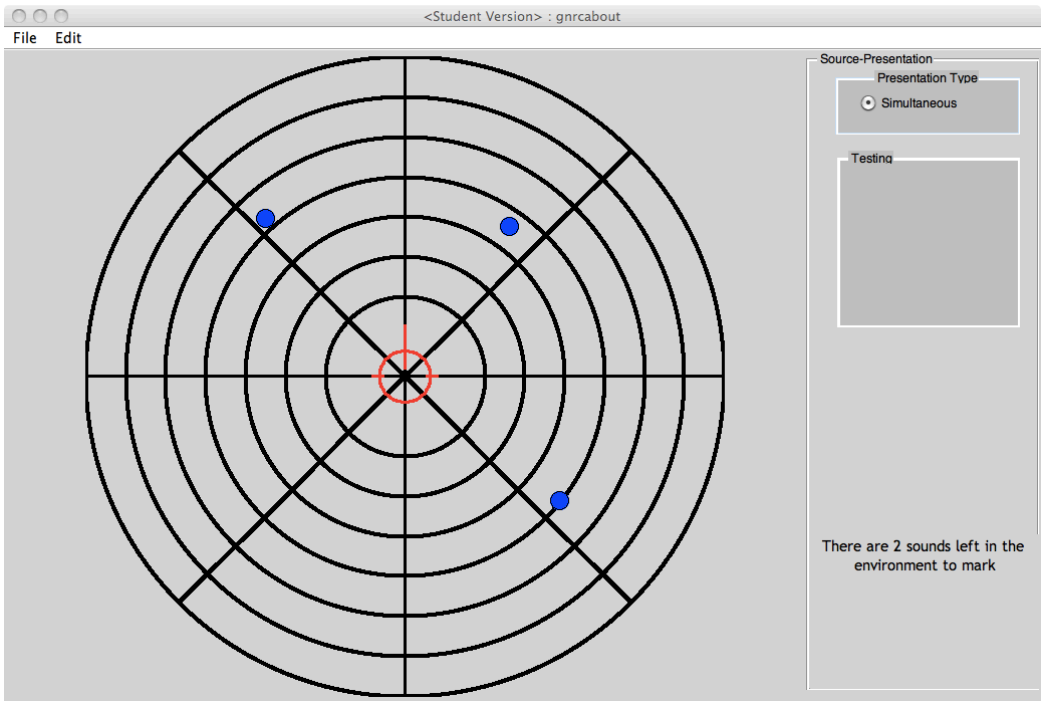


Figure 6.10: Polar reference frame experimental condition in which the participant has explored the environment and marked three of the five sound sources (in blue).

error during the three visual augmentation conditions. A significant difference in positioning error was observed as an effect of visual augmentation [$F_{2,1497}$=43.57, $p<0.05$]. A Tukey LSD multiple comparison test showed that positioning error was significantly lower with either visual augmentation.

Similar results were observed in the comparison of angular error. Figure 6.12 shows the mean positioning error during the three visual augmentation conditions. A significant difference in angular error was observed as an effect of visual augmentation [$F_{2,1497}$=2992.3, $p<0.05$]. A Tukey LSD multiple comparison test showed that angular error was significantly lower when either visual augmentation was used, as compared to no visual augmentation.

As in the previous two analyses, the labeling accuracy follows a similar trend. Figure 6.13 shows that there was a significant difference in labeling error as an effect of visual augmentation [$F_{2,1497}$=26.22, $p<0.05$]. A Tukey LSD multiple comparison

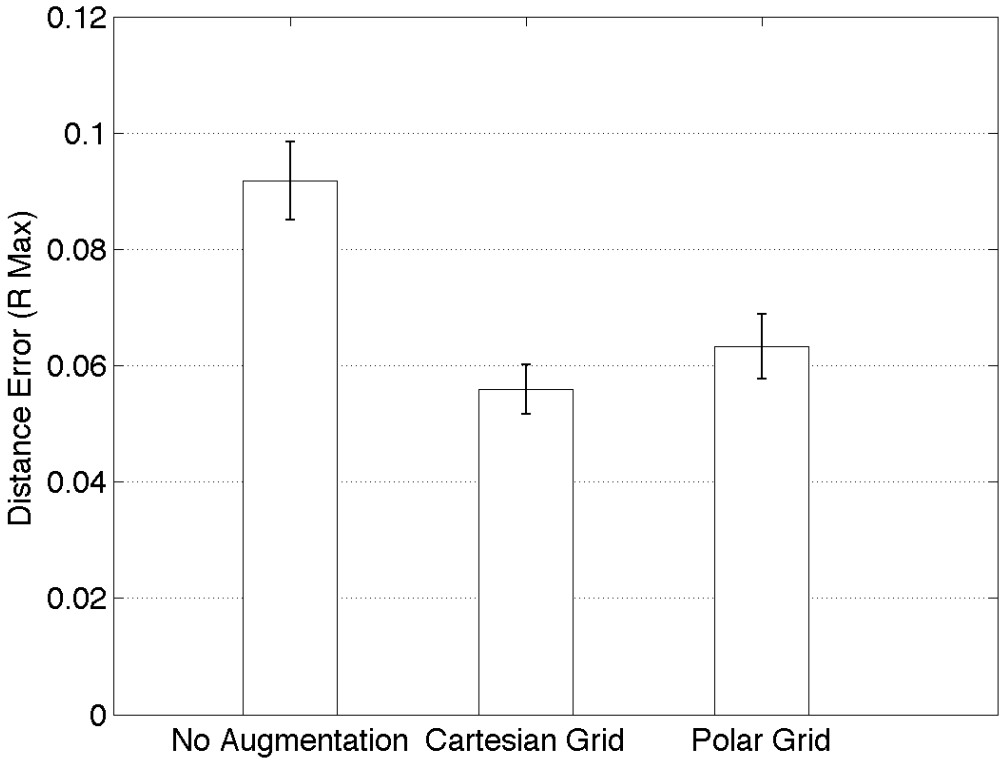


Figure 6.11: Effects of visual augmentation on positioning accuracy. Along the abscissa are the augmentation conditions and along the ordinate is the positioning error.

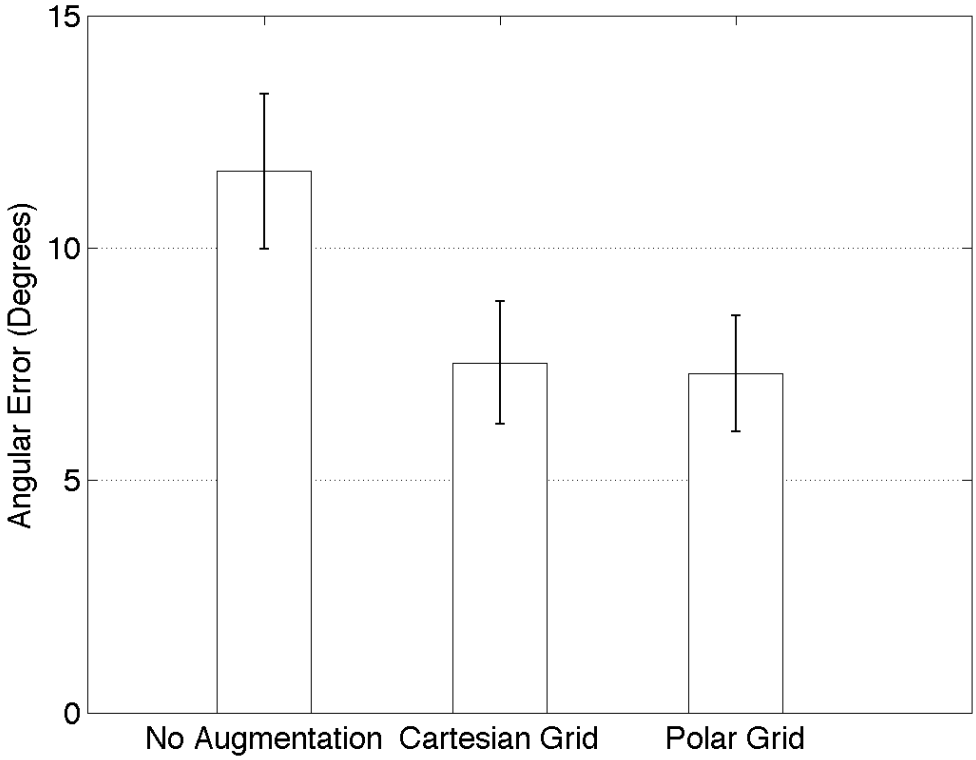


Figure 6.12: Effects of visual augmentation on angular accuracy. Along the abscissa are the augmentation conditions and along the ordinate is the angular error.

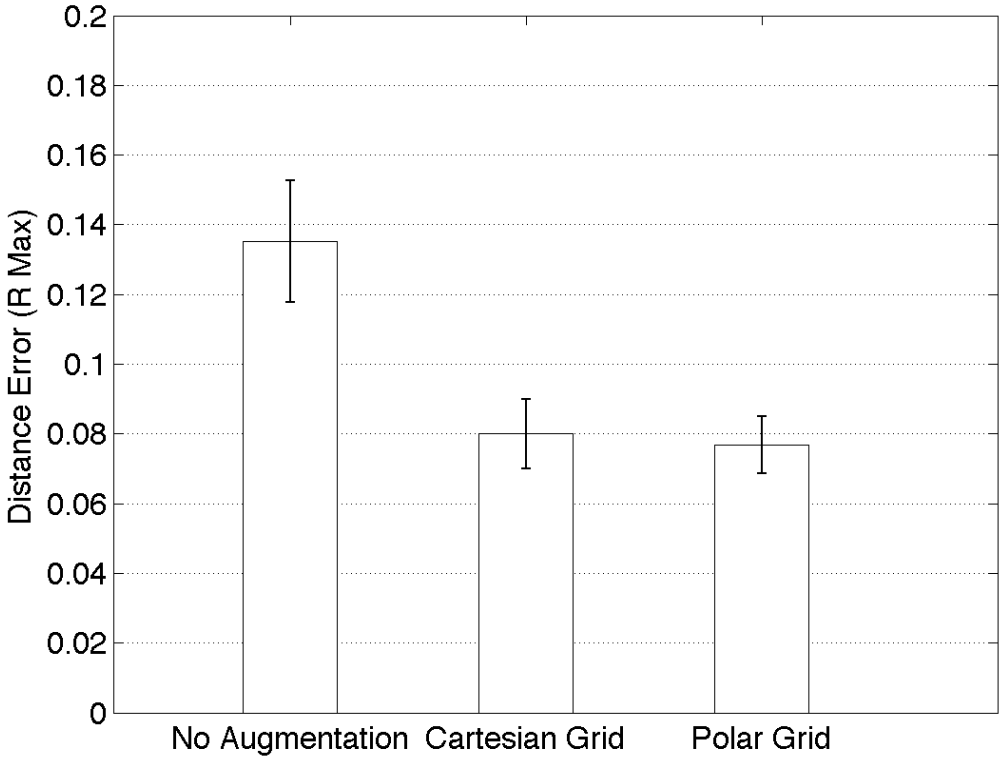


Figure 6.13: Effects of visual augmentation on labeling accuracy. Along the abscissa are the augmentation conditions and along the ordinate is the labeling error.

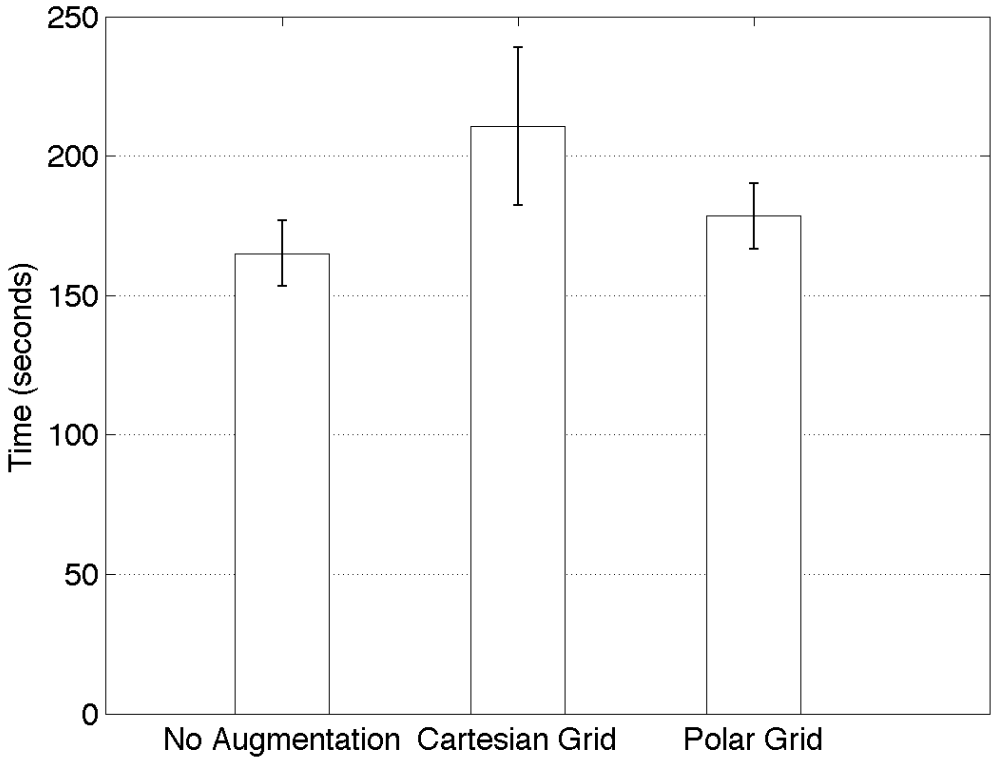


Figure 6.14: Effects of visual augmentation on exploration time. Along the abscissa are the augmentation conditions and along the ordinate is the total exploration time.

test showed that angular error was significantly lower when either visual augmentation was used, as compared to no visual augmentation.

6.3.2.2 Exploration Time

Figure 6.14 shows the average amount of time listeners explored the environment using a polar, Cartesian, or no reference frame. As in the previous analyses, there was a significant difference in exploration time as an effect of visual augmentation [$F_{2,297}$=6.02, $p<0.05$]. A Tukey LSD multiple comparison test showed that exploration time was higher in the Cartesian grid condition than the other two conditions.

The data was further broken down by subject (Figure 6.15) and an additional

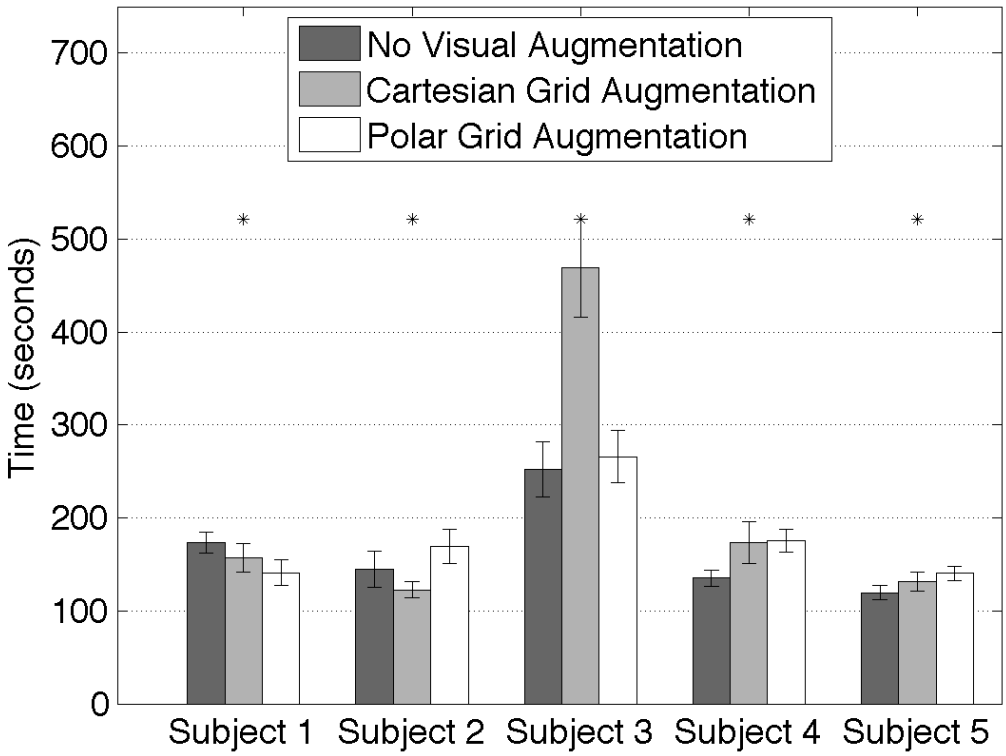


Figure 6.15: Exploration time as affected by visual augmentation. Along the abscissa are the subjects and along the ordinate is the total exploration time.

ANOVA was performed. Visual augmentation significantly affected each subject's exploration time, however there was no trend observed.

6.3.3 Discussion

Participants exhibited significantly higher positioning, angular and labeling accuracy when an auditory environment was augmented with a Cartesian or polar reference frame. No effect was seen across the two types of visual augmentation.

It is interesting to note that some researchers have found that visual objects are more easily remembered with a salient visual reference frame. For example, *Carlson* (2008) and *Kelly et al.* (2010) found that incorporating this type of salient cue enables a participant to form a more accurate mental spatial representation of the environment. *Tversky* (1993) states that in environments where spatial orientations are difficult to remember, heuristics, such as reference points, were used to anchor figures to locations, making them easier to remember. *Von Wright et al.* (1978) found that the use of external spatial frames of reference aided in the encoding of attributes of visual objects in young children. In the auditory domain, *Zahorik* (2002) notes that it is important to recognize the contributions of non-acoustical (visual) factors that affect the perception of auditory space. On the other hand *Albert et al.* (1999) indicates that visual reference frames have little, if any, impact on the acquisition of spatial relationships.

Participants exhibited significantly higher recall accuracy when any visual augmentation was used. There was no difference in accuracy between the two visual augmentations. One might have expected angular accuracy to be significantly higher in the polar augmentation condition, due to the structure of the reference frame. Perhaps reference frames are used as general-purpose landmarks. Their shape may not influence sound search and recall.

Interestingly, participants spent significantly more time exploring the Cartesian

augmented environment, followed by the polar augmented environment. The higher exploration time may have been a result of participants counting the boxes in an attempt to memorize the exact coordinate or square in which each source was located.

Findings from this experiment suggest that VAE designers should incorporate visual reference frames into their visual positional display of auditory spaces. However, they must note that an increase in search time may be observed if the system uses an augmentation that encourages the operator to memorize features of the reference frame. If the operator is performing a task that is not time-sensitive, any frame could be used.

6.4 Effects of Attenuation Models

The previous experiments were performed within a given radius on a free-field geometry. In the environment, every sound was detectable from every position, using inverse-squared attenuation. In a real-life environment where a system operator needed to monitor a large number of sound sources, this type of interface could potentially become noisy and crowded. This would most likely degrade sound search time and accuracy. Perhaps a more drastic attenuation model is needed for dense environments. It is critical to study the effects of a larger geometry.

A larger geometry could potentially aid sound search. The search path analysis in Chapter IV indicated that listeners primarily use attenuation cues to localize sounds. A drastic attenuation model makes the attenuation change more pronounced and (possibly) easier to detect; which may aid the listener's search. A drastic attenuation model could also lower the signal to noise ratio in dense environments, making search easier.

On the other hand, a drastic attenuation model could hinder spatial sound search. Since every source would not be detectable at any point, some sources could become hidden and/or require significantly more time to locate.

The present experiment examines how a drastic attenuation and an absent attenuation model affect listener accuracy and exploration time. In the drastic attenuation condition, we aim to explore the performance effects of a sharp attenuation model. In the non-attenuated condition, our goal is to determine if search and recall are degraded (or even possible) if all sounds are equally as loud.

To test the effects of attenuation models, listeners explored a five-source VAE that had no attenuation or a drastic (inverse-eighth attenuation). Positioning accuracy, angular accuracy, labeling accuracy, and exploration time were measured and compared to the DPDL condition of the experiment in Chapter V.

6.4.1 Methods

6.4.1.1 Participants, Stimuli and Apparatus

The five observers from the previous experiment participated in the current experiment. The experiment required about 4 hours of listening and was completed in a single session, apart from the previous experiment.. The present study used the same real-time MATLAB-based spatial auditory system and stimuli from Chapter IV and Chapter V.

6.4.1.2 Procedure

Before beginning each condition, the participants completed training to orient themselves to the auditory environment. The training procedure used was identical to that used in Chapter V. The user completed training using the same attenuation model as their assigned condition.

A balanced design experiment was conducted to investigate the influence of attenuation modeling (Figure 6.16) on auditory spatial memory within a VAE. The participant performed a modified version of the *Delayed Positioning and Delayed Labeling* (DPDL) experimental condition outlined in the previous chapter. The par-

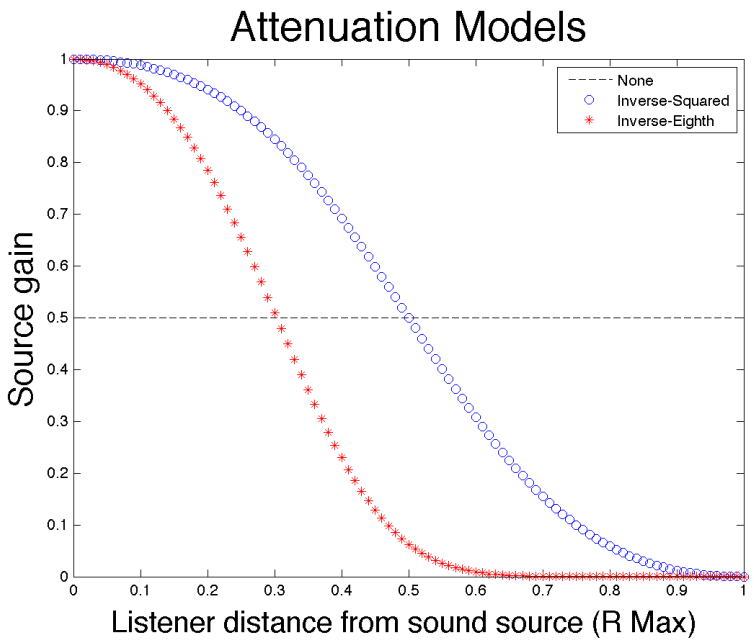


Figure 6.16: Attenuation Models used in each experimental condition.

ticipant walked around the auditory environment and memorized the locations of the five sound sources. After the participant indicated that they had learned the spatial configuration of the environment, they were aurally cued with a sound from the interface. After hearing the sound, the participant clicked the position in the interface where they remembered hearing the sound. Essentially, the participant performed the *Delayed Positioning and Delayed Labeling* (DPDL) experimental condition while the sound sources were non- or inverse-eighth attenuated.

6.4.2 Results

Results of the current experiment were compared with the results of the DPDL condition of the experiment in Chapter V, which used a standard inverse-squared attenuation model.

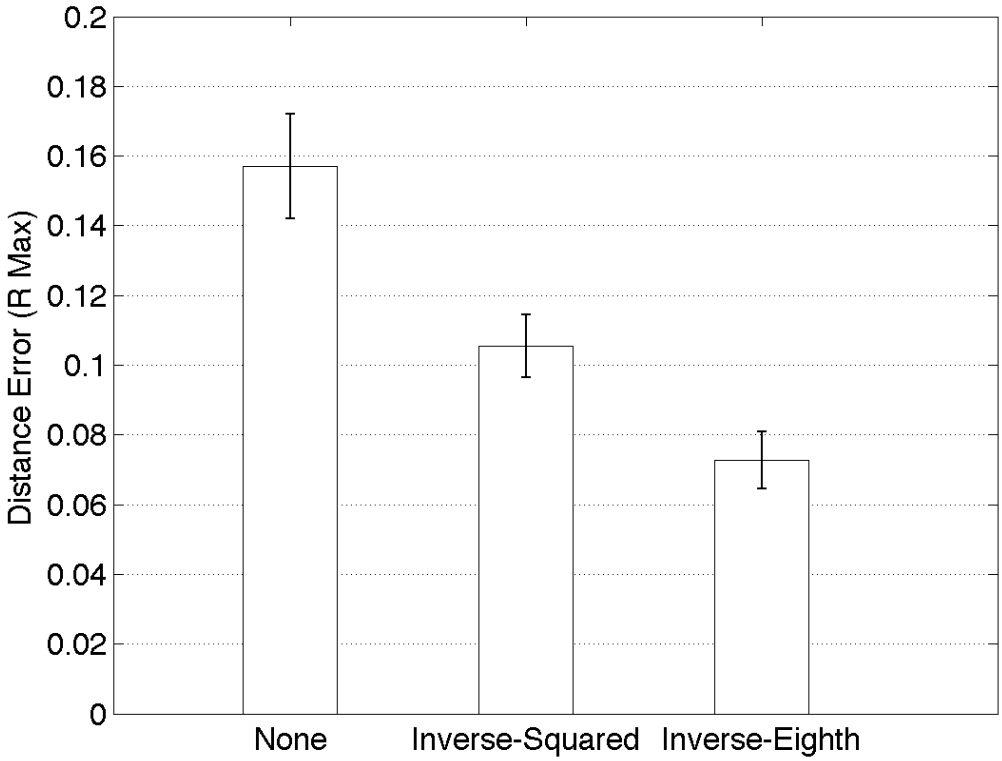


Figure 6.17: Effects of attenuation modeling on positioning accuracy. Along the abscissa are the attenuation models and along the ordinate is the positioning error.

6.4.2.1 Accuracy

The effects of attenuation model on positioning accuracy were examined through a comparison of error during the aforementioned condition and the error during the DPDL experiment from Chapter V.

Figure 6.17 shows that the choice of attenuation model has a significant effect on positioning error [$F_{2,1497}$=56.18, $p<0.05$]. A Tukey LSD multiple comparison test showed that positioning error was the lowest when an inverse-eighth attenuation model was used. Additionally, positioning error was lower using inverse-squared attenuation than no attenuation.

Figure 6.18 shows the effect of attenuation model on angular error. Similar to

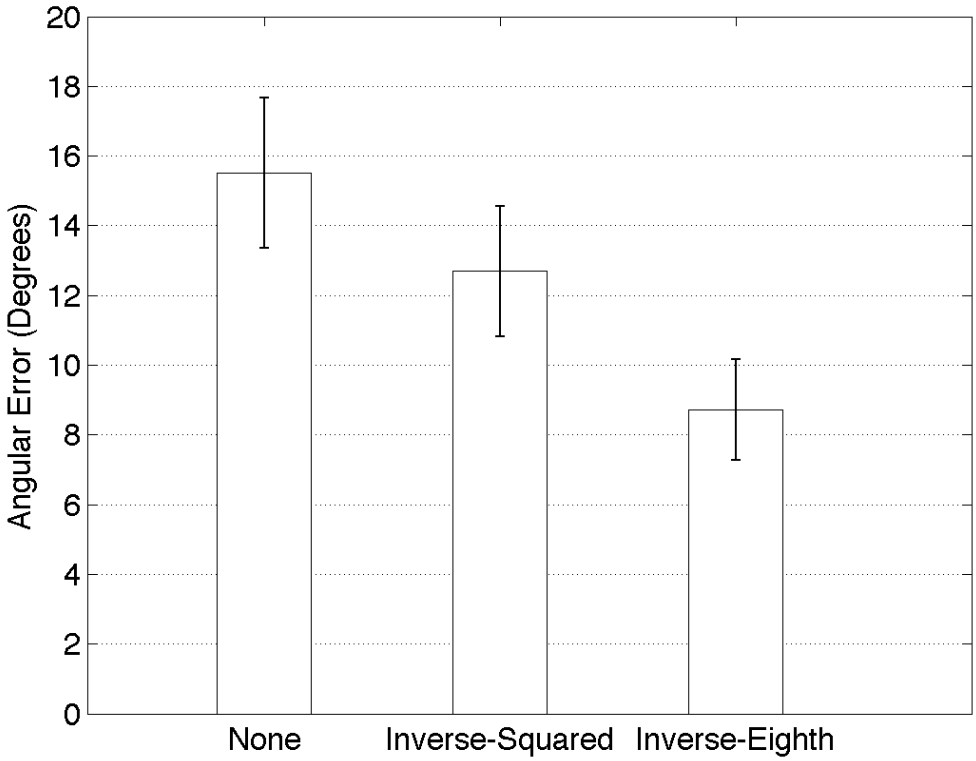


Figure 6.18: Effects of attenuation modeling on angular accuracy. Along the abscissa are the attenuation models and along the ordinate is the angular error.

positioning error results, a significant difference was observed between conditions [$F_{2,1497}$=13.12, $p<0.05$]. A Tukey LSD multiple comparison test showed that angular error was the lowest when an inverse-eighth attenuation model was used. Additionally, angular error was lower using inverse-squared attenuation than no attenuation.

Figure 6.19 shows the effect of attenuation model on labeling error. The analysis shows that attenuation model significantly affects labeling error [$F_{2,1497}$=70.40, $p<0.05$]. A Tukey LSD multiple comparison test showed that labeling error was the highest when no attenuation model was used. The data was further broken down by subject (Figure 6.20) and an additional ANOVA was performed. For two of the subjects, the attenuation model did not affect labeling error. Of the three affected subjects, only two showed significantly higher labeling error with no attenuation model.

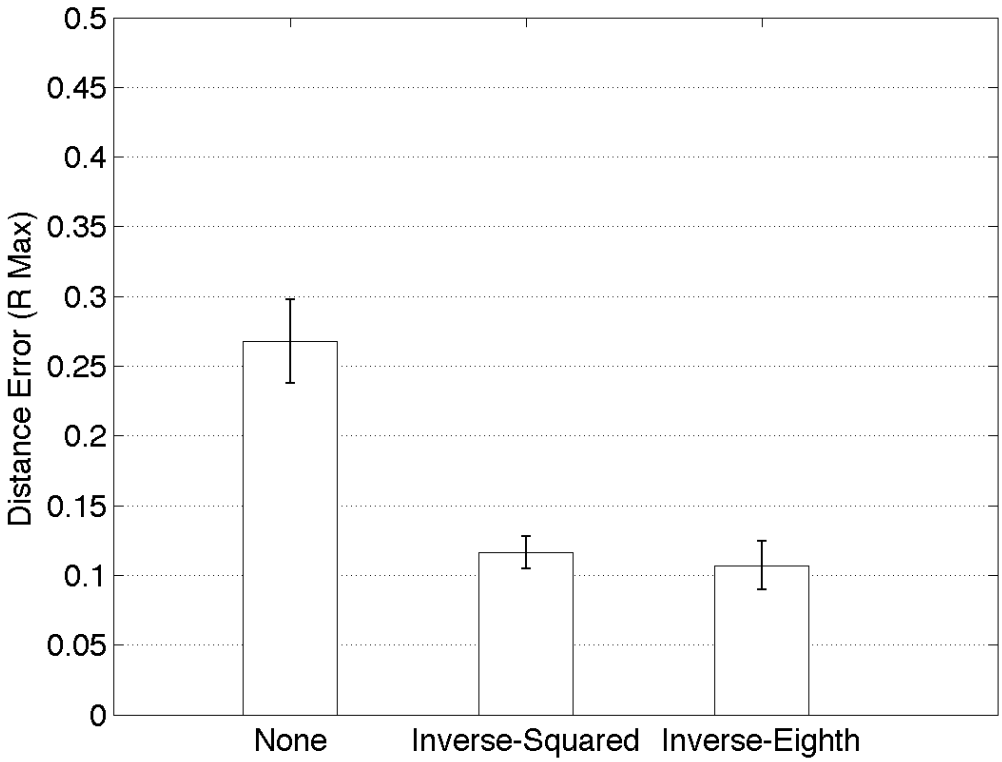


Figure 6.19: Effects of attenuation modeling on labeling accuracy. Along the abscissa are the attenuation models and along the ordinate is the labeling error.

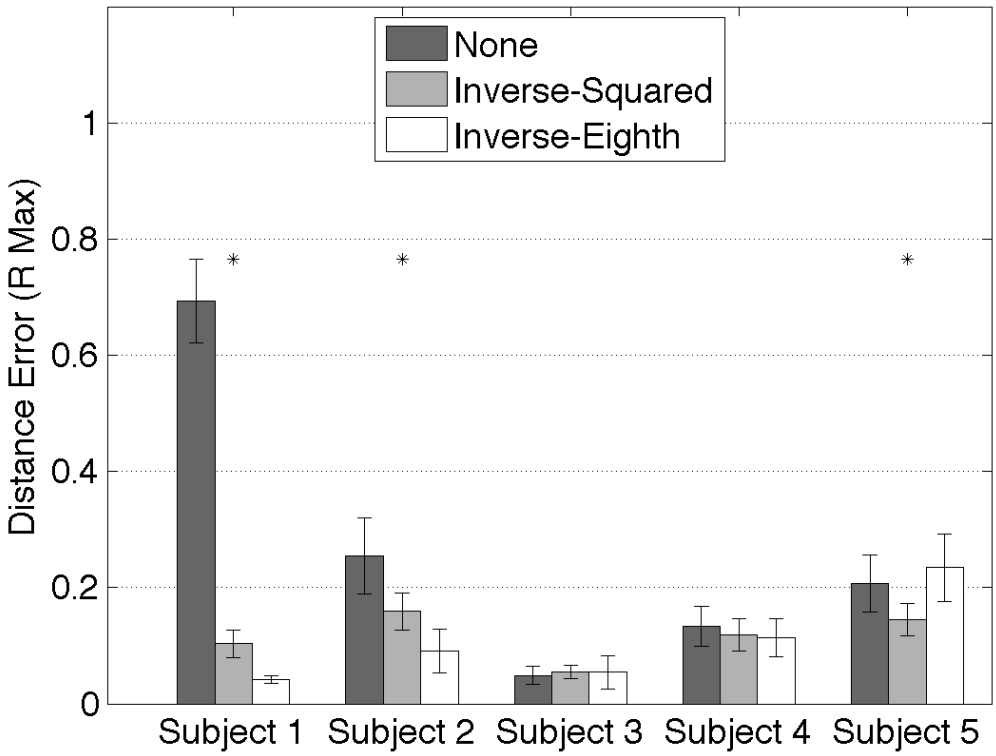


Figure 6.20: Effects of attenuation modeling on labeling accuracy, by subject. Along the abscissa are the subjects and along the ordinate is the labeling error.

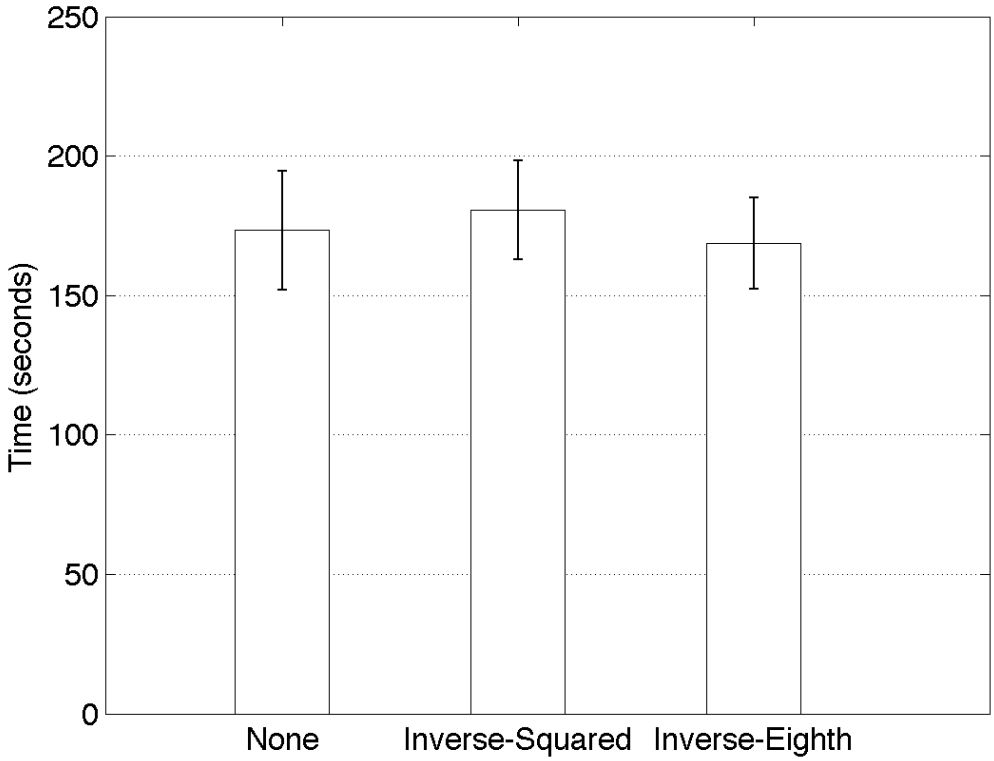


Figure 6.21: Effects of attenuation modeling on exploration time. Along the abscissa are the attenuation models and along the ordinate is the the total time listeners explored the environment while localizing the five sound sources.

6.4.2.2 Exploration Time

Figure 6.21 shows the effect of attenuation model on search time. The analysis shows that the choice of attenuation model does not affect search time [$F_{2,297}$=0.42, p=0.66].

6.4.3 Discussion

The results suggest that VAE designers should use more drastic attenuation models in VAEs to aid search and recall. The findings suggest that accuracy increases as attenuation increases; however, there may be a point at which an attenuation

increase is no longer beneficial. An extremely drastic attenuation model could hide some sounds for long periods of time, making them undetectable. Generally speaking, however, some attenuation modeling is necessary, because in the cases of no attenuation, performance was the worst. Additionally, a design tradeoff must be made since drastic attenuation may hinder sound source detection in applications where it is critical that objects are detected as soon as they enter the environment.

Using inverse-eighth attenuation resulted in higher positioning, angular, and labeling accuracy. This may have occurred because the environment was quieter during search and there was less interference from competing sources in the environment. Perhaps high attenuation narrows the search space of the possible locations of a sound source. With inverse-eighth attenuation, the sound source is only heard once the listener is within a very close proximity of the source. Generally speaking, in this condition, if the listener can hear the source, they must be fairly close to it. In the inverse-squared and flat attenuation conditions, sound sources were heard from more locations. This widened the search space of possible sound locations. The interface with no attenuation modeling had the lowest positioning, angular, and labeling accuracy. This was also the loudest interface.

The results also showed that listeners need similar amounts of exploration time for each attenuation model. This is particularly surprising, given that there were significant differences in accuracy across conditions. One would expect the amount of time needed to locate the sources to significantly differ as well. The results suggest that it may take just as long to find a drastically attenuated source as to narrow down the exact position of an easily detected sound source. Nevertheless, our findings indicate that VAE designers can increase the attenuation model of the system, which leads to more accurate search and recall, without affecting exploration time.

6.5 Summary

The experiments of the current chapter assessed the implications of three practical issues on the recall of VAE objects. It was discovered that when localizing unknown sounds, system operators could expect to experience no significant difference in search and recall accuracy as compared to known sounds. A significant increase in the amount of time needed to search the environment should be expected. The findings also recommend that VAE designers incorporate a coordinate system visual augmentation into positional displays. The data also suggests that VAE designers should use drastic attenuation models in the design of VAEs. Designers should be cautious when choosing the specific attenuation model, as the upper bound of drastic attenuation has not been assessed.

CHAPTER VII

Conclusion and Future Directions

7.1 Summary

This dissertation showed that listeners can train to locate and recall the positions of sound sources in controlled and practical situations within a VAE. In particular, the present work discovered many design considerations to aid creation of VAEs that augment spatial displays of information.

This dissertation has created initial footsteps to expand the understanding of how humans can navigate and use virtual spatial audio cues to learn an environment. The discoveries of this dissertation can be used to offload some visual perception tasks to the auditory sense to improve task efficiency and reduce operator error.

7.1.1 Contributions

The main contributions of this dissertation are:

1. A quick and inexpensive subjective-selection procedure to create reliable customized HRTFs for VAE listeners. It was found that listeners can judge the spatial qualities of sounds, even when using non-individualized HRTFs. The judgments were used to create customized HRTFs for the listeners that participated in the experiments of this dissertation, without incurring the cost

associated with creating individualized HRTFs. Our procedure created an alternative to direct measurement for those researchers and researchers who do not have the resources or access to an HRTF measurement facility.

2. A training procedure to teach listeners to locate virtual sounds in a spatial auditory environment. We were able to characterize the localization strategies used by listeners when searching for sources in a VAE. The strategies were used to create a training procedure which significantly helped listeners to localize spatial sound sources.

3. An assessment of auditory spatial memory. We found that listeners recall VAE configuration more accurately during free recall as compared to cued recall. We found that the length of time that listeners are required to remember the locations of sounds should be minimized.

4. We discovered that listeners can locate and identify unknown sound sources at the same accuracy as known sound sources. However, listeners needed significantly more time to encode the positions of the unknown sources. VAE designers should incorporate coordinate-based visual reference frames when representing positional information. Drastic attenuation models of sound improved search and recall within the auditory environment.

7.2 Extensions of the Present work

There are various ways in which to extend the present work:

- The HRTF selection procedure of Chapter III used stationary sound source judgments in order to select an HRTF for each listener. A possible extension could examine a listener's directional judgment of a moving sound source.

- Each listener in the HRTF selection procedure identified multiple HRTFs with good spatial qualities. Further analysis could include analyzing the HRTFs chosen at the end of the selection procedure to discover common characteristics among the chosen HRTFs. This would provide additional insight to identify specific auditory cues that listener deems important in spatial perception.

- In Chapter IV, listeners trained to locate five sound sources. An additional study to examine practice effect should also be performed. In the study, a group of listeners would perform the Walk and Mark procedure twice (without receiving training) to determine if listener performance changes after performing the initial task.

- In Chapter V, listeners recalled the locations and labels of five sound sources. To delve deeper into the capacity of audiospatial memory, the upper limits of concurrent audiospatial memory should be examined. It would be interesting to perform an additional study to assess listener recall performance of larger quantities of sound sources.

- The results of Chapter VII indicated that the use of reference frames or drastic attenuation modeling improved search and recall of spatial auditory objects. An interesting line of inquiry could assess listener performance in a VAE with visual augmentation and drastic attenuation modeling. An additional line of investigation is needed to assess the upper-limit of attenuation modeling. Specifically, further exploration should determine the point at which drastic attenuation impairs search and recall accuracy.

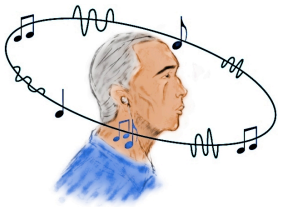
In all, the present work has raised and addressed many questions in using VAEs for search and recall of spatial information. Many audiences will benefit from the findings of this dissertation. Video game designers, psycho-acousticians, VAE designers and multi-modal interface designers may benefit from applying the results of this

research. This dissertation provides the groundwork to shape the design of VAEs, thus promoting richer interaction to aid the perception and surveillance of spatial information.

APPENDICES

APPENDIX A

Selected Experiment Scripts



College of Engineering, Dept. of EECS
University of Michigan
2260 Hayward St.
Ann Arbor, MI 48109-2121
Contact: Kyla McMullen, 734-763-0313,
kyla@umich.edu

Experiment Details

Hello my name is Kyla McMullen and today I will be conducting an experiment in spatialized audio. Spatialized audio involves processing a sound in order to give you, the listener, the perception that the sound is being emitted from a source located in 3-D space. This experience may already be familiar to you. For example, when you listen to music over headphones, you may have the sensation that the actual performance is being emitted from a spot inside of your head. We would like to research ways in which to extend this feeling so that a sound (delivered over headphones) may appear to be coming from a source in 3-D space that could be literally anywhere around you.

The sounds that you will experience today will be a lot simpler than those heard in a concert, or within a crowd. You will hear sounds in isolation. Please do not judge the "naturalness" or "realness" of the environment's sound. Please attend to the spatial qualities of the sound sources within the interface.

The experiment consists of three tasks:

1. Externalization

You will first hear the baseline signal (monophonic) followed by 5 discrete test signals randomly positioned. Select ALL of the intervals that appear to be externalized (located outside the head). If none can be discriminated, please check the 'none" checkbox. If you would like to replay an interval, you may click its corresponding letter in the interface. When you are done choosing the intervals, you will click "Submit', then 'Make Trial' to begin the next trial.

2. Elevation

You will hear 5 pairs of signals. Each pair contains signals presented at two different elevations (+30/-30). Select ALL of the intervals for which you perceive a difference in elevation. If none can be discriminated, please check the 'none" checkbox. If you would like to replay an interval, you may click its corresponding letter in the interface. When you are done choosing the intervals, you will click "Submit', then 'Make Trial' to begin the next trial.

3. Front / Back

You will hear 5 pairs of front/back back/front signals. Select ALL of the intervals for which you discriminate front from back. If none can be discriminated, please check the 'none" checkbox. If you would like to replay an interval, you may click its corresponding letter in the interface. When you are done choosing the intervals, you will click "Submit', then 'Make Trial' to begin the next trial.

Completing the steps from 1-3 indicates the termination of the experiment. Please alert the experimenter when each task has been completed. Please take a break between the experimental tasks if necessary.

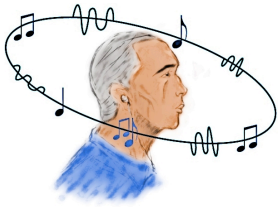
If you have any questions during the task(s), please remove the headphones and alert the experimenter. Please silence any electronics that may emit a noise during the experiment. You will be compensated with $10 (cash) per hour, rounded to the nearest 15 minutes. Do you have any questions?

We will not report any information that could potentially identify you. All of the data that you provide will be stored securely, then shredded and/or destroyed when it is no longer needed. This study has been approved by the IRB (Institutional Review Board). You are free to decide not to participate in this study or to withdraw at any time. If you choose to withdraw, you will be compensated for the time spent performing the study.

If you have questions about your rights as a research participant, or wish to obtain information, ask questions or discuss any concerns about this study with someone other than the researcher(s), please contact the University of Michigan Health Sciences and Behavioral Sciences Institutional Review Board, 540 E Liberty St., Ste 202, Ann Arbor, MI 48104-2210, (734) 936-0933 [or toll free, (866) 936-0933], irbhsbs@umich.edu"

Signature: _____

Date: _____



College of Engineering, Dept. of EECS
University of Michigan
2260 Hayward St.
Ann Arbor, MI 48109-2121

Contact: Kyla McMullen, 734-763-0313,
kyla@umich.edu

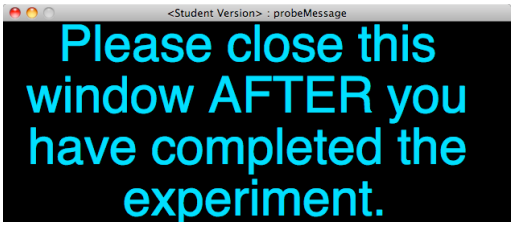
Assessing the Use of Auditory Interfaces to Aid Human Navigation in a Virtual Environment

Experiment Script

Hello my name is Kyla and today we will be conducting an experiment in spatial audio. You will be a part of a study that seeks to determine the effects of training when using an Auditory interface. In this experiment you will be placed within an auditory environment that consists of **5** sounds and you will move about the environment, marking the location of each sound source. Then you will train to use the environment. Afterwards, you will perform the first task an additional time.

First, you will begin by being placed in the center of an auditory environment. There will be five sound sources randomly scattered in the environment. It is your job to navigate through the environment, finding each sound source. The left and right arrow keys may be pressed to rotate your heading. When you are standing in the same location as the sound source, please press the spacebar. You are now free to move on and mark the locations of the other sound sources. Once all of the sources have been marked, you will see the actual locations of the sound sources in the environment (marked in green). Press Submit to save your work and then press New Environment to continue. You will repeat this procedure twenty times.

Next you will see a **message screen** that says:



Please only close this window AFTER the conclusion of each experiment. After the window is closed, the training task will begin. You will be asked to select an HRTF. Please open the folder that says "__Subject_Individualized_HRTFs" and select the file that says [your_Name].mat.

In the training task, you will be presented with a single sound source that will be located directly in front of, behind, to the left, or to the right of you. You must use the mouse to navigate to the sound source and mark its location using the spacebar. Training will continue until you have reached a predetermined criterion. You will save the data, close the **message screen,** open the folder that says "__Subject_Individualized_HRTFs" and select the file that says [your_Name].mat.

In the second portion of the training task, you will be situated in an auditory environment where the sound source to be located will be situated virtually anywhere around you. Similar to the first training task, you must use the mouse to navigate

to the sound source and mark its location using the spacebar. Training will continue until you have reached a predetermined criterion. You will save the data, close the **message screen,** open the folder that says "__Subject_Individualized_HRTFs" and select the file that says [your_Name].mat.

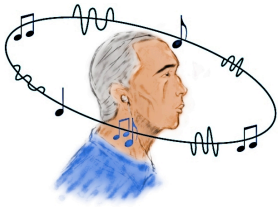
Finally, you will perform the first task an additional time.

Please complete these tasks as accurately as possible, while still being mindful of the amount of time spent answering. If you have any questions during the task(s), please remove the headphones and alert the researcher.

We will not report any information that could potentially identify you. All of the data that you provide will be stored securely, then shredded and/or destroyed when it is no longer needed. This study has been approved by the IRB (Institutional Review Board). You will be compensated hourly for your participation. You are free to decide not to participate in this study or to withdraw at any time. If you choose to withdraw, you will be compensated for the time spent performing the study. Do you have any questions?

Signature: _____

Date: _____



College of Engineering, Dept. of EECS
University of Michigan
2260 Hayward St.
Ann Arbor, MI 48109-2121

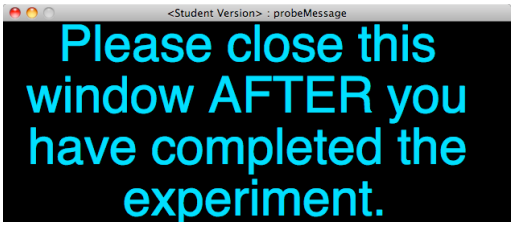
Contact: Kyla McMullen, 734-763-0313,
kyla@umich.edu

Assessing the Use of Auditory Interfaces to Aid Human Navigation in a Virtual Environment

Experiment Script

Hello my name is Kyla and today we will be conducting an experiment in spatial audio. You will be a part of a study that seeks to determine the effects of training when using an Auditory interface. In this experiment you will first complete auditory search training. Next you will be placed within an auditory environment that consists of **5** sounds and you will move about the environment, marking the location of each sound source. After marking the locations of each sound source, you will be asked to label the sound source.

First in the training task, you will be presented with a single sound source that will be located directly in front of, behind, to the left, or to the right of you. You must use the mouse to navigate to the sound source and mark its location using the spacebar. Training will continue until you have reached a predetermined criterion. You will save the data, close the **message screen,** open the folder that says "__Subject_Individualized_HRTFs" and select the file that says [your_Name].mat.



message screen

In the second portion of the training task, you will be situated in an auditory environment where the sound source to be located will be situated virtually anywhere around you. Similar to the first training task, you must use the mouse to navigate to the sound source and mark its location using the spacebar. Training will continue until you have reached a predetermined criterion. You will save the data, close the **message screen,** open the folder that says "__Subject_Individualized_HRTFs" and select the file that says [your_Name].mat.

Finally, you will perform the experimental task where you will walk around and mark each sound encountered. You will begin by being placed in the center of an auditory environment. There will be five sound sources randomly scattered in the environment. It is your job to navigate through the environment, finding each sound source. When you are standing in the same location as the sound source, please press the spacebar to mark it. You are now free to move on and mark the locations of the other sound sources. Once all of the sources have been marked, you will be presented with 5 drop down lists and

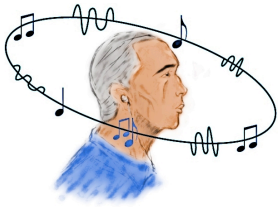
asked to mark the identity of each sound source (as represented by a number on each source marked within the interface). After you have labeled all of the sound sources, please press Submit to save your data. You will then be presented with the actual locations of each sound source (marked in green), labeled with the first letter of the sound source name within the interface. Please Press [New Environment] to continue. You will perform this procedure twenty times before concluding the experiment.

Please complete these tasks as accurately as possible, while still being mindful of the amount of time spent answering. If you have any questions during the task(s), please remove the headphones and alert the researcher.

We will not report any information that could potentially identify you. All of the data that you provide will be stored securely, then shredded and/or destroyed when it is no longer needed. This study has been approved by the IRB (Institutional Review Board). You will be compensated hourly for your participation. You are free to decide not to participate in this study or to withdraw at any time. If you choose to withdraw, you will be compensated for the time spent performing the study. Do you have any questions?

Signature: _____

Date: _____



College of Engineering, Dept. of EECS
University of Michigan
2260 Hayward St.
Ann Arbor, MI 48109-2121

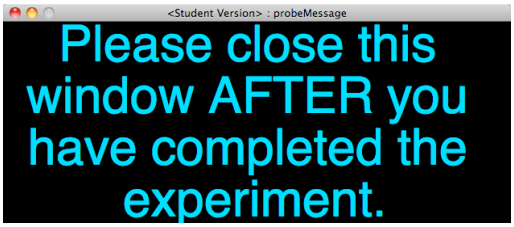
Contact: Kyla McMullen, 734-763-0313,
kyla@umich.edu

Assessing the Use of Auditory Interfaces to Aid Human Navigation in a Virtual Environment

Experiment Script

Hello my name is Kyla and today we will be conducting an experiment in spatial audio. You will be a part of a study that seeks to determine the effects of training when using an Auditory interface. In this experiment you will first complete auditory search training. Next you will be placed within an auditory environment that consists of **5** sounds and you will move about the environment, taking note of the location of each sound source. Once you are familiar with the environment, you will mark the locations of each source. After marking the locations of each sound source, you will be asked to label the sound sources.

First in the training task, you will be presented with a single sound source that will be located directly in front of, behind, to the left, or to the right of you. You must use the mouse to navigate to the sound source and mark its location using the spacebar. Training will continue until you have reached a predetermined criterion. You will save the data, close the **message screen,** open the folder that says "__Subject_Individualized_HRTFs" and select the file that says [your_Name].mat.



message screen

In the second portion of the training task, you will be situated in an auditory environment where the sound source to be located will be situated virtually anywhere around you. Similar to the first training task, you must use the mouse to navigate to the sound source and mark its location using the spacebar. Training will continue until you have reached a predetermined criterion. You will save the data, close the **message screen,** open the folder that says "__Subject_Individualized_HRTFs" and select the file that says [your_Name].mat.

Finally, you will perform the experimental task where you will walk around and take note of each sound encountered. You will begin by being placed in the center of an auditory environment. There will be five sound sources randomly scattered in the environment. It is your job to navigate through the environment, finding each sound source. When you have learned the locations of each of the sounds, you will press the [RETURN] key. After pressing return, you may mark locations of the

IRB IRB Number: HUM00026479 **Document Approved On: 01/06/2010**

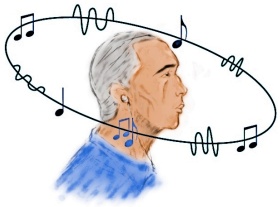
sounds in any order you'd like. Once all of the sources have been marked, you will be presented with 5 drop down lists and asked to mark the identity of each sound source (as represented by a number on each source marked within the interface). After you have labeled all of the sound sources, please press Submit to save your data. You will then be presented with the actual locations of each sound source (marked in green), labeled with the first letter of the sound source name within the interface. Please Press [New Environment] to continue. You will perform this procedure twenty times before concluding the experiment.

Please complete these tasks as accurately as possible, while still being mindful of the amount of time spent answering. If you have any questions during the task(s), please remove the headphones and alert the researcher.

We will not report any information that could potentially identify you. All of the data that you provide will be stored securely, then shredded and/or destroyed when it is no longer needed. This study has been approved by the IRB (Institutional Review Board). You will be compensated hourly for your participation. You are free to decide not to participate in this study or to withdraw at any time. If you choose to withdraw, you will be compensated for the time spent performing the study. Do you have any questions?

Signature: _____

Date: _____



College of Engineering, Dept. of EECS
University of Michigan
2260 Hayward St.
Ann Arbor, MI 48109-2121

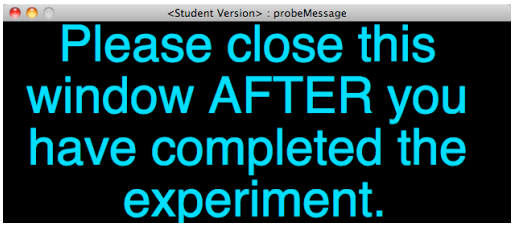
Contact: Kyla McMullen, 734-763-0313,
kyla@umich.edu

Assessing the Use of Auditory Interfaces to Aid Human Navigation in a Virtual Environment

Experiment Script

Hello my name is Kyla and today we will be conducting an experiment in spatial audio. You will be a part of a study that seeks to determine the effects of training when using an Auditory interface. In this experiment you will first complete auditory search training. Next you will be placed within an auditory environment that consists of **5** sounds and you will move about the environment, taking note of the location of each sound source. Once you are familiar with the environment, you will mark the locations of each source. After marking the locations of each sound source, you will be asked to label the sound sources.

First in the training task, you will be presented with a single sound source that will be located directly in front of, behind, to the left, or to the right of you. You must use the mouse to navigate to the sound source and mark its location using the spacebar. Training will continue until you have reached a predetermined criterion. You will save the data, close the **message screen,** open the folder that says "__Subject_Individualized_HRTFs" and select the file that says [your_Name].mat.



message screen

In the second portion of the training task, you will be situated in an auditory environment where the sound source to be located will be situated virtually anywhere around you. Similar to the first training task, You must use the mouse to navigate to the sound source and mark its location using the spacebar. Training will continue until you have reached a predetermined criterion. You will save the data, close the **message screen,** open the folder that says "__Subject_Individualized_HRTFs" and select the file that says [your_Name].mat.

Finally, you will perform the experimental task where you will walk around and take note of each sound encountered. You will begin by being placed in the center of an auditory environment. There will be five sound sources randomly scattered in the environment. It is your job to navigate through the environment, finding each sound source. When you have learned the locations of each of the sounds, you will press the [RETURN] key. After pressing return, you may mark locations of the

IRB IRB Number: HUM00026479 **Document Approved On: 01/06/2010**

sounds in any order you'd like. Once all of the sources have been marked, you will be presented with 5 drop down lists and asked to mark the identity of each sound source (as represented by a number on each source marked within the interface). After you have labeled all of the sound sources, please press Submit to save your data. You will then be presented with the actual locations of each sound source (marked in green), labeled with the first letter of the sound source name within the interface. Please Press [New Environment] to continue. You will perform this procedure twenty times before concluding the experiment.

Please complete these tasks as accurately as possible, while still being mindful of the amount of time spent answering. If you have any questions during the task(s), please remove the headphones and alert the researcher.

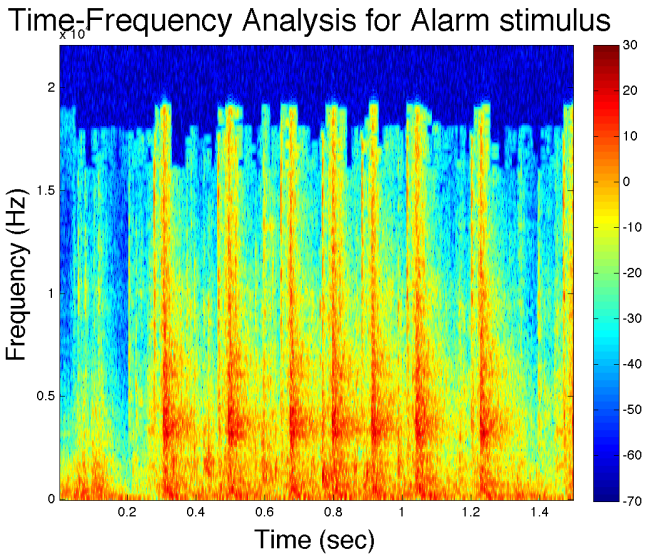
We will not report any information that could potentially identify you. All of the data that you provide will be stored securely, then shredded and/or destroyed when it is no longer needed. This study has been approved by the IRB (Institutional Review Board). You will be compensated hourly for your participation. You are free to decide not to participate in this study or to withdraw at any time. If you choose to withdraw, you will be compensated for the time spent performing the study. Do you have any questions?

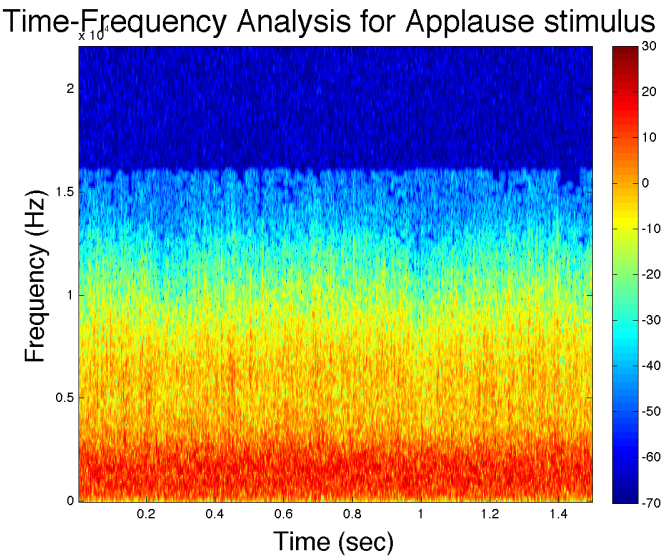
Signature: _____

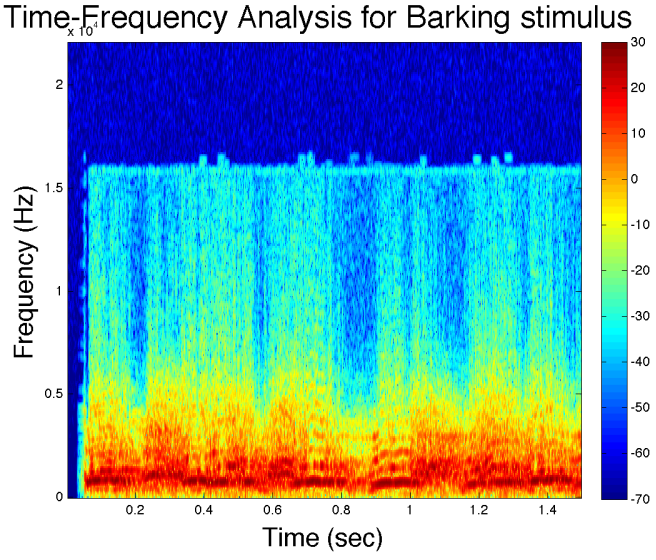
Date: _____

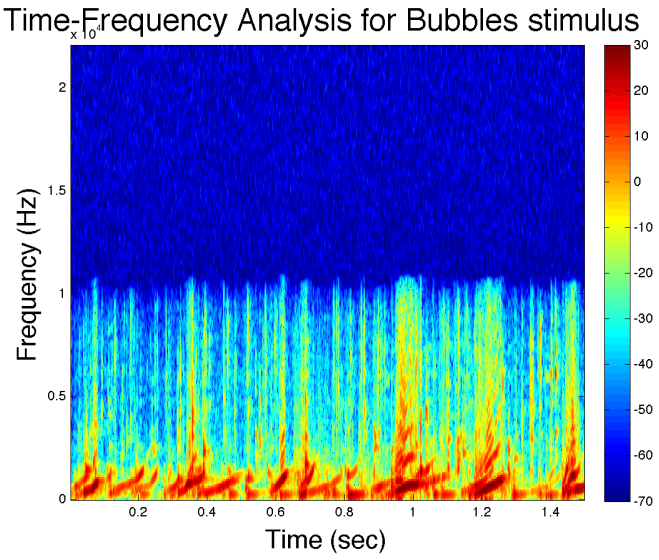
APPENDIX B

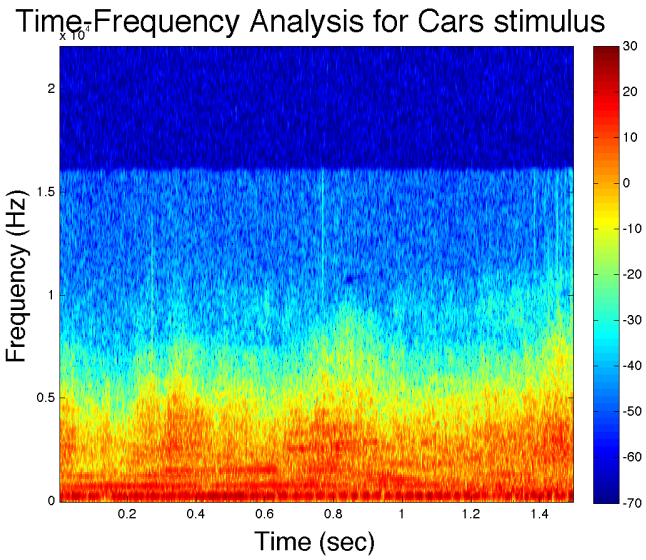
Experimental Stimuli Time-Frequency Analysis

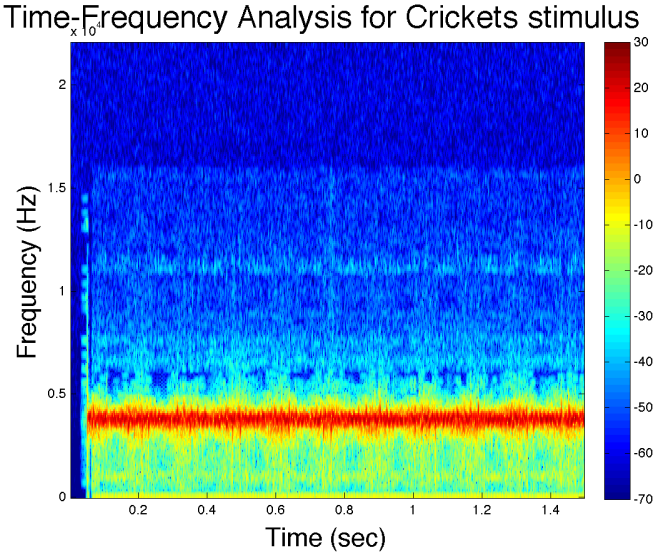

Time-Frequency Analysis for Alarm stimulus
x 10^4
Frequency (Hz)
Time (sec)
0
0.5
1
1.5
2
0.2
0.4
0.6
0.8
1
1.2
1.4
30
20
10
0
-10
-20
-30
-40
-50
-60
-70

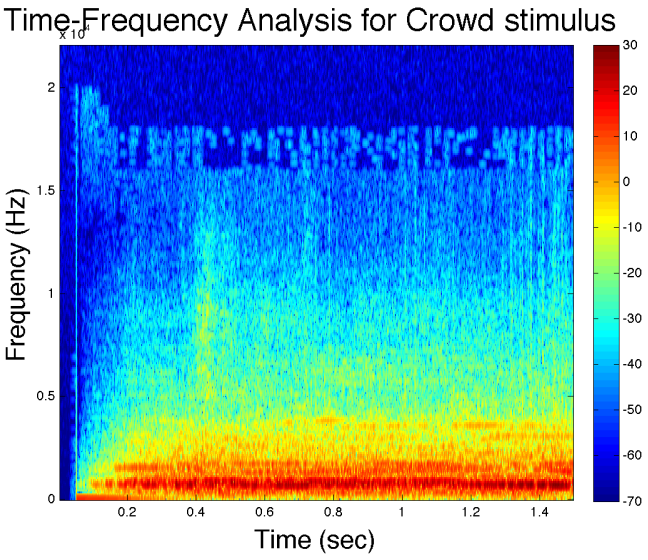

Time-Frequency Analysis for Applause stimulus
x 10^4
Frequency (Hz)
Time (sec)
0
0.5
1
1.5
2
0.2
0.4
0.6
0.8
1
1.2
1.4
30
20
10
0
-10
-20
-30
-40
-50
-60
-70

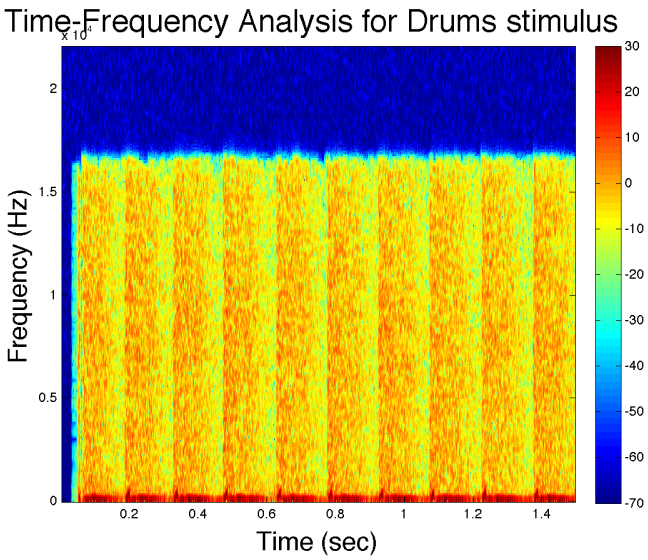

Time-Frequency Analysis for Barking stimulus
x 10^4
Frequency (Hz)
Time (sec)
0
0.5
1
1.5
2
0.2
0.4
0.6
0.8
1
1.2
1.4
30
20
10
0
-10
-20
-30
-40
-50
-60
-70

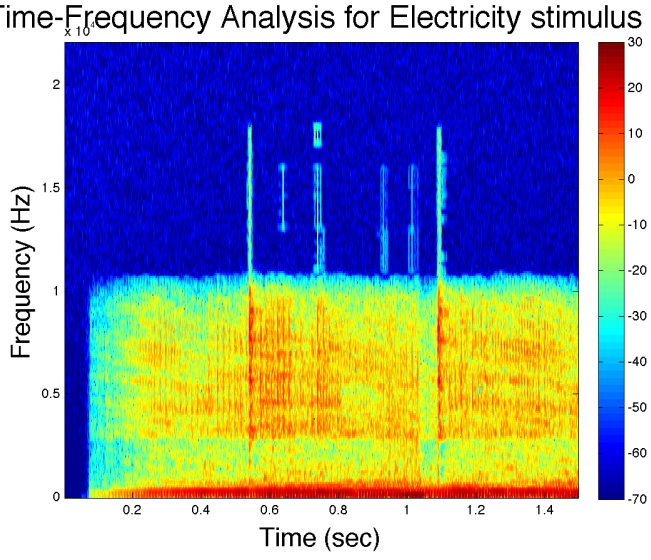

Time-Frequency Analysis for Bubbles stimulus
x 10^4
Frequency (Hz)
Time (sec)
0
0.5
1
1.5
2
0.2
0.4
0.6
0.8
1
1.2
1.4
30
20
10
0
-10
-20
-30
-40
-50
-60
-70

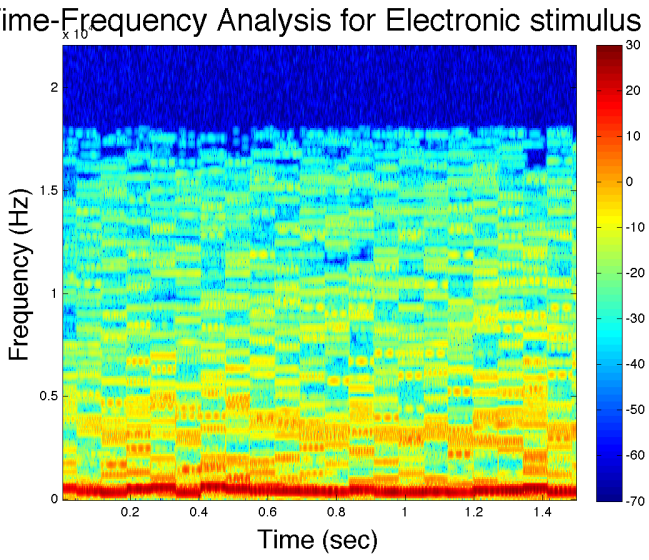

Time-Frequency Analysis for Cars stimulus
x 10^4
Frequency (Hz)
Time (sec)
0
0.5
1
1.5
2
0.2
0.4
0.6
0.8
1
1.2
1.4
30
20
10
0
-10
-20
-30
-40
-50
-60
-70

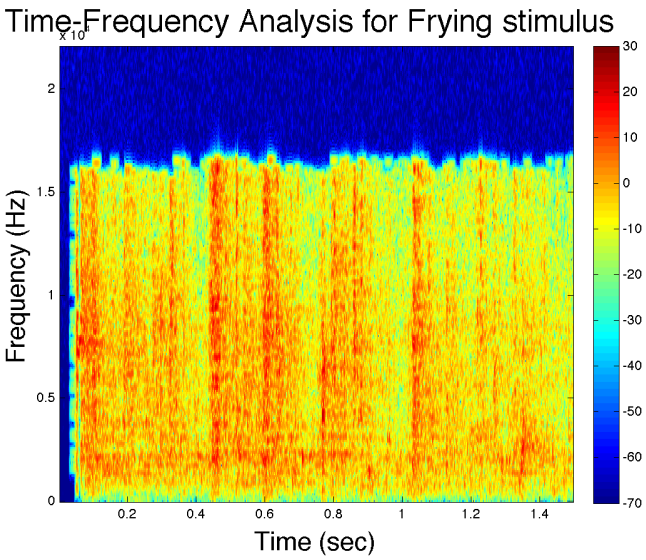

Time-Frequency Analysis for Crickets stimulus
x 10^4
Frequency (Hz)
Time (sec)
0
0.5
1
1.5
2
0.2
0.4
0.6
0.8
1
1.2
1.4
30
20
10
0
-10
-20
-30
-40
-50
-60
-70

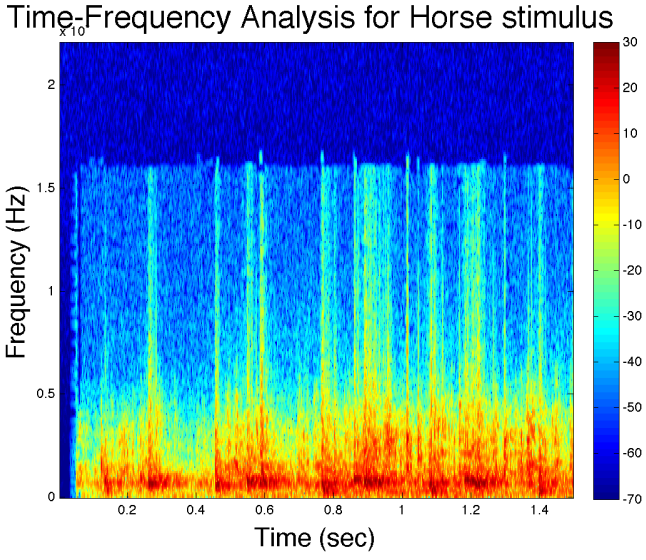

Time-Frequency Analysis for Crowd stimulus
x 10^4
Frequency (Hz)
Time (sec)
0
0.5
1
1.5
2
0.2
0.4
0.6
0.8
1
1.2
1.4
30
20
10
0
-10
-20
-30
-40
-50
-60
-70

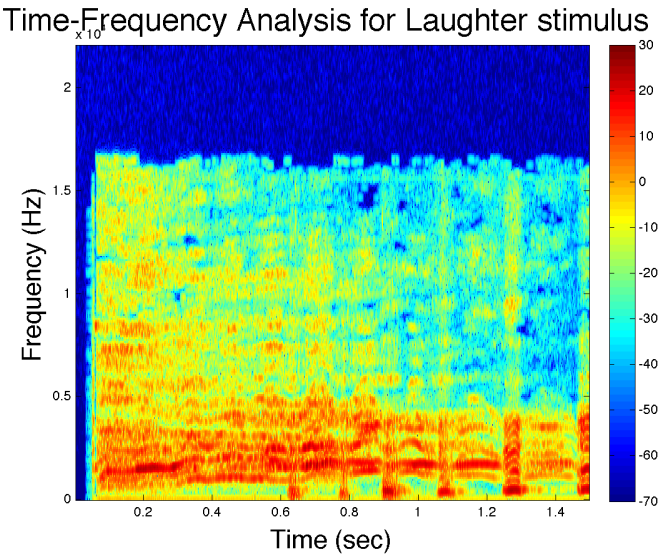

Time-Frequency Analysis for Drums stimulus
x 10^4
Frequency (Hz)
Time (sec)
0
0.5
1
1.5
2
0.2
0.4
0.6
0.8
1
1.2
1.4
30
20
10
0
-10
-20
-30
-40
-50
-60
-70

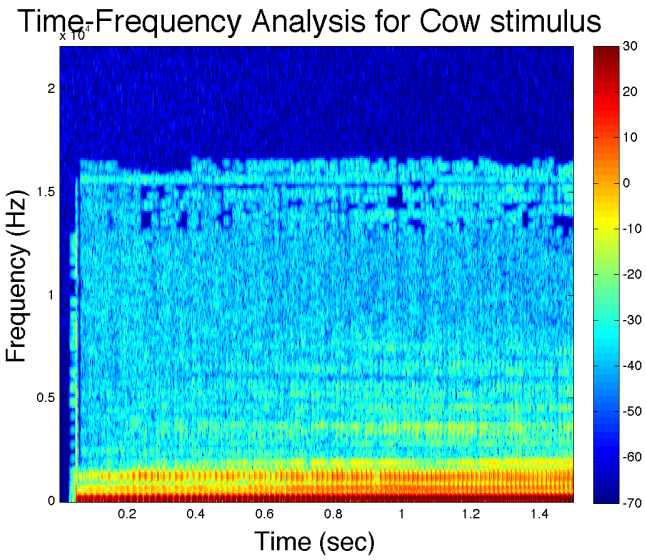

Time-Frequency Analysis for Electricity stimulus
x 10^4
Frequency (Hz)
Time (sec)
0
0.5
1
1.5
2
0.2
0.4
0.6
0.8
1
1.2
1.4
30
20
10
0
-10
-20
-30
-40
-50
-60
-70

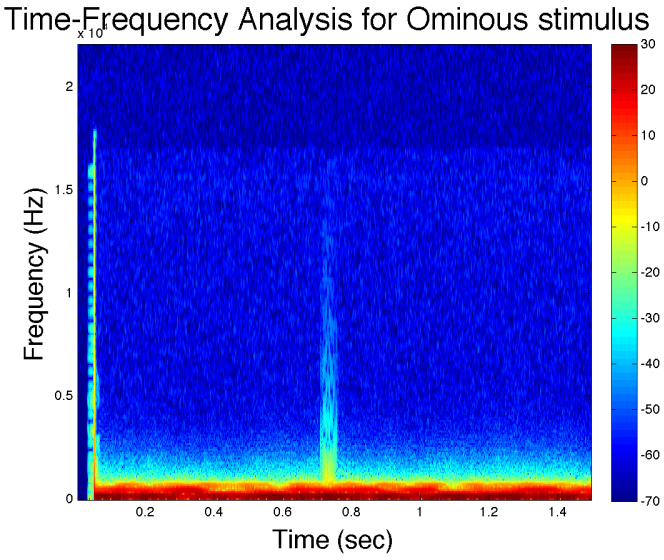

Time-Frequency Analysis for Electronic stimulus
x 10^4
Frequency (Hz)
Time (sec)
0
0.5
1
1.5
2
0.2
0.4
0.6
0.8
1
1.2
1.4
30
20
10
0
-10
-20
-30
-40
-50
-60
-70

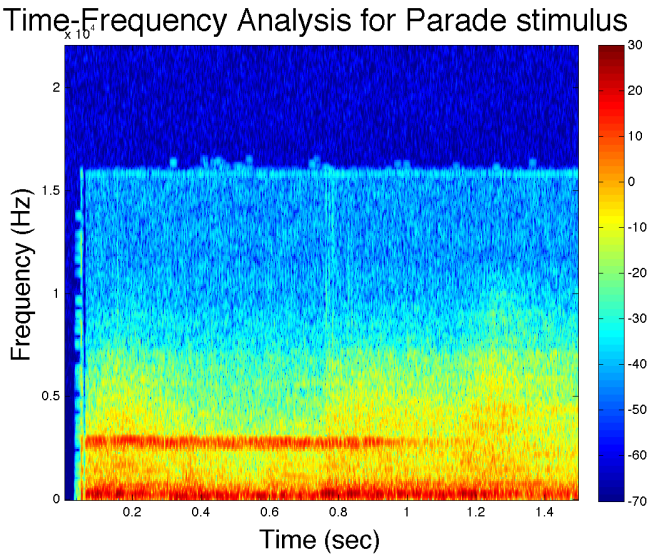

Time-Frequency Analysis for Frying stimulus
x 10^4
Frequency (Hz)
Time (sec)
0
0.5
1
1.5
2
0.2
0.4
0.6
0.8
1
1.2
1.4
30
20
10
0
-10
-20
-30
-40
-50
-60
-70

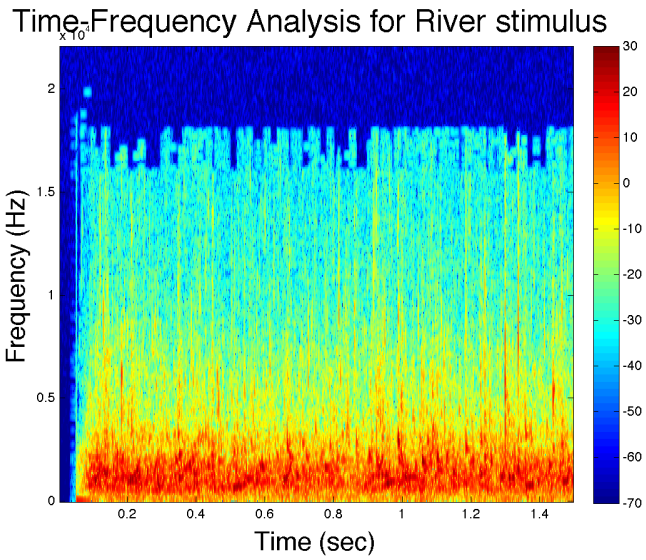

Time-Frequency Analysis for Horse stimulus
x 10^4
Frequency (Hz)
Time (sec)
0
0.5
1
1.5
2
0.2
0.4
0.6
0.8
1
1.2
1.4
30
20
10
0
-10
-20
-30
-40
-50
-60
-70

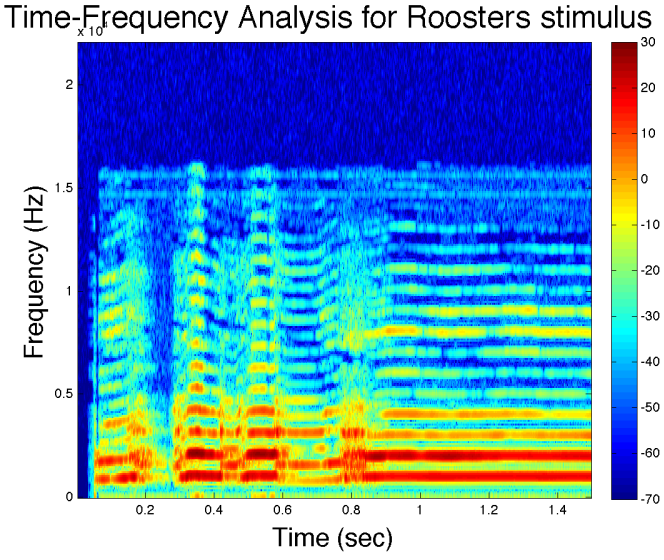

Time-Frequency Analysis for Laughter stimulus
x 10^4
Frequency (Hz)
Time (sec)
0
0.5
1
1.5
2
0.2
0.4
0.6
0.8
1
1.2
1.4
30
20
10
0
-10
-20
-30
-40
-50
-60
-70

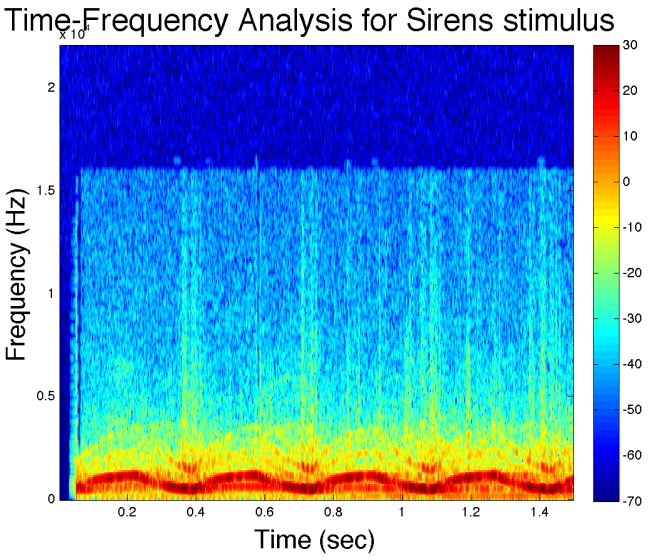

Time-Frequency Analysis for Cow stimulus
x 10^4
Frequency (Hz)
Time (sec)
0
0.5
1
1.5
2
0.2
0.4
0.6
0.8
1
1.2
1.4
30
20
10
0
-10
-20
-30
-40
-50
-60
-70

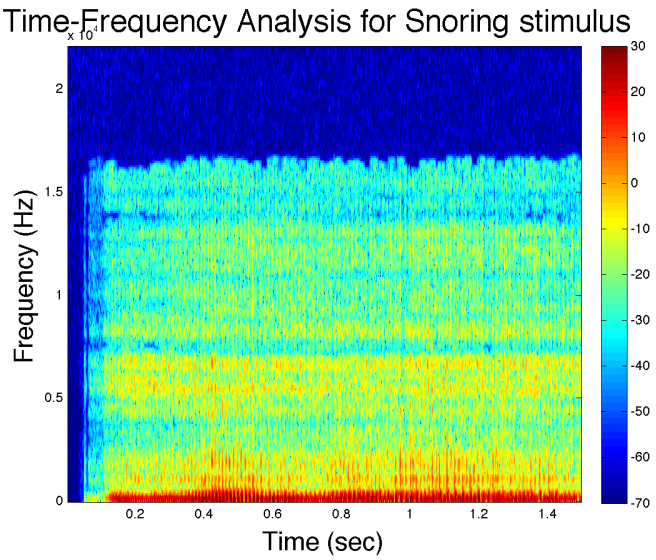

Time-Frequency Analysis for Ominous stimulus
x 10^4
Frequency (Hz)
Time (sec)
0
0.5
1
1.5
2
0.2
0.4
0.6
0.8
1
1.2
1.4
30
20
10
0
-10
-20
-30
-40
-50
-60
-70

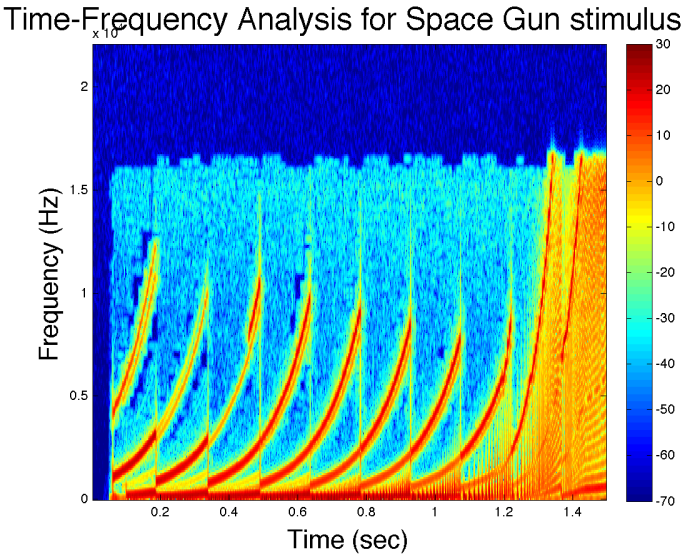

Time-Frequency Analysis for Parade stimulus
x 10^4
Frequency (Hz)
Time (sec)
0
0.5
1
1.5
2
0.2
0.4
0.6
0.8
1
1.2
1.4
30
20
10
0
-10
-20
-30
-40
-50
-60
-70

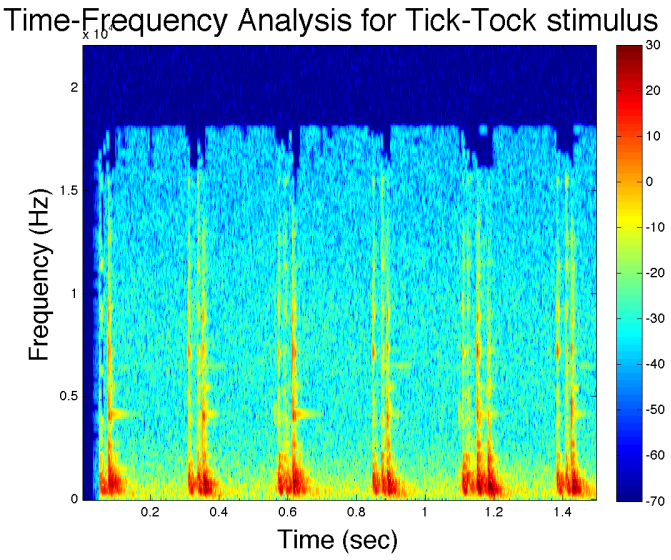

Time-Frequency Analysis for River stimulus
x 10^4
Frequency (Hz)
Time (sec)
0
0.5
1
1.5
2
0.2
0.4
0.6
0.8
1
1.2
1.4
30
20
10
0
-10
-20
-30
-40
-50
-60
-70

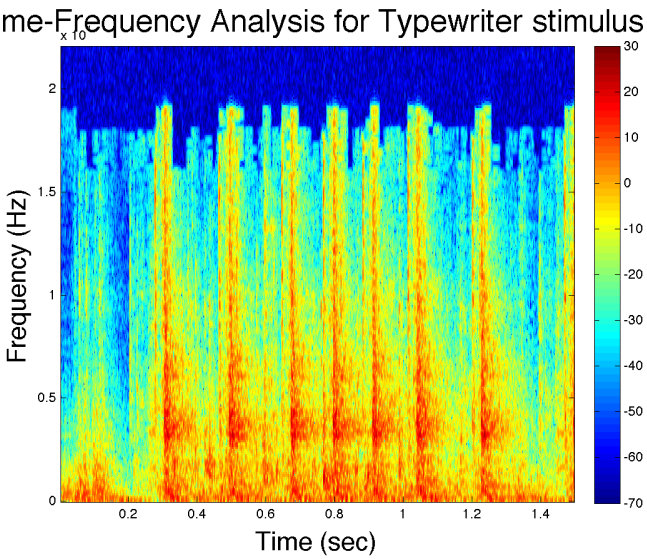

Time-Frequency Analysis for Roosters stimulus
x 10^4
Frequency (Hz)
Time (sec)
0
0.5
1
1.5
2
0.2
0.4
0.6
0.8
1
1.2
1.4
30
20
10
0
-10
-20
-30
-40
-50
-60
-70

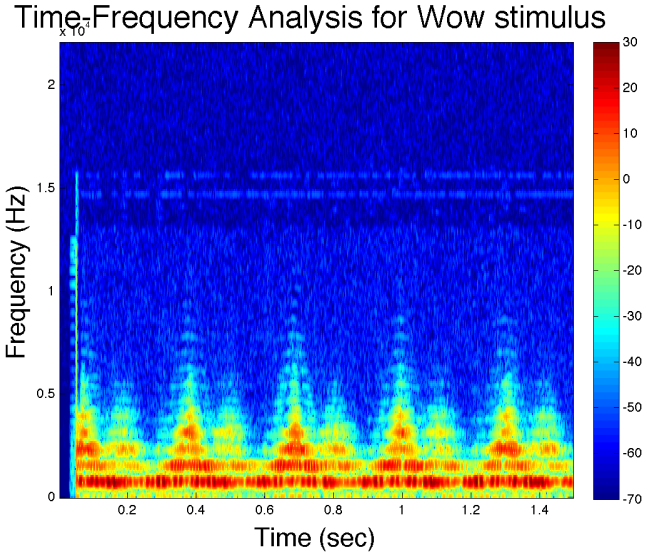

Time-Frequency Analysis for Sirens stimulus
x 10^4
Frequency (Hz)
Time (sec)
2
1.5
1
0.5
0
0.2
0.4
0.6
0.8
1
1.2
1.4
30
20
10
0
-10
-20
-30
-40
-50
-60
-70

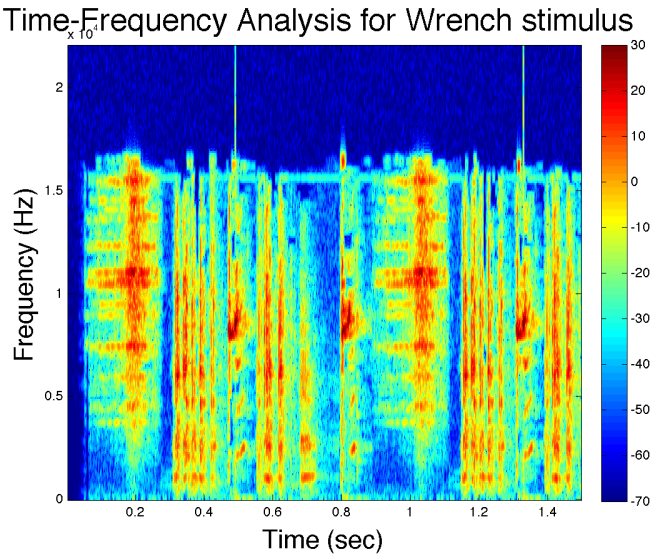

Time-Frequency Analysis for Snoring stimulus
x 10^4
Frequency (Hz)
Time (sec)
0
0.5
1
1.5
2
0.2
0.4
0.6
0.8
1
1.2
1.4
30
20
10
0
-10
-20
-30
-40
-50
-60
-70


Time-Frequency Analysis for Space Gun stimulus
x 10^4
Frequency (Hz)
Time (sec)
0
0.5
1
1.5
2
0.2
0.4
0.6
0.8
1
1.2
1.4
30
20
10
0
-10
-20
-30
-40
-50
-60
-70


Time-Frequency Analysis for Tick-Tock stimulus
x 10^4
Frequency (Hz)
Time (sec)
0
0.5
1
1.5
2
0.2
0.4
0.6
0.8
1
1.2
1.4
30
20
10
0
-10
-20
-30
-40
-50
-60
-70


Time-Frequency Analysis for Typewriter stimulus
x 10
Frequency (Hz)
Time (sec)
0
0.5
1
1.5
2
0.2
0.4
0.6
0.8
1
1.2
1.4
30
20
10
0
-10
-20
-30
-40
-50
-60
-70


Time-Frequency Analysis for Wow stimulus
x 10
Frequency (Hz)
Time (sec)
0
0.5
1
1.5
2
0.2
0.4
0.6
0.8
1
1.2
1.4
30
20
10
0
-10
-20
-30
-40
-50
-60
-70


Time-Frequency Analysis for Wrench stimulus
x 10^4
Frequency (Hz)
Time (sec)
0
0.5
1
1.5
2
0.2
0.4
0.6
0.8
1
1.2
1.4
30
20
10
0
-10
-20
-30
-40
-50
-60
-70

BIBLIOGRAPHY

BIBLIOGRAPHY

Albert, W. S., R. A. Rensink, and J. M. Beusmans (1999), Learning relative directions between landmarks in a desktop virtual environment, *Spatial Cognition and Computation*, *2*, 131–144.

Algazi, V., and R. Duda (2001), Estimation of a spherical-head model from anthropometry, *Journal of the Audio Engineering Society*, *49*(6), 472–478.

Algazi, V., R. Duda, and D. Thompson (2002), Use of head and torso methods for improved spatial sound synthesis, *Proceeding of AES 113th Convention*.

Anderson, M., D. Begault, M. Godfroy, J. D. Miller, A. Roginska, and E. Wenzel (2006), Design and verification of headzap, a semi-automated hrir measurement system, in *Audio Engineering Society Convention 120*.

Arons, B. (1992), A review of the cocktail party effect, *JOURNAL OF THE AMERICAN VOICE I/O SOCIETY*, *12*, 35–50.

Ashmead, D. H., D. L. Davis, and A. Northington (1995), Contribution of listeners approaching motion to auditory distance perception., *Journal of Experimental Psychology: Human Perception and Performance*, *21*(2), 239–256.

Banks, M. S., and D. M. Green (1973), Localization of high- and low-frequency transients, *The Journal of the Acoustical Society of America*, *53*(5), 1432–1433, doi: 10.1121/1.1913489.

BBC (1991), *The BBC Sound Effects Library*, vol. 1-40, Princeton, NJ: Films for the Humanities and Social Sciences, Princeton, NJ.

BDTI (2003), *DSPs Adapt to New Challenges*, Berkeley Design Technology Inc.

Begault, D. (1994), *3-D sound for virtual reality and multimedia*, Cambridge: Academic Press Professional.

Begault, D. (2000), 3-d sound for virtual reality and multimedia.

Blanco-Martin, E., S. Merino Saez-Miera, J. J. Gomez-Alfageme, and L. I. Ortiz-Berenguer (2011), Repeatability of localization cues in hrtf data bases, in *Audio Engineering Society Convention 130*.

Blauert, J. (1983), *Spatial hearing: the psychophysics of human sound localization*, MIT Press.

Blauert, J., M. Brueggen, A. W. Bronkhorst, R. Drullman, G. Reynaud, L. Pellieux, W. Krebber, and R. Sottek (1998), The audis catalog of human hrtfs, *Journal of the Acoustical Society of America*, *103*, 3082.

Bonanni, R., P. Pasqualetti, C. Caltagirone, and G. A. Carlesimo (2007), Primacy and recency effects in immediate free recall of sequences of spatial positions, *Perceptual and Motor Skills*, *105*, 483–500.

Bronkhorst, A. (1995), Localization of real and virtual sound sources, *Journal of the Acoustical Society of America*, *98*(5), 2542–2553.

Brungart, D., and B. Simpson (2001), Auditory localization of nearby sources in a virtual audio display, *Applications of Signal Processing to Audio and Acoustics*, pp. 107–110.

Buechner, S. J., C. Hölscher, and J. M. Wiener (2009), Search strategies and their success in a virtual maze, *Proceedings of the 31th Annual Conference of the Cognitive Science Society*, pp. 1066–1071.

Carlile, S. (1996), *The Physical and Psychophysical Basis of Sound Localization*, chap. 2, RG Landes.

Carlson, L. (2008), On the "whats" and "hows" of "where": The role of salience in spatial descriptions, in *Spatial Cognition VI. Learning, Reasoning, and Talking about Space*, *Lecture Notes in Computer Science*, vol. 5248, pp. 4–6, Springer Berlin / Heidelberg.

Casali, J. G., and W. Wierwille (1986), Vehicular simulator-induced sickness (report no. ntsc-tr86-012 (ad-a173 226), *Tech. Rep. 3*, Arlington, VA: Naval Training Systems Center.

Cheng, C. I., and G. H. Wakefield (2001), Moving sound source synthesis for binaural electroacoustic music using interpolated head-related transfer functions (hrtfs), *Computer Music Journal*, *25*(4), 57–80.

Cohen, P. R., and D. R. McGee (2004), Tangible multimodal interfaces for safety-critical applications, *Communications of the ACM*, *47*(1), 41–46.

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

Eyre, J., and J. Bier (2000), The evolution of dsp processors.

Gardner, W. G., and K. D. Martin (1995), Hrtf measurements of a kemar, *Journal of the Acoustical Society of America*, *97*(6), 3907–3908.

Grantham, D. W. (1995), Spatial hearing and related phenomena, *Hearing*, *12*(1), 297–345.

Green, D. M. (1976), *An Introduction to Hearing*, Hillsdale, NJ: Erlbaum.

Gumerov, N. A., R. Duraiswami, and Z. Tang (2002), Numerical study of the influence of the torso on the hrtf, *Acoustics Speech and Signal Processing 2002 Proceedings ICASSP 02 IEEE International Conference on*, *2*, 1965–1968.

Hahm, J., K. Lee, S.-L. Lim, S.-Y. Kim, H.-T. Kim, and J.-H. Lee (2007), Effects of active navigation on object recognition in virtual environments., *Cyberpsychology and Behavior*, *10*(2), 305–308.

Hammershoi, D., H. Moller, M. F. Sorensen, and K. Larsen (1992), Head–related transfer functions: Measurements on 24 subjects, in *92nd Audio Engineering Society Convention*.

Hartmann, W. M., and A. Wittenberg (1996), On the externalization of sound images, *Journal of the Acoustical Society of America*, *99*(6), 3678–3688.

Hass, J., J. Gibson, and C. Cook (2003), Introduction to midi and computer music: Digital audio concepts, webpage.

Hebrank, J., and D. Wright (1974), Spectral cues used in the localization of sound sources on the median plane, *Journal of the Acoustical Society of America*, *56*(6), 1829–1834.

Hill, E., J. J. Reiser, M. M. Hill, and J. Haplin (1993), How persons with visual impairments explore novel spaces: strategies of good and poor performers, *Journal of Visual Impairment and Blindness*, *87*(8), 295–301.

Hofman, P. M., A. J. V. Opstal, and J. G. A. V. Riswick (1999), Relearning sound localization with new ears, *Journal of the Acoustical Society of America*, *105*(2), 1035–1035, doi:DOI:10.1121/1.424942.

Holland, S., D. R. Morse, and H. Gedenryd (2002), Audiogps: Spatial audio navigation with a minimal attention interface, *Personal Ubiquitous Comput.*, *6*(4), 253–259, doi:http://dx.doi.org/10.1007/s007790200025.

Inoue, N., T. Kimura, T. Nishino, K. Itou, and K. Takeda (2005), Evaluation of hrtfs estimated using physical features, *Acoustical Science And Technology*, *26*(5), 453–455.

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

Jin, C., P. Leong, J. Leung, A. Corderoy, and S. Carlile (2000), *Enabling individualized virtual auditory space using morphological measurements*, pp. 235–238, IEEE PacificRim Conference on Multimedia.

Kay, L. (2011), Bay advanced technologies ltd., *http://www.batforblind.co.nz/*.

Kelly, J., M. Avraamides, and T. McNamara (2010), Reference frames influence spatial memory development within and across sensory modalities, in *Spatial Cognition VII*, Lecture Notes in Computer Science, pp. 222–233, Springer Berlin / Heidelberg.

Klatzky, R. L., Y. Lippa, J. M. Loomis, and R. G. Golledge (2002), Learning directions of objects specified by vision, spatial audition, or auditory spatial language., *Learning and Memory*, *9*(6), 364–367.

Klatzky, R. L., Y. Lippa, J. M. Loomis, and R. G. Golledge (2003), Encoding, learning, and spatial updating of multiple object locations specified by 3-d sound, spatial language, and vision., *Experimental brain research Experimentelle Hirnforschung Expérimentation cérébrale*, *149*(1), 48–61.

Lackner, J. (1990), Human orientation, adaptation, and movement control, *Tech. rep.*, Washington, DC - Ballistic Research Laboratory.

Lewald, J. (2002), Rapid adaptation to auditory-visual spatial disparity, *Learning Memory*, *9*(5), 268–278.

Loomis, J. M. (1985), Digital map and navigation system for the visually impaired, unpublished.

Loomis, J. M., C. Hebert, and J. G. Cicinelli (1990), Active localization of virtual sounds, *The Journal of the Acoustical Society of America*, *88*(4), 1757–1764, doi: 10.1121/1.400250.

Loomis, J. M., R. L. Klatzky, R. G. Golledge, J. G. Cicinelli, J. W. Pellegrino, and P. A. Fry (1993), Nonvisual navigation by blind and sighted: assessment of path integration ability., *Journal of experimental psychology General*, *122*(1), 73–91.

Loomis, J. M., R. L. Klatzky, J. W. Philbeckand, and R. G. Golledge (1998), Assessing auditory distance perception using perceptually directed action.

Loomis, J. M., Y. Lippa, R. L. Klatzky, and R. G. Golledge (2002), Spatial updating of locations specified by 3-d sound and spatial language, *Journal Of Experimental Psychology. Learning Memory And Cognition*, *28*(2), 335–345.

Martin, R. L., P. Flanagan, K. I. McAnally, and G. Eberle (2011), Memory for the locations of environmental sounds, *Journal of the Acoustical Society of America*, *129*(6), 3873–3883.

McNamara, T., J. Sluzenski, and B. Rump (2008), Human spatial memory and navigation, in *Learning and Memory: A Comprehensive Reference*, edited by J. H. Byrne, pp. 157 – 178, Academic Press, Oxford, doi:10.1016/B978-012370509-9.00176-5.

Middlebrooks, J. C. (1999), Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency., *Journal of the Acoustical Society of America*, *106*(3 Pt 1), 1493–1510.

Moller, H., M. F. Sorensen, C. B. Jensen, and D. Hammershoi (1996), Binaural technique: do we need individual recordings, *Journal of the Audio Engineering Society*, *44*(6), 451–469.

Nisbett, R. E., and T. D. Wilson (1977), Telling more than we can know: Verbal reports on mental processes, *Psychological Review*, *84*(3), 231–259.

Parmentier, F. B., and D. M. Jones (2000), Functional characteristics of auditory temporal-spatial short-term memory: evidence from serial order errors, *Journal Of Experimental Psychology. Learning Memory And Cognition*, *26*(1), 222–238.

Perrett, S., and W. Noble (1997), The effect of head rotations on vertical plane sound localization., *Journal of the Acoustical Society of America*, *102*(4), 2325–2332.

Peterson, L., D. Cory, W. Wiener, and A. Brooks (1998), Orientation and mobility services for all individuals with functional mobility limitations, *Proceedings of the 9th International Mobility Conference*, pp. 366–370.

Pollack, I., and M. Rose (1967), Effect of head movement on the localization of sounds in the equatorial plane, *Attention, Perception, and Psychophysics*, *2*, 591–596.

Postma, A., and E. H. De Haan (1996), What was where? memory for object locations., *Quarterly Journal of Experimental Psychology*, *49*(1), 178–199.

Raycal, N. (2006), Nurion raycal, *http://www.nurion.net/LC.html*.

Rayleigh, L. (1907), On our perception of sound direction, *Philosophical Magazine*, *13*(74), 214–232.

Reason, J. T., and E. Diaz (1971), imulator sickness in passive observers, *Tech. rep.*, Flying Personnel Research.

Recanzone, G. H. (1998), Rapidly induced auditory plasticity: The ventriloquismaftereffect, *Proceedings of the National Academy of Sciences of the United States of America*, *95*(3), 869–875.

Roginska, A., G. H. Wakefield, and T. S. Santoro (2010a), User selected hrtfs: Reduced complexity and improved perception, *Tech. rep.*, Undersea Human Systems Integration Symposium.

Roginska, A., G. H. Wakefield, T. S. Santoro, and K. A. McMullen (2010b), Effects of interface type on navigation in a virtual spatial auditory environment, in *Proceedings of the International Conference on Auditory Displays*.

Roginska, A., G. H. Wakefield, and K. McMullen (2011), Searching for sources from a fixed point in a virtual auditory environment, *The 17th Annual Conference on Auditory Display*.

Rundus, D. (1971), Analysis of rehearsal processes in free recall, *Journal of Experimental Psychology*, *89*, 63–77.

Runkle, P., A. Yendiki, and G. H. Wakefield (2000), Active sensory tuning for immersive spatialized audio, *Proceeding of International Conference on Auditory Display*.

Rutherford, P., and D. Withington (2001), The application of virtual acoustic techniques for the development of an auditory navigation beacon used in building emergency egress, *Proceedings of the 2001 International Conference on Auditory Displays*.

Saito, K., and S. Iwaya, Y. (2004), The technique of choosing the individualized head-related transfer function based on localization, *Tech. rep.*, IEICE.

Seeber, B., and H. Fastl (2003), Subjective selection of non-individual head-related transfer functions, *Proceeding of ICAD 2003*, pp. 259–262.

Shilling, R., and B. G. Shinn-Cunningham (2000), Virtual environments hanbook, *Handbook of Virtual Environment Technology*.

Silzle, A. (2002), Selection and tuning of hrtfs, *AES*, pp. 1–14.

Sobey, E. J. C. (2006), *A field guide to roadside technology*, Chicago Review Press.

Stanney, K. M., and P. Hash (1998), Locus of user-initiated control in virtual environments: Influences on cybersickness, *Presence: Teleoperators and Virtual Environments*.

Tellevik, J. M. (1992), Influence of spatial exploration patterns on cognitive mapping by blindfolded sighted persons, *Journal of Visual Impairment and Blindness*, *92*, 221 – 224.

Terai, K., and I. Kakuhari (2003), Hrtf calculation with less influence from 3-d modeling error: Making a physical human head model from geometric 3-d data, *Acoustical Science and Technology*, *24*(5), 333–334.

Thinus-Blanc, C., and F. Gaunet (1997), Representation of space in blind persons: Vision as a spatial sense?, *Psychological Bulletin*, *121*, 20–42.

Tran, T. V., T. Letowski, and K. S. Abouchacra (2000), Evaluation of acoustic beacon characteristics for navigation tasks, *Ergonomics*, *43*, 807–827.

Tretter, S. A. (2008), Overview of the hardware and software tools, in *Communication System Design Using DSP Algorithms with Laboratory Experiments for the TMS320C6713^{TM} DSK*, edited by J. K. Wolf, R. J. McEliece, J. Proakis, and W. H.

Tranter, Information Technology: Transmission, Processing and Storage, pp. 1–28, Springer US.

Tversky, B. (1993), Cognitive maps, cognitive collages and spatial mental models, in *Spatial Information Theory: A Theoretical Basis for GIS – Proceedings of COSIT'93*, edited by A. Frank and I. Campari, pp. 14–24, Springer, Berlin.

Von Wright, J. M., P. Loikkanen, and P. Reijonen (1978), A note on the development of recall of spatial location., *British Journal of Psychology*, *69*(2), 213–216.

Walker, B. N., and J. Lindsay (2005), Navigation performance in a virtual environment with bonephones, *Proceedings of the 11th International Conference on Auditory Display (ICAD2005)*, pp. 260–263.

Wallach, H. (1940), The role of head movements and vestibular and visual cues in sound localization, *Journal of Experimental Psychology*, *27*(4), 339–368.

Warren, R. M., and J. A. Bashford (1981), Perception of acoustic iterance: pitch and infrapitch., *Perception And Psychophysics*, *29*(4), 395–402.

Wenzel, E., F. Wightman, D. Kistler, and S. Foster (1988a), Acoustic origins of individual differences in sound localization behavior, *Journal of the Acoustic Society of America*.

Wenzel, E., M. Arruda, D. Kistler, and F. Wightman (1993), Localization using nonindividualized head-related transfer functions, *Journal of the Acoustical Society of America*, *94*(1), 111–123.

Wenzel, E. M. (1992), Localization in virtual acoustic displays, *Presence: Teleoper. Virtual Environ.*, *1*(1), 80–107.

Wenzel, E. M., F. L. Wightman, and S. H. Foster (1988b), A virtual display system for conveying three-dimensional acoustic information, *Human Factors and Ergonomics Society Annual Meeting Proceedings*, *32*, 86–90(5).

Wightman, F., and D. J. Kistler (1989a), Headphone simulation of free-field listening. part 2: psychophysical validation., *Journal of the Acoustical Society of America*, *85*(2), 868–878.

Wightman, F. L., and D. J. Kistler (1989b), Headphone simulation of free-field listening. part 1: Stimulus synthesis, *Journal of the Acoustical Society of America*, *85*(2), 858–867.

Wightman, F. L., and D. J. Kistler (1999), Resolution of front-back ambiguity in spatial hearing by listener and source movement, *Journal of the Acoustical Society of America*, *105*(5), 2841–2853.

Yost, W. A., F. L. Wightman, and D. M. Green (1971), Lateralization of filtered clicks, *The Journal of the Acoustical Society of America*, *50*(6B), 1526–1531, doi: 10.1121/1.1912806.

Zahorik, P. (2002), Auditory display of sound source distance, *Source*, pp. 1–7.

Zahorik, P., P. Bangayan, V. Sundareswaran, K. Wang, and C. Tam (2006), Perceptual recalibration in human sound localization: Learning to remediate front-back reversals, *Journal of the Acoustical Society of America*, *120*(1), 343–359.

Zotkin, D., R. Duraiswami, L. Davis, A. Mohan, and V. Raykar (2002), Virtual audio system customization using visual matching of ear parameters, *Pattern Recognition 2002 Proceedings 16th International Conference on*, *3*, 1003 — 1006 vol.3.

Zotkin, D. Y. N., J. Hwang, R. Duraiswaini, and L. S. Davis (2003), *HRTF personalization using anthropometric measurements*, pp. 157–160, Ieee.