12-1-2012

Assessment of Audio Interfaces for use in Smartphone Based Spatial Learning Systems for the Blind

Shreyans Jain

ASSESSMENT OF AUDIO INTERFACES FOR USE IN SMARTPHONE BASED

SPATIAL LEARNING SYSTEMS FOR THE BLIND

By

Shreyans Jain

B.E. (Hons), Rajiv Gandhi Technological University, Bhopal, India

A THESIS

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

(in Spatial Information Science and Engineering)

The Graduate School

The University of Maine

December, 2012

Advisory Committee:

Nicholas A. Giudice, Assistant Professor of School of Computing and Information Science, Advisor

Kate Beard-Tisdale, Professor of School of Computing and Information Science

Reinhard Moratz, Associate Professor of School of Computing and Information Science

THESIS ACCEPTANCE STATEMENT

On behalf of the Graduate Committee for Shreyans Jain, I affirm that this manuscript is

the final and accepted thesis. Signatures of all committee members are on file with the

Graduate School at the University of Maine, 42 Stodder Hall, Orono, Maine.

Dr. Nicholas A. Giudice,   Date
Assistant Professor of School of Computing and Information Science

LIBRARY RIGHTS STATEMENT

In presenting this thesis in partial fulfillment of the requirements for an advanced degree at The University of Maine, I agree that the Library shall make it freely available for inspection. I further agree that permission for "fair use" copying of this thesis for scholarly purposes may be granted by the Librarian. It is understood that any copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Signature:

Date:

ASSESSMENT OF AUDIO INTERFACES FOR USE IN SMARTPHONE BASED

SPATIAL LEARNING SYSTEMS FOR THE BLIND

By Shreyans Jain

Thesis Advisor: Dr. Nicholas A. Giudice

An Abstract of the Thesis Presented
in Partial Fulfillment of the Requirements for the
Degree of Master of Science
(in Spatial Information Science and Engineering)
December, 2012

Recent advancements in the field of indoor positioning and mobile computing promise development of smart phone based indoor navigation systems. Currently, the preliminary implementations of such systems only use visual interfaces—meaning that they are inaccessible to blind and low vision users. According to the World Health Organization, about 39 million people in the world are blind. This necessitates the need for development and evaluation of non-visual interfaces for indoor navigation systems that support safe and efficient spatial learning and navigation behavior.

This thesis research has empirically evaluated several different approaches through which spatial information about the environment can be conveyed through audio. In the first experiment, blindfolded participants standing at an origin in a lab learned the distance and azimuth of target objects that were specified by four audio modes. The first three modes were perceptual interfaces and did not require cognitive mediation on the part of the user. The fourth mode was a non-perceptual mode where object descriptions were given via spatial language using clockface angles. After learning the targets through the four modes, the participants spatially updated the position of the targets and localized

them by walking to each of them from two indirect waypoints. The results also indicate hand motion triggered mode to be better than the head motion triggered mode and comparable to auditory snapshot.

In the second experiment, blindfolded participants learned target object arrays with two spatial audio modes and a visual mode. In the first mode, head tracking was enabled, whereas in the second mode hand tracking was enabled. In the third mode, serving as a control, the participants were allowed to learn the targets visually. We again compared spatial updating performance with these modes and found no significant performance differences between modes. These results indicate that we can develop 3D audio interfaces on sensor rich off the shelf smartphone devices, without the need of expensive head tracking hardware.

Finally, a third study, evaluated room layout learning performance by blindfolded participants with an android smartphone. Three perceptual and one non-perceptual mode were tested for cognitive map development. As expected the perceptual interfaces performed significantly better than the non-perceptual language based mode in an allocentric pointing judgment and in overall subjective rating.

In sum, the perceptual interfaces led to better spatial learning performance and higher user ratings. Also there is no significant difference in a cognitive map developed through spatial audio based on tracking user's head or hand. These results have important implications as they support development of accessible perceptually driven interfaces for smartphones.

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Motivation

The last few years have seen a massive change in how people navigate from one place to another in outdoor environments. Global Positioning System (GPS) based in-vehicle navigation systems allow people to reach their destinations easily and on time. There also has been an impressive improvement in smartphone based pedestrian navigation systems. Smartphones with their embedded sensors such as GPS, Wi-Fi, accelerometers, gyroscopes, etc. serve as a great tool for not only navigating users from one place to another, but also for making them aware of their surroundings. A recent study by the Pew Research Centre indicated that about 74% of the total smartphone owners used location based services to get directions and recommendations about places nearby, based on their current position (Zickuhr, 2012). Therefore, the use of outdoor location based systems is on the rise owing to their navigation and exploration applications.

On the other hand, with the rapid urbanization and lack of space in cities, buildings are becoming more and more complex. One example of such a building is the Seattle Central Library- the flagship library of Seattle's public library system. The library opened to rave reviews in 2004. However, the library had to hire a professional "wayfinder" to install navigational signs inside the building as more and more people were getting lost inside the library (Murakami, 2006). Indeed, most of us have lost our way inside large indoor spaces such as a shopping mall, airport, conference center or a library. Outdoor

navigation systems based on GPS do not work inside buildings because the GPS signals are attenuated and scattered by roofs, walls and other objects (Dedes & Dempster, 2005; El-Natour, Escher, Macabiau, & Boucheret, 2005).

Some environmental attributes of indoor spaces make their learning and navigation more difficult than outdoor navigation (Giudice, Walton, & Worboys, 2010). First, the availability and the nature of landmarks are significantly different in outdoor navigation as compared to their indoor counterparts. For outdoor navigation, the user has access to large, permanent landmarks such as mountains and lakes which are accessible from multiple locations and are independent from the route taken; in contrast, indoor navigation affords access to smaller "local" landmarks such as water coolers, paintings, and walls, which are dependent upon the route taken by the user and fail to provide a global view because of occlusion caused due to walls and the user's limited field of view. The second attribute that makes indoor navigation problematic is the absence of a consistent structure (such as a block) and names for hallways and corridors. While in cities we find named streets such as Fifth Street, Broadway, Main street etc.; indoor addresses usually do not go beyond specific room numbers. According to an Environmental Protection Agency (EPA) survey, the average American spends 87 % of their time indoors (Klepeis et al., 2001). Thus, to provide the ability to navigate within buildings similar to how we navigate outdoors is a difficult challenge.

Thanks to recent commercial and research initiatives, we now have several indoor positioning and mapping systems for smartphones. One of the most popular commercial application in this domain is the Google Maps for indoors(McClendon, 2011). Some other applications are Micello (Micello, Sunnyvale CA) and Point inside (Point Inside

Bellevue, WA). While Micello and Point inside only provide indoor maps, Google maps

for indoors (Fig 1.1) also tracks the position of the user in the indoor space with the help

of added sensors in the building.
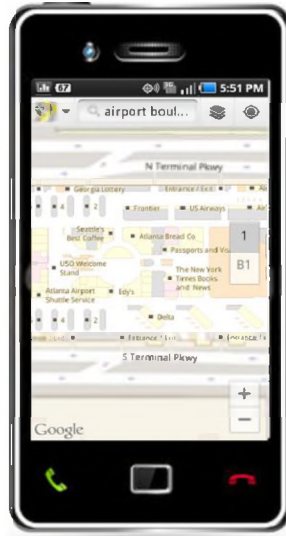
Figure 1.1 Google Maps for Indoors Running on Android Smartphone

The user can also query the location of the nearest point of interests in the building, for

example: the nearest coffee shop or restroom. The system also provides route information

to the user's destination on the map.

Thus smartphone based indoor mapping and navigation systems such as google maps for

indoors have a number of advantages.

1. They utilize off the shelf smartphone devices as the core platform. The use of

   smartphones has already become widespread among the masses. The total number

   of smartphone users in the world recently crossed the 1 billion mark and is

   expected to double by 2015(Yang, 2012). In addition, according to a latest

Nielson survey 50.4% of the total mobile phone subscribers used a smartphone as their main phone(Nielson, 2012).

2. The systems which track the user's location are able to update and zoom the visual content of the map based on the user's location, allowing them to find nearby points of interests. They thus help in the cognitive map development of the space.

3. The smartphone based systems have the ability to route users to their destinations with the help of highlighted routes.

However, the current implementations also have two limitations.

1. These devices use vision as the only output modality for providing spatial content and other information to the users. This limits their use only to the sighted community. To realize why this is a problem, the reader is invited to imagine the following persona.

"Rita is an established researcher in the field of computer science. She is visually impaired and has a guide dog- pluto which helps her to avoid obstacles as she navigates. Rita, being a computer scientist is technology savvy and owns the latest smartphone. She uses her smartphone for note-taking through voice input, and has a calendar application to help her manage her schedule. She owns a very high quality pair of bone conduction head phones, which she uses to listen to music, while commuting."

Now, let us imagine a day in the life of Rita. Being an accomplished scientist, she attends several conferences related to her research interest every year.

"This year she has been invited to San Diego, California to speak at a very large conference at the San Diego convention center. She is visiting San Diego for the first time. She takes a cab from her hotel and reaches the convention center. She is now at the entrance of the building and needs to reach room 116, where her talk has been scheduled. Even though, she has the latest version of Google maps on her phone, she is unable to use it because of its lack of non-visual support. She asks help from the information desk and reaches her destination."

This problem is typical for many low vision users who are unable to visually attend to the screen of a smartphone device, as all of the above applications provide only a visual interface to the user. Also, the absence of vision makes it impossible for them to use the visual cues provided by the environment which are otherwise available to the sighted individuals (Giudice & Legge, 2008). These visual cues (e.g. a painting, a signboard etc.) are extremely helpful for the sighted to be able to find their way in indoor environments.

The ability to navigate independently and confidently is an essential part of everyone's life. Every day we perform some form of navigation, be it to the work, grocery store or to school. While it might be comparatively easy to navigate to familiar and predictable places with limited or no vision e.g. finding your correct seat at the movie theatre, or finding the candle in the kitchen drawer when the power goes out, it is extremely difficult to navigate in unfamiliar and unpredictable environments. Imagine you are wearing a blindfold and are standing at the entrance of your local shopping mall. You are now asked to walk to your favorite coffee shop, which you visit on a daily basis. Even though you would have a mental representation of the mall, the concern of hitting obstacles and people while on your way would make your travel difficult. You would probably be able

to navigate to the shop, though you would be considerably slow and would achieve this task with great cognitive effort.

Most sighted people have to deal with situations like finding the ticket/check in counter at an unfamiliar airport and boarding the flight, often under time constraints. Sighted people have access to visual signage and terminal maps, which mitigate this problem to a great degree. Now imagine you are again blindfolded but this time you are left at the gate of an unknown airport. Your task is to locate the ticket counter, buy a ticket to Omaha, Nebraska and then board the plane. This time you will find the task extremely hard or perhaps impossible. It would be harder than navigating in the familiar mall as this time you would have no access to the spatial representation you had in the previous case.

Rita would face the same problem on her return journey at the San Diego airport, as unlike her sighted counterparts, she was not able to learn the layout of the airport on her arrival. In fact most blind people encounter such situations many times in their lives. According to the World Health Organization, about 285 million people are visually impaired worldwide (WHO, 2012). This community is therefore in need of a navigation and environmental awareness tool which is both inexpensive yet accessible.

While the smartphone based systems we discussed earlier in this section provide a ready and comparatively cheap solution compared to specialized assistive equipment, the lack of alternate interfaces makes their use almost impossible for the blind community, who as we discussed are in the greatest need of such a system.

2. The current implementations of smartphone based indoor mapping and positioning systems provide the user information about the location of rooms and pathways (Fig 1.1).

This knowledge is sufficient for sighted users as they can navigate easily once they arrive at smaller rooms, so do not need the system to also describe these spaces. However, blind individuals would need another level of granularity of spatial details about the environment for successful navigation. To understand this clearly, we need to revisit Rita. *"Rita has just finished her highly admired keynote speech. The next item for her today's agenda is to meet with the database specialty group at a meeting room (Room 210) in the convention center. Again with the help at the information desk she is able to find Room 210. However, the next problem for her is to find the location of an empty chair. She was able to find a chair with some help. After the meeting the group decides to go for lunch at the nearest Subway sandwich shop. Rita joins the queue for order and orders a customized sandwich. After getting the sandwich she is not sure about the location of the payment counter at the shop. She again gets help from her colleague and pays the bill."*

To accomplish tasks in small enclosed spaces such as rooms and restaurants, we need to have a spatial representation of the space. For example, as we discussed earlier it is easier for us to find the location of the familiar drawer in an event of a power outage. Now again imagine you are wearing a blindfold and are asked to locate the check-in counter at a hotel lobby. In the absence of a cognitive map, this task is extremely difficult to achieve. The smartphone applications we discussed earlier do not have a provision to provide room level spatial details. As we just discussed, this feature should be an important component of any blind navigation or mapping system.

Therefore though cheap and readily available, the smartphone based navigation systems suffer from two major limitations

(1) Lack of alternatives to vision to support blind and low-vision persons to navigate freely.

(2) Lack of provisions for room or finer level spatial navigation

In this thesis research, I have tried to address these limitations through a series of behavioral experiments. The next section, presents research focus, questions and hypotheses for this thesis.

1.2 Research Focus, Questions and Hypotheses

As we discussed in the last section, there is a growing need for the development of non-visual interfaces for smartphones to provide spatial information to the blind and low-vision community. We also saw the need for development for room level navigation systems. Both of these important issues are addressed in this research.

The main focus in this thesis research was the development and evaluation of smartphone based audio interfaces to help the visually impaired form accurate mental representations of a space and thus support independent, ideally stress-free and effective navigation. This leads us to the first research question (RQ1):

RQ1: Can audio be used as an alternative to vision to help blind individuals in forming accurate cognitive maps of indoor spatial layouts?

We wanted our interfaces to support fast yet effective navigation while putting minimal stress on the blind user. The argument advanced here is that this is best done using perceptual interfaces. Perceptual interfaces are those that directly convey spatial information through spatial senses like vision, audition, or touch and require no cognitive

mediation on the part of the user to accomplish spatial learning. Spatial audio or 3 D audio is one such perceptual interface where the sound has been processed so as to appear to come from the direction and distance of the target. Spatial learning of targets through pointing of a body part (hand/head) or through kinesthesis (sense of knowing the location of a body part) is also considered perceptual (Chapter 2 describes these interfaces in detail). This leads us to the second research question (RQ2):

RQ2: Are there differences in audio based perceptual interfaces in terms of speed and accuracy of mental representations formation?

A spatial language interface is defined as a mode of spatial learning in which directional and distance information is given through words, for example in clock directions such as 2 O Clock or through degrees such as 60 degrees right. This interface is a non-perceptual interface as it requires cognitive mediation of the signal (i.e., you must interpret the words, as they have no intrinsic spatial content) on the user's part to comprehend the direction of the target. This interface is the gold standard to convey spatial information to the users through audio, for example in car navigation system or pedestrian navigation systems.

The previous section noted that blind spatial navigation has inherent mental stress associated with it. We want our interface to help in development of accurate spatial representations while exerting minimal cognitive load on the user. This leads to the third research question (RQ3).

RQ3: Are there differences in perceptual and non-perceptual interfaces in terms of speed and accuracy of mental representations formation?

The spatial audio based interfaces require the user's head to be tracked for more accurate localization and removal of front back confusions (Chapter 4). However, these head trackers are expensive, and require the user to wear additional hardware. Since our system would be implemented on a smartphone, we propose the use of smartphone based hand tracking for immersive 3D audio generation. However, tracking the user's hand instead of their head is a novel approach and needs to be tested to establish the veracity of this concept (see chapter 4). This idea leads to our fourth research question (RQ4):

RQ4: Can head tracking be replaced with hand tracking to generate more immersive spatial audio?

Finally, we wanted to explore user preferences for the use of these non-visual interfaces. We addressed this issue through a fifth research question (RQ5):

RQ5: Are there users' preference differences with respect to effectiveness and usability of interfaces tested in this thesis?

Answers to these research questions were investigated through conceptualization and design of three behavioral experiments. Table 1.1 summarizes the purpose of each behavioral experiment and specific questions it answered.

Experiment Number	Purpose	Research Questions Answered
1.	To compare the efficacy of three perceptual (3 D audio, Hand pointing, Head pointing) and one non-perceptual audio interface (Spatial Language) in conveying spatial information	RQ1, RQ2, RQ3
2.	To compare cognitive map development through vision, 3D audio interface with head motion, 3D audio interface with hand motion	RQ1, RQ4
3.	To compare the efficacy of three perceptual (3 D audio, Kinesthetic, Kinesthetic with 3D audio) and one non-perceptual smartphone based audio interface (Spatial Language) in spatial representation development.	RQ1, RQ2, RQ3, RQ5

Table 1.1 Research Experiments and their Purpose

Hypotheses:

1) Audio based interfaces can be used as an alternative to visual interfaces for spatial information acquisition.

2) Perceptual interfaces lead to faster and more accurate spatial behavior and cognitive map development than non-perceptual language-based interfaces.

3) There is no significant difference in spatial behavior and cognitive map development when using hand tracked 3D audio versus the traditional approach of head tracked 3D audio.

4) The subjective preference ratings will favor perceptual interfaces over non-perceptual language-based interfaces.

1.3 Organization of the Remaining Chapters

The remaining chapters are organized as follows; Chapter 2 provides a brief overview of some of the current and past research using non-visual interfaces for navigation. Chapter 3 gives an overview of research on comparing perceptual and non-perceptual interfaces and describes the methods and results for experiment 1. Chapter 4 starts with explaining the importance of head tracking in 3D audio applications. It then describes a study (experiment 2) which compared head tracked 3D audio with hand tracked 3D audio and vision for a spatial updating task. Chapter 5 introduces kinesthetic interfaces as another mode of conveying spatial information perceptually. We then introduce two new modes for spatial learning based on kinesthetic cues, namely SpeakonTouch and Spatial SpeakonTouch and compare their ability to help build cognitive maps that support spatial behavior of different scenes learned through 3D audio versus spatial language via an empirical study (experiment 3). Finally, this thesis concludes with future directions in chapter 6.

CHAPTER 2

LITERATURE REVIEW

Spatial Information systems allow individuals to form a mental image of the space during or before their travel to an unknown space. The first and most common approach to provide spatial knowledge is the use of maps. These maps are either two dimensional or three dimensional and could be paper based or digital. Paper based maps and atlases have been used for centuries to help humans navigate through unknown territories. In fact, the earliest known world map, 'Imago Mundi' is commonly dated back to 6th century BCE (Raaflaub & Talbert, 2009). The use of paper based maps is still very common for navigation.

More recently, computer based digital maps have become popular. These maps are displayed on a computer screen as in Google maps (Google Maps, 2012) or in MapQuest (MapQuest, 2012). As we discussed in the previous chapter, similar to these outdoor spatial information systems, we now have smartphone based indoor mapping systems (Section 1.1). These systems have only visual interfaces making their use for blind and low-vision users impossible. According to the findings from the 2010 National Health Interview Survey (NHIS), about 21.5 million Americans had low vision(Schiller, Lucas, Ward, & Peregoy, 2012) . Thus, to bridge this gap, there is a need to develop accessible spatial information systems for these smartphone based systems that rely on more than purely visual interfaces. Two non-visual modalities, namely, audio and touch, have been investigated in the past to convey spatial information.

This chapter reviews some of the previous research on accessible interfaces that support cognitive map development for blind and low vision users. Even though the main research focus is in the use of audio, a brief discussion of how touch alone has been used to impart spatial information in the past is described in section 2.1. Section 2.2 then discusses the principle of human sound localization in section as it forms an important component of this thesis. Section 2.3, reviews how audio has been used in the past to aid cognitive map development for blind individuals. Section 2.4, concludes this review and provides broader contexts.

2.1 Tactile Spatial Information Systems

This section describes some of the tactile spatial information systems that have been used to convey map knowledge non-visually.

2.1.1 Braille Tactile Maps

The most common approach for providing spatial information to blind and low-vision users in the past has been the use of braille tactile maps. These maps are created by using a swell technique on heat sensitive paper or by embossing Braille on heavy card stock with the help of a special Braille printer. (Tatham & Dodds, 1988) provides a great overview of the design and construction issues of tactile maps. These maps can be used as a wayfinding support as described in (Golledge, 1991). However, these maps suffer from some significant drawbacks. First, according to the National Federation for the blind, fewer than ten percent of blind Americans can read Braille (Nuckols, 2009). Another problem associated with these maps is that Braille labeling is inflexible due to the fixed size of the cells (Tatham, 1991). A map without labels cannot be used to learn a

14

space effectively. The size and cost of a Braille printer and the non- refreshable nature of the maps produced further add to the problems. Therefore, there is a need to develop alternate systems that are inexpensive, portable, dynamic (that is they allow refreshable maps) and support universal design principles.

2.1.2 Refreshable Tactile Based Systems

The problem of the non-refreshable nature of braille maps led to the development of refreshable tactile systems. (Vidal-Verdú & Hafez, 2007) conceptualized a refreshable tactile screen similar to a computer monitor where the pixels are replaced by taxels which they describe as touch simulation units. The taxels are based on electromagnetic or piezoelectric simulators and help convey information to the blind by mechanical stimulation on touch. They can be further classified as static refreshable devices or dynamic refreshable devices.

The static refreshable devices are designed so that they can be explored with the help of the fingers. They usually comprise of large tactile screen and have many tactile actuators, which get actuated as the user moves their finger on the screen. Most of the commercial devices are based on either piezoelectric actuators for example ABTIM(ABTIM, Wuppertal Germany) or are based on micro solenoids (Schweikhardt & Klöper, 1984). The main problems associated with these devices are their power consumption and cost.

In dynamic refreshable displays the user need not move their fingers on the screen. Instead they use a small array of taxels coupled with a mouse which points to a virtual tactile screen. The pins actuate based on the position of the mouse. Most commercial devices are based on either piezoelectric actuators as with the OPTACON (Linvill &

15

Bliss, 1966), electromagnetic actuator like TACTACT (Kammermeier, Buss, & Schmidt, 2000) or are based on Shape Memory alloy based actuators, for example HAPTAC (Hasser & Roark, 1998). While the cost of dynamic refreshable displays is lower than their static counterparts, it is still higher than ordinary smartphones (target devices for this research). Longer training times and lower recognition rates as compared to static displays are the other drawbacks of these displays.

2.1.3 Force Feedback Devices

The force feedback based displays provide response to the user in the form of a haptic effect. These displays have been used in the past to provide spatial information non-visually. For example, (Rice, Jones, Golledge, & Jacobson, 2003) the authors defined "Virtual Walls" (line of force used to define a shape in a virtual domain) around the campus of University of California at Santa Barbara (UCSB). Each building on campus can be located and their shape determined with the help of a force feedback mouse. Some of the most popular force-feedback devices are the Logitech Wingman Force Feedback mouse (Logitech, Morges, Switzerland) and the Phantom (Sensable, Woburn, MA,).

While force feedback devices have an ability to convey to the user objects that have linear boundary, it is difficult to convey to the user objects that have irregular shape. This is a problem in learning indoor spatial layouts, where irregularly shaped objects are commonplace. They also have a limited extent and require constant map-panning to explore large maps.

Because of the limitations of tactile interfaces such as their non-refreshable screens, higher costs, limited screen extents, or lack of inherent information content bandwidth

they are rarely used in isolation. These interfaces are most commonly combined with an audio interface to better help in the development of cognitive maps for visually impaired. Section 2.3 reviews some previous research describing audio based interfaces.

The goal of this thesis research was the development of perceptual audio interfaces which support accurate spatial behavior and cognitive map development in an intuitive way, using inexpensive and readily available hardware. As described in section 1.2, one such interface is spatial audio or 3D audio which is defined as the sound is processed to give the listener the direction and distance of the source in 3D space. In order to understand the implementation limitations in our target device (i.e., a smartphone), we need to first understand the theory of human sound localization, which is discussed next in section 2.2.

2.2 Human Sound Localization

Although humans have the ability to localize objects through a number of different senses (vision, audio, touch, smell etc.), we primarily localize distal objects through either vision or audio. When using vision, four primary cues help in object localization, as described in more detail in (Mackensen, 2004). First, the brain records the angular displacements of both eyeballs through tactile information provided by the eye muscles when we focus on an object and uses this information to determine the lateral and vertical position of the object. Another cue that helps in determining the location of the object, especially its distance, is provided due to the lateral displacement of the two eyeballs which together form an optical angle. The curvature of our eye's lens also helps in the determination of the sharpness of the optical image which acts as yet another cue for

visual localization. Finally, the relationship between an object and its surrounding environment can help in determining its distance. This is because we already know the size of some known objects and when we see their size as compared to the environment we can judge the distance, called the size constancy effect.

There are a number of cues that enable our auditory system to determine the location of sound. (Mackensen, 2004) classifies the cues primarily into three categories: cues that are based on the sound source, environmental cues, and cues related to the individual listener.

The characteristics of the sound source play an important role in our ability to localize them. For example our ability to localize a sound may depend upon the frequency of the sound source. High frequency sounds have been found easier to localize (Roffler, 1968). Similarly, environmental cues such as sound reflections can play a major role in our ability to localize sound. Our correct perception of the distance of the sound depends on the presence or absence of reflections from surrounding materials such as walls and ceilings.

The cues related to the individual listener's head (e.g., size, shape, etc.) play a crucial role in the localization of sounds. They can be divided into three categories: 1) Interaural cues, 2) Head Related Transfer Function (HRTF) cues, 3) Head motion based cues

2.2.1 Interaural Cues

The duplex theory of sound localization first proposed by Lord Rayleigh in 1907 provided an explanation for human ability to localize sound. The two most important cues that enable us to determine the direction of the sound source originate from the

intensity and time differences of the waves as they reach our ears. These cues are collectively known as the interaural cues and are described in detail in (Rayleigh, 1907).

1. Interaural Time Difference

The sound waves start from a source, travel through a medium and finally reach both of our ears. If the sound source is located directly ahead (or behind) of us, it takes the same amount of time for the sound to reach both ears. If the source is located on the right side, the waves would reach the right ear slightly quicker than the left ear. Similarly, if the sound source is located somewhat on the left side, the waves would arrive at the left ear first. This effect is known as the interaural time difference and is an important cue in sound localization (Fig 2.1).

Sound Source

time (a) > time (b)
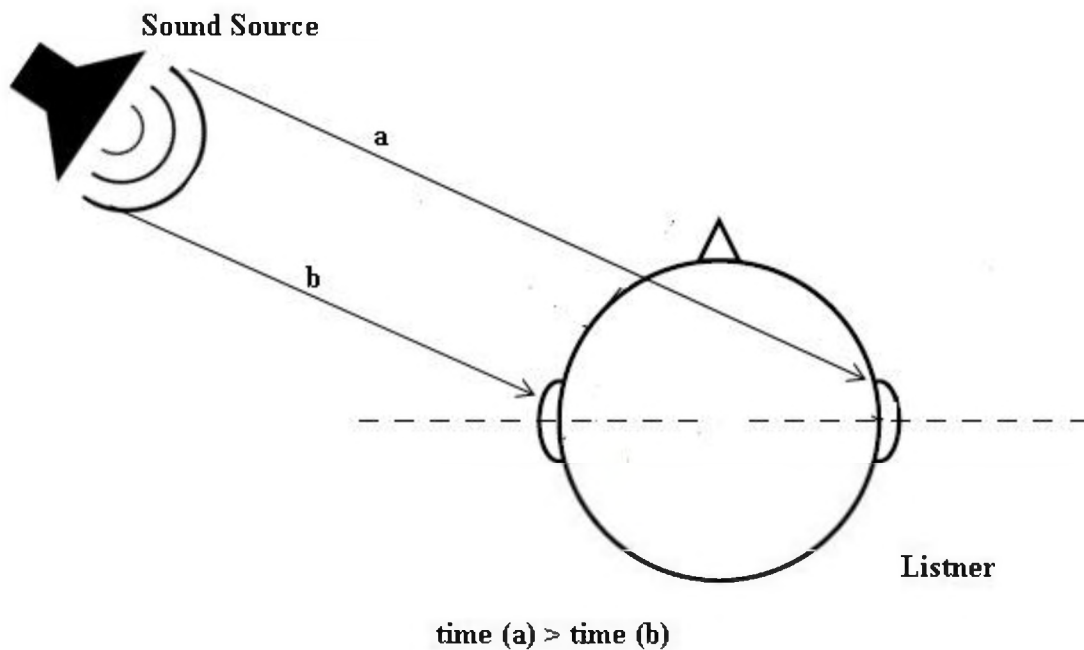
Figure 2.1 An Illustration of Interaural Time Difference Effect.

2. Interaural Level Difference

Another interaural cue that plays a major role in sound localization is the relative loudness of the sound reaching each ear. If the source is directly in front of the listener, the sound will have an equal level in both ears. However, if the sound source is more to the left side of the listener, the sound in the left ear would be louder than the right. Similarly an object on the right would sound louder in the right ear as compared to the left ear. This is because a sound shadow is formed on the far ear because of the blocking of the sound's line of path by the head (Fig 2.2). This effect is more prominent for high frequency sound waves, as the low frequency sound waves are less affected by this phenomenon because of their ability to bend around large objects.
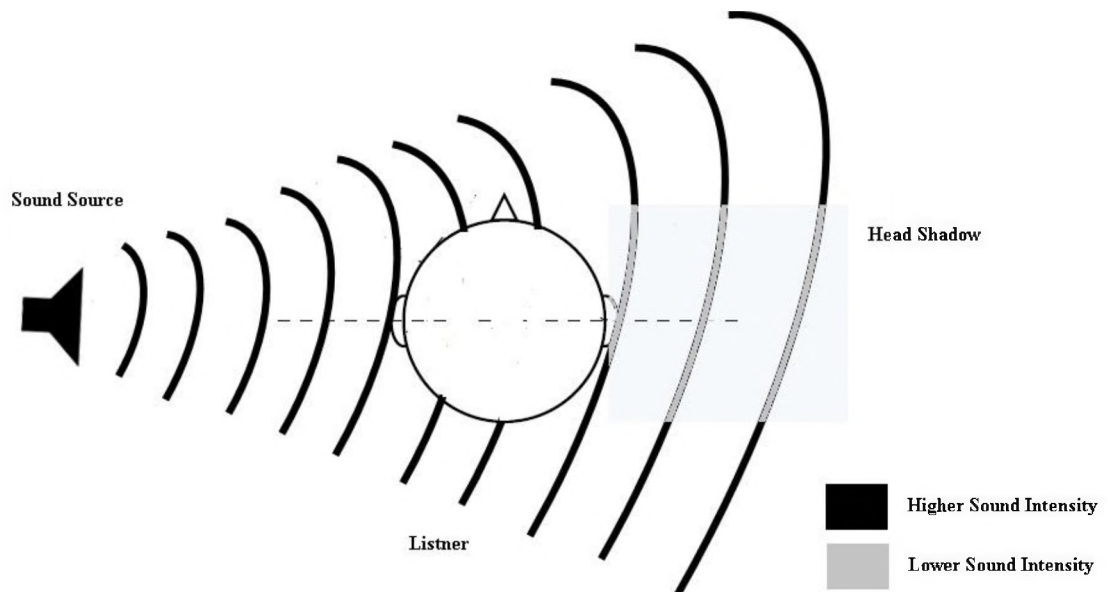
Figure 2.2 An Illustration of Interaural Level Difference Effect

2.2.2 Head Related Transfer Functions

Even though the duplex theory described by Lord Rayleigh is successful in explaining the binaural component of sound localization (cues derived from time/intensity differences between the two ears), a monaural component (cues derived by a single ear) critical in sound localization also exists. These monaural cues arise from the modification of the sound due to interaction with various parts of the human body such as head, shoulders, torso, and in particular our pinnae (outer ear) before entering the ear canal for further processing (Begault, 1994).

Previous research has tried to capture these cues through physical modeling (Shaw, 1974), empirical studies (Frederic L. Wightman, 1989) or through computer simulations(Kahana, Nelson, Petyt, & Choi, 1999). The captured parameters are called Head Related Transfer Functions (HRTF) and they encode the directional component of sound through monaural signals alone. Therefore separate HRTFs for left and right ear exists which describe the modification of the sound before it enters the left or right ear canal.

These HRTFs have been used extensively for generating spatial audio through headphones for example in (Bronkhorst, 1995; Wenzel, 1993). Even though they are a significant improvement over spatial audio synthesized using just the binaural cues, there are still a few problems. First, virtual sound sources located directly ahead of the listener sound "inside" their head (Griesinger, 1999). The second problem relates to the fact that HRTFs measured for one person do not necessarily work for another person(Pralong, 1996). This requires the measurement of HRTFs for each person individually to obtain

the most accurate auditory spatialization. However, HRTF databases such as the CIPIC HRTF database (Algazi, Duda, Thompson, & Avendano, 2001) exist which provide generic HRTFs which can drastically improve localization performance. Most soundcards also have a generic implementation of these HRTFs. Another problem is the inability of the listener to differentiate if the sound came from back of them or in front of them. This problem is commonly referred to as "Front-Back" confusion (Pralong, 1996) and can be resolved by tracking the head motion of the listener. This leads us to the next cue for listener specific cues for spatialization, head motion of the listener.

2.2.3 Head Motion of Listener

As we have discussed earlier, human ability to localize sound is based on the fact that we have two spatially separated ears. While plain interaural cues enable us to determine the direction of the sound source (left or right), they do not help us in knowing if the source is at front, back, above or below (Makous & Middlebrooks, 1990). For example, a source located at 45°, right and front (Point A in Fig 2.3) would have the same values of ITD as 45°, right and back (Point B in Fig 2.3). This virtual cone created at 45°, on the left in the proceeding example is termed as the "Cone of Confusion". A cone of confusion can occur at all positions between directly left and directly right of a listener's head as shown in (Fig. 2.3). In the figure points A and B would have the same intensity, which makes it difficult for the user to judge, if the target is in the front or at the back.
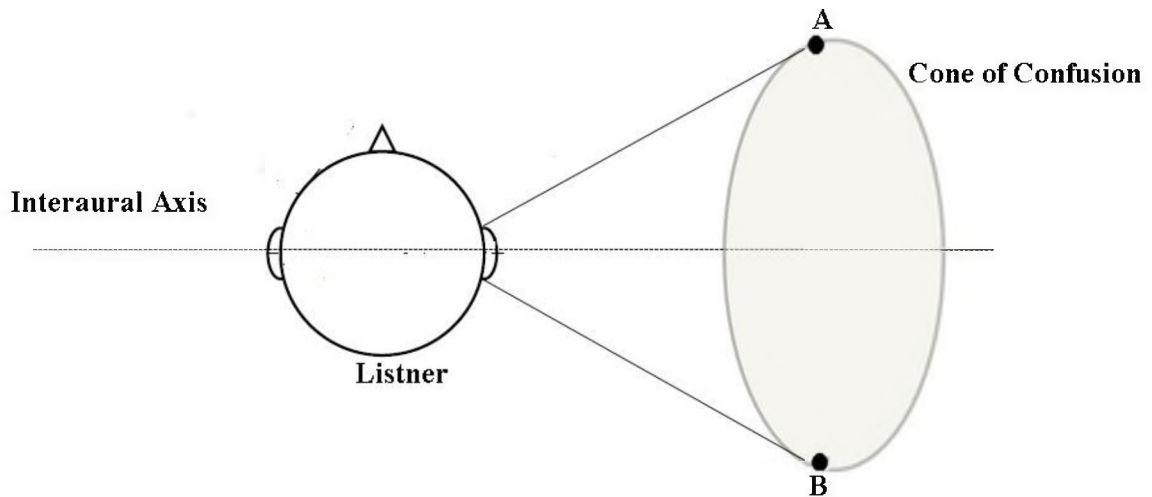
Figure 2.3 Cone of Confusion

To resolve the problem of front back confusion, and to improve the accuracy of sound localization, head motion by the listener has been accepted as a very important perceptual cue for audition (Perrett & Noble, 1997; Thurlow, Mangels, & Runge, 1967; Wallach, 1940). As we have seen (Fig. 2.3), if the sound source is positioned in the median plane, both ears receive the sound at the same instant. Thus there is no interaural time difference (section 2.2.2). This is true for both the front and back plane and the only way to resolve this ambiguity is through movement of the head. For example, if the sound source is located in the front and left, a movement of the head in a counter-clock wise direction would result in the sound as coming first and with more intensity towards the left ear, followed by the right ear. Similarly, on moving the head in a clockwise direction, this effect would be reversed. Thus, head motion acts as an important perceptual cue in determining the direction and distance of the sound. It is also worthwhile to note that to localize the sound correctly, it is important to know the direction of head movement (clockwise or counter-clockwise).

As discussed in chapter 1, this thesis research's goal was the development of perceptual audio interfaces. As discussed, spatial audio, the sound that has been processed to provide the listener with direction and distance information of the sound source, is a perceptual interface as it is intuitive and does not require cognitive mediation on the part of the user. The theory of sound localization is helpful in understanding the implementations of spatial audio interfaces (Chapter 3 and Chapter 5).

2.3 Audio Based Systems

For a blind or low vision user, the auditory sense is one of the key senses to interact with the world. Audio feedback has been used in the past to provide spatial information to the users. This section reviews how the use of audio has been studied for the development of accessible spatial learning systems. The next sections specifically review the systems based on 1) Non-Speech audio 2) Speech based audio 3) virtual or spatialized audio

2.3.1 Non Speech Audio Interfaces

These interfaces make use of non-speech sounds such as sonification or even music to impart navigation information. (Kramer et al., 1999) describes sonification as "transformation of data relations into perceived relations in an acoustic signal for the purposes of facilitating communication or interpretations." One system making use of sonification for navigation is the System for Wearable Audio Navigation (SWAN) developed at the Georgia Institute of Technology (Wilson, Walker, Lindsay, Cambias, & Frank, 2007). This system sonifies the pertinent navigation related data, into non-speech sound beacons which guide the user to reach their destination. Here, the sound beacons appear to be coming from the direction of the next waypoint.

Music has also been used in a number of systems to guide users to their destinations. gpsTunes combined the functionality of a mobile Global Positioning System (GPS) with an MP3 player and directed users to their destination by continuously adaptive music based on their heading relative to the destination(Strachan, Eslambolchilar, & Murray-Smith, 2005). Another example of such music based navigation system is Ontrack (Warren, Jones, Jones, & Bainbridge, 2005) which allowed users to follow routes by keeping track of the volume and perceived direction of music.

2.3.2 Speech Based Audio Interfaces

The use of speech in communicating navigation information is very common. The main advantage of using speech to disseminate spatial information is the preciseness with which information can be presented. For example: "Walk 187 feet ahead, then turn right" is a very accurate instruction. Another advantage of using speech in navigation is the fact that, people already know how to process speech based information, and thus need not learn new instruction set to understand the directions. In fact, most in-car GPS based navigation systems employ this approach to guide the drivers to their destinations. These navigation systems give the users turn by turn navigation information at waypoints. The same idea has been implemented in pedestrian navigation systems, which allow the blind and low vision users to navigate in outdoor settings. One of the early examples of such a system is the Mobility of Blind and Elderly people Interacting with Computers - MoBIC system (Strothotte et al., 1996). This navigation system consisted of two interrelated components. The first component was MoBIC pre-journey system (MoPS) which allowed the users to plan journeys before starting the navigation. The second component called MoBIC outdoor system (MoODS) provided the users with navigation and

orientation assistance. The interface consisted of either cursor keys of the standard keyboard or a touch tablet with tactile grid or a map overlay for input. The output was either in Braille or synthetic speech. The user could place him anywhere on the map, and then freely explore the surroundings. The output consisted of verbal description of the place. Thus the user could learn about any obstacle, in the way or get orientation cues to align oneself on the map.

The Personal Guidance System (PGS) developed by Loomis and colleagues (Golledge, Marston, Loomis, & Klatzky, 2004; Loomis, 1985; Loomis, Golledge, & Klatzky, 1998) at the University of California at Santa Barbra (UCSB) uses differential GPS and compass data to guide the users to their destinations. It also employs a speech based interface to convey route information. A study to evaluate various modes for the PGS navigation system was conducted (Loomis, Marston, Golledge, & Klatzky, 2005), where five different auditory modes were tested. One of the modes namely virtual speech- in which the sound appeared to come from its actual direction, received highest subjective ratings, and shortest travel times. This thesis will review this research in detail in Chapter 3.

2.3.3 Virtual or Spatial Audio Based Interfaces

Thus far, this research has reviewed some of the previous research on providing spatial information by the use of speech or non-speech auditory displays. Some of the methods took advantage of the directional or spatial hearing capacity of humans, which almost always led to better performing audio displays.

In our daily lives, we hear the sounds coming from the exact direction of their source. For example, one can point almost exactly, with their eyes closed the location of a sound source, say a television. This effect can also be experienced, while walking in an open space and hearing the sound of a thunder. The human ability to localize the sound source is a complex phenomenon, and is already discussed in section 2.2.

Since spatial audio interfaces work at the direct perceptual levels, they may act as better interfaces than their non-spatialized counterparts. This effect has been studied extensively, in the field of human computer interaction. (Ho & Spence, 2005) studied the use of spatial audio based warning signals in a simulated driving task. The results from the series of experiments suggest that spatially predictive warning signals are most effective in capturing driver's attention. Another study by (Begault, 1994) compared the acquisition time for capturing visual targets in a flight simulator with the help of heads up auditory display. While the first condition was a standard one earpiece audio display, the second condition had a spatial audio display. The results from the study showed that pilots using the spatial audio displays were able to acquire the visual targets faster, than the pilots who used non-spatialized display.

2.4 Summary

This chapter reviewed some of the previous research on accessible interfaces that supports cognitive map development for blind and low vision users. It briefly described how touch alone has been used to impart spatial information in the past in section 2.1. It then reviewed the theory of human sound localization in section 2.2. This background is important for the implementation of spatial audio displays based on hand motion tracking

for indoor map learning using smartphones. Finally section 2.3, discussed how audio has

been used as a modality in the past to convey spatial information to the users.

CHAPTER 3

COMPARING THE EFFICACY OF AUDITORY MODES FOR LEARNING SPATIAL LAYOUTS

The previous chapters reviewed the effectiveness of audio as an alternate modality to vision for use in navigation systems. This chapter describes our first study, which presents and evaluates some new audio based perceptual interfaces for learning indoor spatial layouts. Section 3.1 reviews other comparable literature which has investigated perceptual audio interfaces and their efficacy in spatial learning and updating. I then describe three perceptually directed audio modes to learn indoor spatial layouts namely: Auditory Snapshot, Head motion triggered audio interface, and Hand motion triggered audio interface. I also describe spatial language, a non-perceptual mode and how it was used in this experiment as a benchmark to test against the other three novel interfaces. Section 3.2 describes in detail the methods employed in the study. Results are described in section 3.3. I discuss the implications of this research and provide conclusions in section 3.4.

3.1 Introduction

In this section I provide motivation to the work. I also describe previous research in this domain

3.1.1 Motivation and Related Work

One of the most common approaches to convey spatial information to end-users is through the use of spatial language. In fact, most in-car navigation systems employ this approach to guide drivers to their destination. For example "Drive 500 yards then, turn

right on Main Street". This method has also been used for many pedestrian based navigation systems (Heinroth & Buhler, 2008). More recently, this technique is also being employed in smartphone based pedestrian navigation systems; for example, in the Google Maps application for Android (Melanson, 2010). One of the challenges with spatial language is that it does not comprise a direct perceptual channel and requires cognitive mediation and working memory demands on the user's part because of the need to interpret metric, topological and other spatial information embedded in the linguistic signal (Klatzky, Marston, Giudice, Golledge, & Loomis, 2006). Since spatial language interfaces lack the intuitive component and require more working memory demands to be used effectively, it may not be the interface of choice especially for blind individuals in high cognitive effort or spatially demanding situations, such as when:

a) The blind user is engaging in a dynamic interaction with another person or the world, for example: The user is involved in a conversation with their friend or thinking about something they just passed while simultaneously navigating to their destination.

b) The blind user is navigating in surroundings which require a high level of attention to avoid obstacles, for example: the user is navigating in a mall.

c) The user has cognitive load introduced from something beyond the current spatial demands, for example: the user is about to give a keynote speech in a conference and is under time or pressure constraints or the user is at an airport and has to catch a flight.

As discussed in section 1.1, blind navigation inherently requires more cognitive mediation on the user's part to access and interpret environmental information as

compared to their sighted counterparts. As was described in (Giudice & Legge, 2008) there are various other differences that can make blind navigation a difficult task. Blind navigators need to learn to interpret non-visual sensory signals in order to traverse safely in the environment and avoid obstacles. They also need to constantly keep track of their current location and heading in the environment with respect to their final destination. These tasks require significant moment by moment problem solving and therefore require mental effort (Rieser, Guth, & Hill, 1986). In sum, blind navigation is an effortful endeavor requiring a lot of cognitive resources to accomplish safely and effectively. Therefore, there is a need for non-visual interfaces which require less cognitive mediation and which can convey more direct perceptual information.

Spatial audio or 3D audio is a technique in which the sound has been processed in such a way that the perceived azimuth of the sound source indicates the target direction, and the perceived intensity of the sound gives target distance (even though accurate distance perception has been found difficult to achieve (Zahorik, 2002)). Spatial audio works at a more direct perceptual level than spatial language and does not interfere with other competing cognitive tasks in situations described above. (Loomis et al., 1998) compared the guidance performance of this approach with a synthetic speech display. The results of the study indicated that the spatial audio based approach fared best in both user route guidance performances (less distance travelled, faster travel times) and user preferences. Another study by (Klatzky et al., 2006) compared the guiding performance of a spatial language interface with a spatial audio interface for following a route in the presence of additional cognitive load introduced through a vibrotactile N back task. While the guiding performance by the two modes did not differ significantly in the no-load

condition, it improved significantly in the spatial audio condition in the presence of cognitive load, whereas performance in the language condition was significantly worse in the presence of load.

Another interface known as the Haptic Pointer Interface (HPI), is described in (Loomis et al., 2005). When using this interface the user holds a rectangular stick, dubbed the pointer, with an electronic compass attached at the tip. Whenever the user points towards a landmark or a waypoint, within a tolerance range of 10°, they hear the information about that landmark or waypoint through speech or tone based audio as described in points 3 and 4 below. In the study, five different interfaces based on spatial audio and haptic pointing device were investigated. These interfaces were:

1) Virtual Speech: The spatial audio interface in which instructions to the next waypoint or landmark were given in the form of spatialized speech. The participant wore headphones, with an electronic compass (for head tracking) attached to the strap. The computer continuously gave synthesized speech indicating the distance left to the next waypoint. This distance (e.g., 32 feet) was uttered 72 times per minute. As the participant moved towards the target, the intensity of the sound increased, and the azimuth of target updated.

2) Virtual Tone: Again the participant wore headphones with the electronic compass on the strap. However, instead of hearing speech they heard tones, which were spatialized and thus appearing to come from the direction of the next waypoint. If the participant's head pointed within ten degrees on either side, they would hear an on course tone, which appeared five times each second with a duration of 160 milliseconds and a gap of 40

milliseconds between tones. If the relative bearing was more than ten degrees on either side, they would hear an off course tone which was a frequency swept tone and was played 2.3 times every second. Spatialized speech indicated the distance to the next waypoint and was provided every 8 seconds. Again, as the participant approached the target, the intensity of all the three sounds (On-course, Off-course and Speech) increased.

3) HPI tone: With the haptic pointer interface, the instructions were delivered in the form of a tone and were based on the pointing direction of the hand held stick. Whenever the user pointed the hand held pointer within 10 degrees of the direction of the next waypoint, they heard a sequence of beep tones. These beeps were the same as the on-course signal used in the previous interface. The sounds in this interface were solely based on proprioceptive information based on hand/arm orientation and did not include spatialized information. The auditory output was provided through a shoulder mounted speaker. Also, a non-spatialized speech sound indicated the amount of distance left every eight seconds. Whenever the relative bearing became more than 90 degrees, the user would hear a speech message indicating the correct bearing (e.g., 110 degrees left) rounded to the nearest 10 degrees.

4) HPI speech: This interface was similar to the HPI tone mode, except for the fact that the user now heard speech instead of tones. Thus whenever they were pointing within 90 degrees of the correct route, they would hear the word "straight" from the shoulder mounted speaker. Whenever their relative bearing was more than 10 degrees on either side, they heard left or right. When the bearing from their arm exceeded 90 degrees from the original bearing they heard their bearing in the form of speech (e.g., 100 degrees left).

5) Body pointing: This interface was similar to the HPI tone mode, except that the electronic compass was now mounted on the torso at the waist. Thus now instead of pointing their arms the users had to now point their body/torso to hear the beeps indicating that they were on course. Again, the sound was not spatialized and was delivered through a speaker mounted on the shoulder of the user.

The results of the study indicated that the virtual speech mode led to the shortest travel times and highest subjective ratings. Both of the spatialized audio displays (virtual speech and virtual tone) led to fastest travel times. According to the authors, the probable reason of the superiority of the spatial audio displays was perceptual localization. Whenever the participants reached a waypoint, the next waypoint was available immediately in the spatialized modes, as compared to the other modes where either their hand (HPI tone and HPI speech) or the body (Body pointing) had to be in line with the next waypoint.

This study mainly compared several perceptual interfaces in their route guiding performance. The participants were asked to follow a route, without the need of forming a global structure of the space in their minds (a cognitive map). While these perceptual interfaces are effective in guiding the users to their destination, their performance in helping form a cognitive map needs to be tested, which is one goal of this thesis work. It is important for the blind and low-vision users to form this global picture in their mind as it would help them to travel to the same destination again in the future, even without the aid of the navigation device. Having such a representation also supports more complex spatial behaviors like spatial inference, detours, shortcuts, and other cues which are important in daily life but are not possible from a simple route level representation. It is also important to test the efficacy of these interfaces in helping blind individuals learn

spatial structures such as rooms (e.g. office spaces, kitchens etc.) and lobbies; learning which forms an important part of their daily lives. The study described in this chapter investigated the efficacy of perceptual interfaces (namely 3D audio, hand motion triggered audio and head motion triggered audio) compared to non-perceptual interface (spatial language) in terms of cognitive map development.

Another goal of this experiment was to evaluate the spatial updating performance of the participants when using different audio interfaces to learn target arrays. Spatial updating refers to the ability of a moving person to mentally update the location of a target initially seen, heard or touched from a stationary point (Loomis, Lippa, Golledge, & Klatzky, 2002). Several studies in the past have demonstrated people's ability to update an internal representation of visual targets (Easton & Sholl, 1995), auditory targets (Ashmead, Davis, & Northington, 1995) and haptic targets (Holllns & Kelley, 1988). This ability to update our mental image of the objects is a very important phenomenon as it allows us to act on the objects even though our position might change from the learning location. As an example, imagine that a sighted person is in a kitchen working with a sharp knife. They stop to go and drink water from tap when suddenly the power goes off making the room completely dark. The person would still be able to keep track of the knife and avoid injury when they return back to their original position. This spatial updating phenomenon also occur at large scales and is crucial for navigating in large and complex environments to prevent getting lost. Blind and low-vision people are at a considerable disadvantage compared to their sighted counterparts because vision provides important cues regarding not only the motion of the user, but also about the global layout of the environment, both being important sources of information for effective spatial updating. However, as has

been shown in previous research (mentioned above), spatial updating is also possible and accurately performed when the targets have been acquired through non-visual modalities such as touch and audition.

Spatial updating performance can be evaluated by a number of different tasks. In one such spatial updating task, a user learns the position of an object through any modality (vision, touch, or sound), and then is asked to walk to the target with their eyes closed, from either the point where they learned the object, or from a different point from the learning perspective. Being able to walk to the target after learning it from a different point requires the person to update the spatial location of the target with respect to their new position. This can only be achieved if the mental image of the object has been updated with respect to the new location.

(Loomis et al., 2002) describes a study in which the participants learned a single target by means of spatial language or spatial audio. They then walked towards the target, either directly or indirectly. The authors propose an "image updating" model for this task. The first part of the task was "encoding" where the participants' formed an image of the object and its location in their mind. The next phase was updating, where the participants updated the location of the image with respect to their own position. They found that spatial updating of the verbally described targets (through spatial language) had the same characteristics as the updating of targets described through spatial audio, which suggests that spatial updating depends on the spatial image which in turn is independent of modality.

Spatial updating is said to be automatic if it occurs without explicit instruction or intention. Studies described in (R. F. Wang, 2004) found that spatial updating of real objects acquired through a perceptual channel (vision or hand pointing) was "automatic" as compared to the updating of objects acquired through verbal descriptions, a non-perceptual mode.

The current study, evaluated target learning and spatial updating performance with three perceptually directed interfaces (namely Auditory Snapshot, Head motion triggered interface and hand motion triggered interface, described in the next section) and spatial language, a non-perceptual interface. The purpose of the study was to extend previous research by evaluating cognitive map development with these "audio only" interfaces, with an ultimate aim of implementing them on handheld smartphone devices.

3.1.2 Audio Modes for this Study

This section, introduces the three perceptual interfaces by which the participants learned the experimental environments in the study and discusses how spatial language, a non-perceptual interface was used in this study.

1) Auditory Snapshot

This interface is based on spatial audio. The target name along with its distance appears to come from the direction of the target location. The auditory snapshot starts with the object on the left most part of the scene playing first, followed by the next object and so on. One snapshot is said to be completed when a person has heard all the object names along with the associated distances flowing from left to the right. Each utterance of the object name is coupled with the distance of the object in the current implementation.

Each target name is spoken twice. In contrast to traditional spatialized audio, which requires head motion, this condition does not require the user to move their head at all, as the signal itself is moving. That is, the azimuth information of the object is provided to the user as embedded in the spatial audio signal. As an example: Suppose, a scene (Fig. 3.1) consists of three objects, a table, a lamp, and a chair, placed at 8 feet, 4 feet and 4 feet at angles: +30°, - 60° and +60°, respectively. The auditory snapshot of the scene would sound like: Lamp 4 feet- Lamp 4 feet, Table 8 feet- Table 8 feet, and Chair 4 feet- Chair 4 feet, where each sounds to the listener as if the objects were placed at the respective angles in the real world.
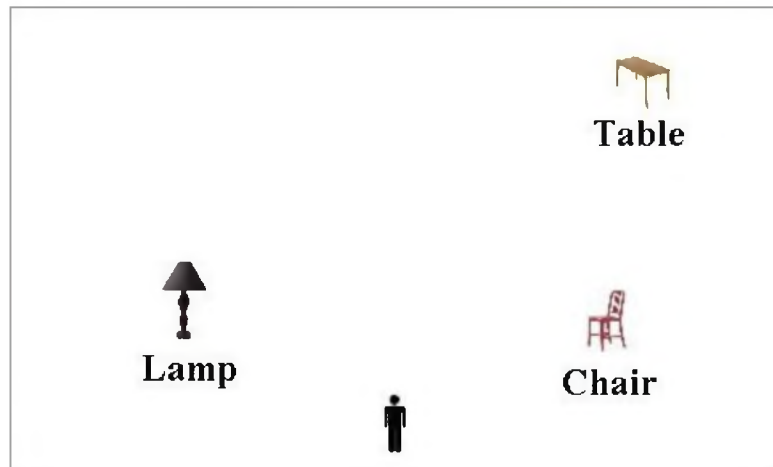
Figure 3.1 Sample Scene for Study 1

This interface is thus useful in providing a global view of the scene to the user. The user is able to learn the objects and their spatial locations in the scene in a natural way through the virtual soundscape of the scene as created by spatialized audio. Since spatial audio based interfaces are perceptual, this interface requires minimal cognitive mediation on the user's part to comprehend the object array. Since the user does not require moving any

part of their body, they can quickly learn objects in a room in an effortless manner. For example, imagine this audio mode implemented on a smartphone belonging to our friend Rita (Chapter 1).

"Rita decides to use this mode on an indoor room description application and learns the spatial layout of the Subway sandwich shop at the San-Diego convention center to locate the counter. In such a situation Rita (wearing a pair of headphones) would stand at the door of the restaurant and would start the application. The key objects in the restaurant start speaking their name and distance from the door starting from the object at the left, Wall 5 feet, Sandwich counter 10 feet, Soda fountain 12 feet, Cash Counter 7 feet. She now has an idea of the location of the key objects in the restaurant and has the requisite information about object relations to form a cognitive map of the space. She now heads to the sandwich counter with ease, orders her sandwich, gets a drink cup which she is able to fill herself from the fountain previously described, as she has updated her location within the cognitive map. Finally after getting her soda, she walks to the counter and pays the bill"

Even though this mode requires the object locations to be known in the database it is computationally simpler than tracking the user's position using external sensors which are expensive and still inaccurate. If this mode works efficiently in helping the cognitive map development of the users and is preferred by the users, it would be really a beneficial interface.

2) Head motion triggered mode

This auditory mode is based on the head motion of the user. An inertial head tracker is placed on the listener's head to track its motion. The target name and the distance in feet are uttered twice as the user faces an object of interest. For example, for the scene in Fig 3.1, the user would hear- Table 8 Feet- Table 8 Feet, as the user aligns their head to 30° on his right. This mode does not feature directional audio as does the previous interface, but is still a perceptual interface as it allows the user to learn spatial layouts based on their head orientation. Indeed, use of this additional proprioceptive cues derived from head movement can be very effective, as described in the earlier study (Loomis et al., 2005).

The user starts exploring the spatial layout by orienting their head to the extreme left of the space (Fig 3.2). A voice "Start" informs the user that he is in the initial position. He then is instructed to slowly sweep his head, from left to right, keeping the lower part of the body fixed. As the user comes across an object of interest, its name and distance are uttered twice. Eventually, as the user reaches the right end of the space, they hear a sound "Stop", informing them, that they have reached the right most extent of the space. This completes one exposure to the room. The user then moves their head to orient back to the start position (Fig 3.2). While reorienting, the user does not hear any sounds, until the initial state is reached. As the user reaches the initial starting point, a voice utters "Start" again; to let the user know that they have once again reached the initial position and can perform a second sweep of the space.

Figure 3.2 Head Motion Triggered Mode

This interface, tries to simulate human vision with audition. As persons localize an object by facing directly towards it (though we also have the ability to move our eye to expand our field of view), we imagined this interface to help blind individuals localize objects by facing them directly. Since there was no directional audio (the sound output was delivered equally to both ears), the localization of the objects in this interface is based purely on the orientation of the head. The users thus form the spatial image of the scene perceptually by remembering the target name and distance coupled with the bearing of their head. Rita would learn the cognitive map of the sandwich shop by moving her head. We assume that Rita by some means is able to convey her head motion information to the application (e.g. headphones etc. See chapter 4 for a review on head tracking technology).

"Rita goes to the shop and plugs in her head tracking headphones to her smartphone. She now rotates her head to the left to learn the location of the wall; next she learns the location of the sandwich bar and the drink fountain. Finally she learns the location of the cash counter."

Rita again has access to the cognitive map of the shop, but this time it is built up from auditory messages derived through the head motion of when she is directly facing the objects of interest.

3) Hand motion triggered mode

The hand motion triggered mode is based on the movement of the user's arm. This interface was implemented by placing the orientation tracking device on a stick (Fig. 3.3). A user can point this device in any direction. As the user points towards the direction of an object of interest, the object name and distance are uttered. For example, if the user's arm is at 60° they would hear Chair 4 feet- Chair 4 feet denoting that they are currently pointing towards the chair.
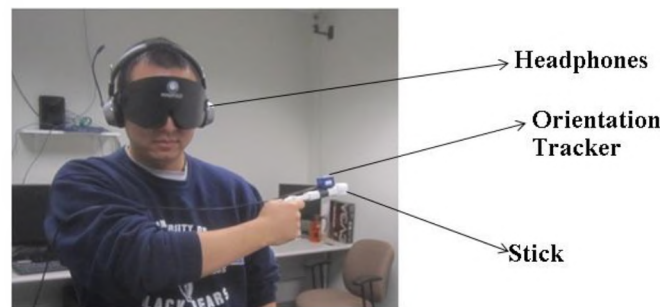
Figure 3.3 Hand Motion Triggered Mode

As in the previous modes, the user starts exploring the spatial layout by pointing the device to the left-most direction. A voice "Start" conveys to the user that they can now start exploring the objects in the room by moving their arm from left to right and stopping each time they hear an object to confirm its location. As the user points to an object of interest, its name and distance are spoken. When the user points their arm to the extreme right, they have completed a sweep and have learned all the objects in the current scene.

42

At this point, they hear a voice announcing "Stop". They can now move their arm back in the left direction, until they hear "Start" again, and can thus repeat the learning of the room.

This mode can be implemented on current off the shelf smartphones. The users would have the ability to scan a room or any other spatial layout by pointing to various objects. For example Rita would now learn the cognitive map of the shop in the same way as the previous mode, except that she would now use her smartphone as a pointing device instead of her head.

This mode is easier to implement on smartphones than the head motion triggered interface described earlier, because of the ease of tracking of user's arm as compared to their head without the need of extra sensors. This mode is also aesthetically more preferable as it does not need the user to wear any extra equipment on their head.

4) Spatial Language mode

A spatial language interface is implemented to describe a non-visual mode to support spatial learning and navigation by the use of verbal descriptions of spaces. This is a standard way to support non-visual spatial learning, behavior, and cognitive map development. The efficacy of spatial language in supporting these tasks and helping individuals build correct spatial relationships between targets has been widely studied in previous research (Ferguson & Hegarty, 1994; Giudice, Bakdash, Legge, & Roy, 2010; Kulhavy, Schwartz, & Shaha, 1983).

In the current study, we provided information about target names and distances in terms of clock face angles. Standing at an origin position, the participant heard the digit

43

indicating the clock angle of the target followed by the object label and its distance. This spatial language scheme is similar to the study described in (Klatzky, Lippa, Loomis, & Golledge, 2002). The participants faced 12 O'clock while hearing the spatial language utterance. Thus 3 O' clock meant 90° on the right and 10 O'clock meant 60° towards the left.

As was done in the other conditions, each utterance of the scene started from the left and swept rightward across the object array. As an example, the scene in figure 3.1 sounded as:

"10 O' clock 4 feet Lamp, 1 O' clock 8 feet Table, 2 O' clock 4 feet Chair"

The participant heard two utterances of the scene. This exposure lasted 19 seconds which was consistent with the time exposure of the previous modes. Let us continue with our persona:

"Rita starts the spatial language mode on her smartphone and hears the following description about the room. 9 O clock 5 feet Wall, 12 O clock 10 feet sandwich counter, 1 O clock 12 feet Soda fountain, 3 O clock 7 feet cash counter"

She can then perform the tasks we described before but to do so, she will have first needed to convert the cognitively mediated, non-perceptual verbal messages into a spatial form.

3.2 Method

This section describes the methodology for the study which compared the spatial updating performance of the participants with the modes: Auditory Snapshot, Head

44

motion triggered mode, Hand motion triggered mode and the Spatial Language mode. The study was approved by the University of Maine's Institutional review Board (IRB) and took about 1.5 hours to complete for each participant.

3.2.1 Participants

Sixteen sighted University of Maine students (8 female, mean age= 24.9 years) participated voluntarily in the study and signed informed consent forms. All the participants reported normal hearing and were monetarily compensated for their time and effort.

Sighted participants have a different spatial experience as compared to their blind counterparts, owing to the use of a different modality (vision) in learning and exploring the surrounding environment. However, for the current study, we considered only sighted participants (wearing blindfolds) as they are more readily recruited and evidence from previous studies suggests that there is little difference in learning between blindfolded-sighted and blind participants through non-visual modalities as they are equally accessible to both groups (Giudice, Betty, & Loomis, 2011; Loomis et al., 2002; Walker & Mauney, 2010). This study served as a preliminary indicator for the success of the investigated interfaces.

3.2.2 Apparatus

This study was conducted in a lab room having dimensions 4.26 m by 5.71 m. The participants were blindfolded for the entire experiment (Mindfold, Inc. Tucson, AZ). The participants wore Creative HS-1200 (Creative Technology Ltd. USA) wireless headphones during the study to listen to instructions and stimuli. An inertia cube

(Intersense , LLC Billerica, Massachusetts) was attached to the Headphones/Stick to determine orientation of the user's Head/Hand during the head motion and hand motion triggered conditions The inertia cube is a head-tracking device developed by Intersense Inc. and is based on nine miniature inertial sensing elements and uses Kalman Filters to provide head orientation with an accuracy of 1°.

A battery powered Light Emitting Diode (LED) was placed on the wireless headphones and allowed us to track the precise position of the participant, using an optical Precision Position Tracker (PPT) system (WorldViz inc., Santa Barbra, CA). This LED tracker also allowed us to measure the virtual positions of the targets in order to generate the Virtual auditory Environments (VAE).

The Virtual Auditory Environments were generated using Vizard 3.13(WorldViz inc., Santa Barbra, CA), using Python 2.4 (Python Software Foundation, 2012). The participants recorded their responses using a Nintendo Wii (Nintendo Inc.) remote ("Wiimote"). The A button of the wiimote was termed as "Start" and the B button was termed as "Stop" and was used to record the current state (Position and Orientation) of the participant. The Virtual Environment was generated and the study controlled through a computer (Intel i7 2.65 GHz processor). The wiimote and the headphones were connected wirelessly to the controlling station with Bluetooth.

3.2.3 Stimuli

The target stimuli were names of objects found commonly in households and in office spaces. They were selected from the list of common stimuli as discussed in (Snodgrass & Vanderwart, 1980). The complete list of target names is given in Table 3.1.

S.no	Target Name
1	Box
2	Book
3	Lamp
4	Bed
5	Ball
6	Guitar
7	Dog
8	Table
9	Chair
10	Couch
11	Cat
12	Desk

Table 3.1 Target Names for Experiment 1

In the Auditory Snapshot, Head motion triggered, and Hand motion triggered conditions, the stimulus consisted of the target name followed by the distance. For example, Table 4 feet. In the spatial language mode, the stimulus consisted of the direction of the target, in terms of clock angle followed by the target name and distance. Therefore, a table at 8 feet ahead located at 60°, was heard as 2 O Clock 8 feet Table.

The stimuli were recorded as Wave files using the online AT&T Text to Speech Converter(AT & T Labs, Inc., 2010) using the US English Female voice Crystal. The

stimuli were then edited in Audacity ("Audacity," 2010) to ensure consistent duration and waveform.

The polar coordinates of the target locations (Fig. 3.4) were $\pm90°/1.21$ m, $\pm90°/2.43$ m, $\pm60°/1.21$ m, $\pm60°/2.43$ m, $\pm30°/1.21$ m, and $\pm30°/2.43$ m. There were two drop off points located on each side of the origin at a distance of 0.5 m each, labeled as A and B in Figure 3.4. Across participants the target labels were counterbalanced.
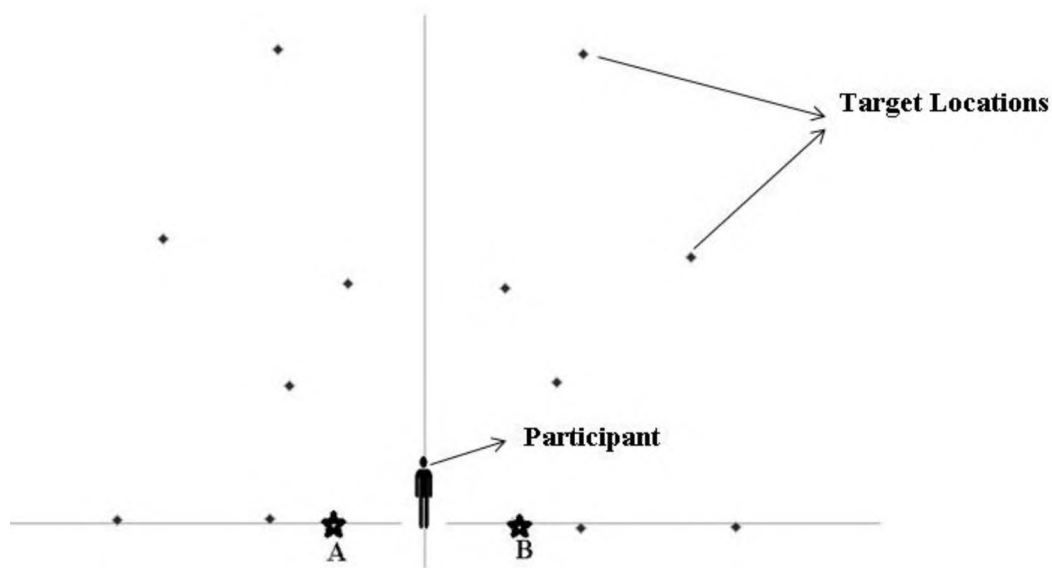
Figure 3.4 Top View of Target Locations for Experiment 1

3.2.4 Procedure

The design of the study was completely within subjects, with each participant being exposed to each of the four learning modes. The order of the three new learning modes was counterbalanced. As an exception, the Spatial Language condition was always presented last, as it provided the angular information about the targets directly (e.g., 60 degrees), whereas the other conditions used perceptual cues to convey the azimuth. To

48

avoid this leading to any confounds, the spatial language was run last. The overall procedure for the study consisted of five phases. The study began with familiarization of the equipment to the participants. They were then given an opportunity to walk to distances 4 feet and 8 feet with their eyes closed, and were given corrective feedback on their performance. Once the participants were comfortable in using the equipment and with the blind walking task, we started the experimental trials.

a) Learning Phase I

The first phase of the study was a multi-trial learning phase. The participants learned 3 target objects selected from the pool of stimuli (3.2.3), depending upon the learning mode. They were exposed to the same target array twice. In the auditory snapshot mode, the participant heard the 3 target names and their distances, with the sound appearing to come from the direction of the object. A snapshot was completed when the participant heard the name of all the three targets from left to right. They were exposed to two snapshots in a single trial. Similarly, in the head motion and hand motion conditions the participants moved their head/hand, in the horizontal left—right plane until they learned the three objects. They again learned the target array twice. Finally, in the spatial language mode, the participants learnt the array via speech output, with the direction being described in terms of clock angles. They were given two exposures to the target array.

b) Learning Criterion

To ensure that the participant learned the array with sufficient accuracy, a learning criterion phase was introduced. Participants started this phase by orienting themselves at

49

an initial heading of 0° (Laboratory 'North' heading). The participant's heading was measured with the help of an inertia cube placed on their head. To orient to the starting orientation, the participant held a wiimote that vibrated whenever they were facing within $\pm 2°$ of the heading of the start position. This helped the participant to orient to the initial heading after each response.

The computer randomly selected a target name from the three targets that the participant had learned from phase a) and spoke its name through computer-generated speech. The participant was then asked to turn to this randomly selected target (E.g., "Turn to the Table"). They would then orient themselves to face the target object. They pressed the "Stop" button on the wiimote to indicate completion of their response. This was repeated for all the three targets in a given trial. If the absolute angle error, in the three trials was less or equal to 15°, the participant was assumed to have successfully learned the target array. If they did not pass the learning criterion, they were asked to re-learn the same array using the same mode. After this re-learning period, they were once again asked to perform the learning criterion phase. This process was repeated until either the participant passed the criterion successfully or they had performed the learn-test criterion sequence for six times, whichever was first.

c) Walking Phase I

Once the participant passed the learning criterion, they entered the next phase of the study which was the first walking phase. In this phase, the participant stood at the tactile landmark termed as the "Origin". The experimenter asked them to sidestep to either the left drop off point (A) or the right drop off point (B). They were then asked to press the

"Start" button on the wiimote. As the participant pressed the start button, they heard instructions to walk to a randomly selected target from that location) For Example "Walk to the Chair". The participant would then start walking from the drop off point directly towards the target. Instead of walking directly to the target, we asked the participants to walk from drop off points. To be able to accurately walk to the target from a new location, the participants would need to have formed an accurate "spatial image" of the target array. This spatial image would then allow them to mentally calculate updated angles and distances to walk to the targets from the drop off points.

When they reached the location where they thought the target was, they pressed the "Finish" button. The experimenter then guided them back to the Origin. The participant realigned themselves to face north, with the help of the tactile landmark.

We recorded the absolute position of the participant when they thought they had reached a target (indicated by pressing the "Finish" button). We also recorded the time taken by the participant to walk from the drop off point to the target.

d) Learning Phase II (Re-exposure)

This phase was similar to the first learning phase. The only difference between learning phase I and II was that in Learning phase II, the participant was exposed to the target only once as opposed to being exposed to the targets twice in learning phase I. This phase was provided to the participants to allow them to refresh their mental model for the scene one more time after walking from the first drop off point.

e) Walking Phase II

This phase was similar to Walking phase I, and the participants walked to the same targets as learned in phase a). The difference was that this time the participant walked to the targets from the remaining drop-off point (A or B) not used in Walking Phase I. Thus the participants walked to each target location from two points equidistant from the origin and located on either side of it. This ensured that there was no directional bias in responses to the targets located on either side of the origin.

Again the location of the response and the response times for marking the response were recorded.

3.3 Results

We analyzed the performance with the four modes in the study mainly for:

1) Number of trials required to reach the learning criterion

2) Distance Errors

3) Azimuth Errors

4) Target to Response Distance Errors

5) Response Times

6) User Ratings

3.3.1 Number of Trials to Reach Criterion

As discussed in 3.2.4, a learning criterion ensured that the participant accurately learned the target array before performing the updating tasks requiring blind walking to the targets. The participants passed the learning criterion if their average pointing error for

the targets was less than or equal to 15 degrees or if they learned the target array 6 times whichever was first.

A repeated measures Analysis of Variance (ANOVA) was conducted on the number of trials needed to achieve the learning criterion using the variable of modality. The effect of modality did not reach significance, $F(3,15)= 1.007$, $p=0.396$, $\eta^2_p = 0.048$. The means and standard deviations for the number of trials required to reach criterion is given in Table 3.2.

S. No.	Condition	Mean	Standard Deviation
1.	Auditory Snapshot	2.19	1.601
2.	Hand Motion Triggered	1.63	0.806
3.	Head Motion Triggered	2.00	0.730
4.	Spatial Language	2.31	1.401

Table 3.2 Number of Trials to Reach Criterion

3.3.2 Distance Error

The distance error was calculated as the difference between the distance of the target from the origin to the distance between response and origin. The distance errors were analyzed in two ways: a) signed and b) unsigned. In both cases there was no significant effect of the two drop off points. So the results from the two start points were collapsed while calculating means. Outliers, defined here as values greater than 2.5 Standard deviations from the mean, were removed (n=8, 2.08%) and were replaced with the mean value prior to averaging.

53

1) Signed Distance Error

This was calculated as the difference in distance between the target and origin and the distance between response and origin. A repeated measures ANOVA with factors of interface mode and distance showed that there was no modality effect, $F(3,15)= 0.062$, $p=0.980$, $\eta^2_p = 0.0$. However, there was an effect of the distance of target from the origin on error, $F(3,15)= 313.529$, $p< 0.01$, $\eta^2_p = 0.456$. Subsequent t-tests suggest the participants were better in walking to the near targets located at 4 feet ($M=0.298$, $SD=0.400$) than to the far targets located at 8 feet ($M=-0.464$, $SD= 0.437$), $t(184)= 16.808$, $p<0.01$.

The signs of the means suggest that while the participants overestimated the near targets located at 4 feet, they generally under estimated the far targets located at 8 feet.

S. No.	Condition	Mean Signed Distance Error(m)	Standard Deviation
1.	Auditory Snapshot	-0.073	0.575
2.	Hand Motion Triggered	-0.097	0.579
3.	Head Motion Triggered	-0.073	0.495
4.	Spatial Language	-0.062	0.615

Table 3.3 Signed Distance Error

2) Absolute Distance Error

The absolute error was calculated as the absolute value of the difference in distance between the target and origin and the distance between response and origin. Again no significant effect of modality on absolute distance errors was found, $F(3,15)= 1.055$,

p=0.368, η^2_p = 0.008. There was again a significant effect of the distance of the targets, F(3,15)= 12.809, p< 0.01, η^2_p = 0.033. As in the previous findings, the participants walked to the 4 feet targets (M=0.400, SD= 0.299) better than the 8 feet targets (M=0.523, SD=0.364).

S. No.	Condition	Mean Abs. Distance Error (m)	Standard Deviation
1.	Auditory Snapshot	0.469	0.337
2.	Hand Motion Triggered	0.465	0.356
3.	Head Motion Triggered	0.412	0.282
4.	Spatial Language	0.498	0.363

Table 3.4 Absolute Distance Error

3.3.3 Angle Error

The angle between origin and target and origin and response was calculated using the circular statistic method (Mahan, 1991). The angle error was calculated as the difference in angle between the origin-target and origin-response vectors. We analyzed both signed and absolute angle errors. Again there was no significant difference between the angle errors from the two drop off points.

    1)   Signed Angle Error

An Analysis of variance on signed angle error with the variable of interface modality showed no significant effect of the mode, F(3,15)= 1.908, p= 0.128, η^2_p = 0.020. The sign of the Grand mean (M= +3.937°, SE= 0.967°) was positive.

S. No.	Condition	Mean Signed Angle Error (degrees)	Standard Deviation
1.	Auditory Snapshot	4.925	19.769
2.	Hand Motion Triggered	5.318	20.903
3.	Head Motion Triggered	10.103	32.041
4.	Spatial Language	2.089535	18.73103

Table 3.5 Signed Angle Error

2) Absolute Angle Error

The absolute angle error for each trial was calculated as the absolute value of the signed angle error. ANOVA results showed that there was an effect of modality on the absolute angle errors, $F(3,15)= 9.697$, $p<0.01$, $\eta^2_p = 0.073$. Subsequent pairwise t tests revealed that the Auditory Snapshot mode (M= 13.76°, SD= 10.78°) fared better than the Head motion mode (M=21.80°, SD= 16.70°), $t(95) = -4.366$, $p<0.01$. The Hand motion mode (M= 16.30°, SD= 12.67°) also led to lower absolute angle errors than head tracked condition, $t(95)= 2.763$, $p=0.007$. The spatial language mode (M=12.51°, SD= 10.22°), also led to better performance than the head motion mode, $t(96)= 4.901$, $p<0.01$. The absolute angle errors for the four modes are depicted in the graph.
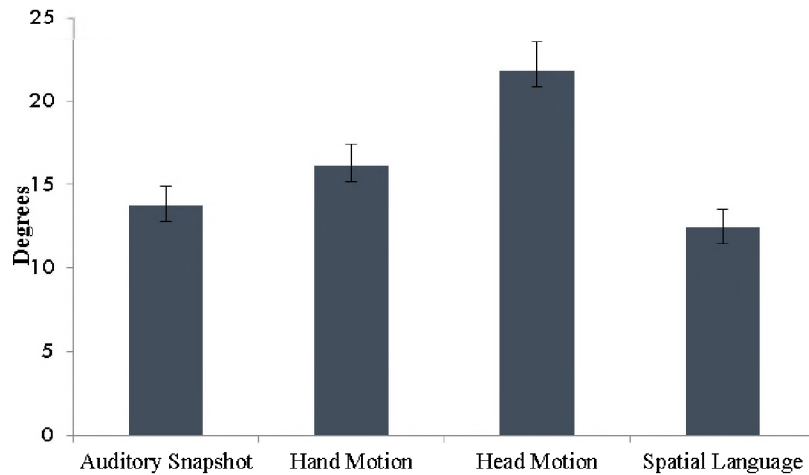
Figure 3.5 Absolute Angle Errors for Experiment 1

3.3.4 Target to Response Distance

The Target to Response distance for the walking phase was calculated as the distance between the target location and the response position for each target and response pair This measure gives us an estimate on how near the participants responses were to the targets and is always positive as it represents the Euclidian distance between the two points. Again the responses from left and right drop-off points were collapsed.

A within subjects ANOVA was conducted to compare the effect of interface modes on the Target to Response Distance. No significant differences were found in the target to response distances for the four modes, $F(3,15)= 2.272$, $p=0.08$, $\eta^2_p = 0.018$.
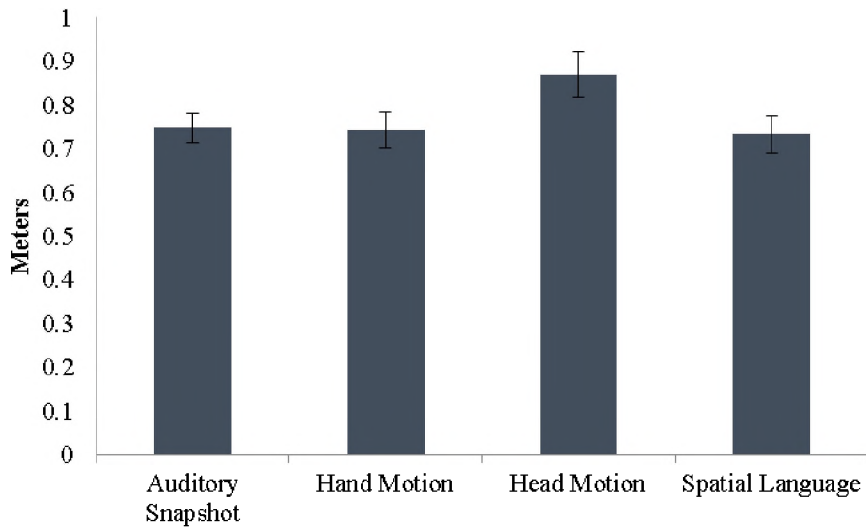
Figure 3.6 Target to Response Error Graph

3.3.5 Response Time

The response time for the walking phase was calculated as the time taken by the participant to walk from the drop-off point to the target. The response time is an indication of the cognitive load on the participant in marking the response. Higher response times mean higher mental effort in remembering the target locations. The mean values for the mean response times for the four modes are depicted in the graph in Fig. 3.7.

A within subjects ANOVA was conducted to compare the effect of presentation modes on the Response Times. No significant differences were found in the Response times for the four modes, $F(3,15)= 2.112$, $p=0.098$, $\eta^2_p = 0.017$.
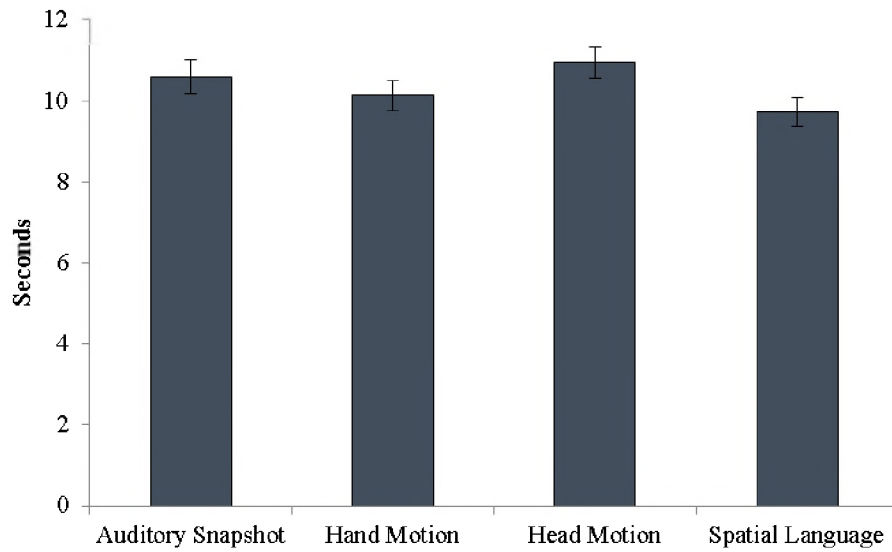
Figure 3.7 Response Times for Experiment 1

3.3.6 Preference Ratings

After completion of all the phases of the study the participants were asked to rank the modes in order of their preference (most preferred=1 and least preferred =4). The mean values of user ratings are depicted in Fig 3.8.

An analysis of variance showed no effect of modality on preference level, $F_{(3,15)}= 1.622$, $p=0.194$, $\eta^2_p = 0.075$.
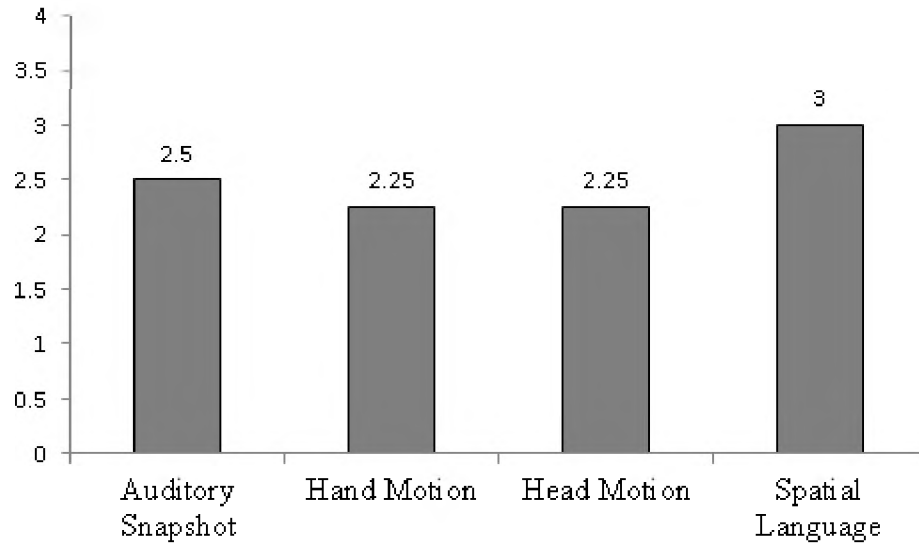
Figure 3.8 Mean Preference Ratings

3.4 Discussion and Conclusion

In this chapter, I started by reviewing the factors that lead to additional cognitive demands for blind navigation. I then described how perceptual spatial interfaces such as spatial audio and hand/head motion triggered modes, with their capability to convey spatial information with minimal cognitive effort, can be used to aid spatial behaviors like updating and cognitive map development in a non-visual manner. I conducted an experiment that compared the spatial updating performance of the three perceptual modes with each other and with spatial language, a non-perceptual mode which requires cognitive mediation on the user's behalf to convey spatial information.

No significant differences were found in learning rates, response times and distance errors in the three perceptual interfaces described in this chapter. However, participants incurred significantly more absolute angle errors with the head motion triggered interface as compared to the other two perceptual interfaces namely auditory snapshot and the

hand motion triggered interface. This means that the participants had difficulty in remembering the azimuth of the targets when they learned it through the motion of their heads as compared to the other two perceptual interfaces. This result is an interesting finding as it implies participants formed more accurate spatial representations from arm movement triggered audio and spatial audio than from head movement triggered audio. Another interesting result is the functional equivalence of spatial images formed by hand motion triggered audio and auditory snapshot. While I implemented auditory snapshot on a smartphone device (Chapter 5), I left further investigation of hand triggered audio interface on a smartphone for future research because of the inability of the current sensors to provide correct orientation (See chapter 6 for further discussion).

While no significant differences between the perceptual and non-perceptual modes in learning rates, response times and distance errors were observed, to my surprise, I found that participants incurred significantly less absolute angle error with spatial language as compared to the head motion triggered mode, one of our perceptual modes. As described in section 3.2, I used spatial language as a control condition and therefore administered it at the last for each trial. In retrospection, we realize that this procedural decision may well have led to an artificially elevated level of spatial learning performance by the participants as compared to the head motion triggered interface. Also, since our experiment did not involve any additional cognitive load for the participants, the spatial updating performance with spatial language may not have suffered due to cognitive arbitration of the non-perceptual mode.

CHAPTER 4

COMPARING HEAD-MOTION AND HAND-MOTION BASED SPATIAL

AUDIO INTERFACES

The last chapter explores a new perceptual audio interface which we called an "auditory snapshot", which is based on three dimensional or spatial audio. The interface proved to be effective in imparting spatial information about scenes to the participants. As discussed in previous chapters, the spatial information system should employ off the shelf smartphone devices. Chapter 2 discussed the importance of tracking head motion of the user to achieve better spatialization of sound. This chapter explores the issue of unavailability of head motion tracking mechanisms in smartphones and the approach to replace it with hand motion tracking. This chapter also describes spatial updating performance of blindfolded participants after learning targets with spatial audio generated by the traditional approach (with head tracking), our novel approach (with hand tracking) and the visual approach. The structure of this chapter is as follows: section 4.1 provides an introduction and some previous work on this issue. Section 4.2 describes a study which was used to assess the efficacy of our new approach compared with the traditional approach and with the baseline of vision. Section 4.3 provides the results from the study. Finally section 4.4 presents the implications of this research and provides some conclusions and broader contexts for the results.

4.1 Introduction

As described in section 2.2, head motion plays a crucial role in helping to determine the direction and distance of a sound source in the real world by removing front back confusions and improving overall localization accuracy. Head motion also plays a crucial role in the generation of virtual spatialized audio by modifying the audio signals from the rendering machine in accordance to the user's head motion. Section, 4.1.1 reviews some of the related work that underlines the importance of head tracking in 3D audio applications. Section 4.1.2 reviews some of the technologies and methods that have been used in the past to accomplish the feat of tracking a user's head.

4.1.1 Related Work

Some of the earliest research in this domain was done by Wallach in 1938. The author defined the angle between the direction of the sound source and the aural axis as the "lateral" angle. This angle describes the cone of confusion as all the points on the surface would have the same angular measurements. In his experiment, the author proved that the perceived location of the auditory event was independent of the sound source's actual position as long as the changes in the lateral angles were presented in line with the head motion of the user (Wallach, 1940).

(Wenzel, 1996) demonstrated that allowing head motion tracking significantly improved the localization performance of humans, even when non-individual general HRTFs (section 2.2) were used during binaural synthesis. (Sandvad, 1996) used individual HRTFs for binaural synthesis, and when head tracking was enabled the localization results were only slightly worse than real life performance.

(Thurlow et al., 1967) evaluated the impact of induced head motion on sound localization. The authors considered four different modes of induced head motion, namely: rotation, pivot, rotation-pivot, and no head motion. While in the rotation mode, the participants turned their head left and right, in the pivot mode they moved their head in such a way that one ear was higher than the other. The rotation-pivot mode combined the previous two modes. In the no-head motion mode, head motion was not allowed. The results indicated that the rotation and rotation-pivot modes led to better localization performance than pivot and no motion. These results provide further evidence of the importance of horizontal (left to right or right to left) head rotation in sound localization.

In a study described in Perrett & Noble (1997), the authors measured the accuracy with which the participants localized a sound source (a 2 kHz low pass filtered noise burst) with or without head motion. In the without head motion condition, participants made a number of localization errors. However, when allowed to move their head or turn their head to 45°, the number of errors decreased significantly. Similar results were obtained in a study described in F.L. Wightman & Kistler (1999), where the participants were asked to indicate the apparent positions of virtual and real sound sources in the presence or absence of head motion. The authors found that while the front-back confusions were common in the restricted head motion mode, they almost disappeared when head motion was allowed. In yet another study described in Wu, Duh, Ouhyoung, & Wu (1997), the authors found that the ability of the participants to localize sound sources increased by more than 90% in the presence of head motion tracking as compared to the absence of this facility.

Besides helping in sound localization and reducing front back confusions, head motion tracking allows us to move the sound scene along with the user's head. For example, a sound located at 45° right of the user would sound as if at 0° ahead as the user turns his head to 45° towards the right. This is particularly important in non-visual spatial learning systems where the user might not have visual access to the target to confirm its location. It is for this reason that most navigation systems based on spatial audio have some provision for head tracking of the user. For example: The personal guidance system (Loomis, 1985; Loomis et al., 1998, 2005), The Swan project (Wilson et al., 2007),The LISTEN project (Warusfel & Eckel, 2004), 3DAAR (Sundareswaran et al., 2003), Wearable Augmented Reality TestBed for Navigation (Behringer, Tam, McGee, Sundareswaran, & Vassiliou, 2000) etc.

The next section, explores how head tracking can be achieved by various means for the purposes of improving sound localization. The section discusses the prospect of tracking the user's hand instead of the head, which is a necessary change for implementing spatial audio applications on smartphone devices.

4.1.2 Methods to Track Head Motion

Some of the current techniques for head motion tracking are summarized in (Rolland, Baillot, & Davis, 2001). The authors classify head motion tracking techniques as falling into six categories. Table 4.1 summarizes the salient features of the technologies presented by Rolland et al. along with their pros and cons below.

Traditional Approaches

1) Time of Flight techniques

The systems based on these techniques rely on the measure of distances of features attached on one side to a reference and on the other side to a moving target. These distances are determined by time of propagation of ultrasound signals.

Physical Phenomenon	Acoustic Pulse propagation
Orientation Accuracy	0.1-0.6 degrees
Advantages	Small, Light
Disadvantages	High Cost, Sensitive to heat temperature and pressure
Examples	Intersense Cube, Honeywell Hemet tracking system.

Table 4.1 Summary of Time of Flight Techniques

2) Spatial Scan techniques

Spatial Scan trackers are based on the analysis of two dimensional projections of image features using optical cameras.

Physical Phenomenon	Spatial Scan
Orientation Accuracy	1/2800 of the cameras field of view
Advantages	High Update rate
Disadvantages	Sensitive to optical noise, High cost, use of cameras make it impractical for use in portable systems.
Examples	Multitrac from Simulis

Table 4.2 Summary of Spatial Scan Techniques

3) Mechanical linkage techniques

They make use of mechanical parts to calculate the linkage angle between a fixed reference and the user

Physical Phenomenon	Mechanical Linkages
Orientation Accuracy	0.15- 1 degree
Advantages	High Update rate, High accuracy, no effect of environmental noise
Disadvantages	Limitation of motion
Examples	Argonne Remote Manipulator

Table 4.3 Summary of Mechanical Linkage Techniques

4) Phase difference method

These techniques measure the relative phase of an incoming signal and compare it to a signal in a fixed reference system.

Physical Phenomenon	Phase Difference sensing
Orientation Accuracy	variable
Advantages	Less susceptible to noise
Disadvantages	Possible ambiguity in results
Examples	Southerland Head mounted display

Table 4.4 Summary of Phase Difference Method

5) Direct field sensing technique

This method utilizes either magnetic or gravitational fields to calculate the orientation

Physical Phenomenon	Magnetic/ Gravitation fields
Orientation Accuracy	variable
Advantages	Small, Inexpensive
Disadvantages	Highly Susceptible to noise
Examples	Honeywell, Flock of Birds

Table 4.5 Summary of Direct Field Sensing Methods

6) Hybrid Systems

These techniques employ a multitude of different technologies such as direct field sensing, time of flight techniques, mechanical linkages etc. to calculate orientation of the user's head.

Physical Phenomenon	Direct Field Sensing, Inertia etc.
Orientation Accuracy	variable
Advantages	Compact and accurate
Disadvantages	High Cost, Occlusion sensitive and other problems related to the physical phenomenon used
Examples	Inside Out optical tracking system which used three gyroscopes and three accelerometers for head tracking (Azuma, 1995)

Table 4.6 Summary of Hybrid Systems

The above techniques have been used to track user's head motion in laboratories for a long time. However, these techniques suffer from a number of limitations, which make their use in portable navigation systems difficult.

a) High Cost:

Almost all of the techniques described above are costly because they rely on expensive highly specialized equipment. This makes it difficult to install a head

tracker in low cost systems, as is my ultimate goal for the application of my thesis work.

b) Size:

The orientation sensors based on mechanical rotations are bulky and heavy, making their use limited for implementation in systems used for real world navigation

c) Need for extra setup:

The spatial scan systems require installation of additional optical cameras which make their use almost impossible in portable navigation systems as requiring expensive infrastructure modifications for the system to work is impractical for any widespread implementation.

This thesis research project aims to implement a system with off the shelf smartphone devices. Use of these traditional technologies makes it almost impossible to develop an inexpensive navigation system implemented on smartphone devices. Alternative tracking approaches to these traditional techniques are thus reviewed in the next section.

Headphone-based Tracking

One approach to obtaining real-time head tracking data is the use of sensors in the headphones worn by the user, which can then be plugged into the mobile device. Three different sensor technologies can be used to obtain orientation information in such a setup, namely: Acceleration Sensors, Magnetic field sensors and gyroscope sensors. (Christoph, 2007) provides a comparison between these three sensors for use in headphone based tracking.

1. Accelerometer Sensor

An accelerometer measures the linear acceleration of the object to which it is attached. It is a single degree of freedom device which consists of three primary components: 1) a mass, a spring, and a supporting structure with damping properties. In the most common implementation, a mass is mounted on a piezoelectric crystal (a piezoelectric crystal generates electric charge when pressure is applied). When the object on which the sensor is attached is moved, it creates a pressure. The resulting force can be obtained by measuring the voltage on the sides of the crystal. This force is proportional to the acceleration of the body (Force=Mass times acceleration). A double integral of this acceleration yields the current position, assuming the initial position and speed of the body is known (Yazdi, Ayazi, & Najafi, 1998).

The main advantages of this sensor are that it is lightweight and requires no external reference. Using accelerometer only for measuring head rotation however leads to many problems such as:

a) The sensor needs to be calibrated before use.

b) When used in personal navigation devices acceleration due to translation and the earth's gravity may be quite large, causing false rotation values.

c) The rotation angle value is based on the double integration of the differences in the sensor values. Small errors in differences can have a great effect on the calculated angle.

2. Magnetic Sensors

The magnetic field of the earth, defined as the field created by the imaginary magnetic force running from Magnetic North to the South Pole, can be used as an external reference. This approach has been used for thousands of years in the form of magnetic compasses for use in navigation. Usually the magnetic sensors measure two components of the earth's magnetic field, $H_x(t)$ and $H_y(t)$. The orientation angle can be calculated as

$$\varphi(t) = arctan\frac{H_y(t)}{H_x(t)}$$

The above equation can however be applied only when the sensor is horizontal. This makes its use in head motion tracking scenarios difficult as the sensor cannot be guaranteed to always be horizontal. However, this can be compensated for by including an additional tilt sensor which measures the roll θ and pitch ϕ of the user's head. An overview of these type of sensors can be found in (Caruso & Bratland, 1998).

The main advantage of using magnetic sensors is that they are not susceptible to drifts as they use the earth's magnetic field as a reference. However they also have a number of limitations which make their use in head trackers difficult.

a) One important problem associated with these sensors is their susceptibility to distortions caused by the environment, for example by metal surfaces and the electromagnetic fields caused by lights, electric machines, like computers, microwaves etc.

b) Another problem that may arise when the head motion is not strictly in the horizontal plane is the need of an additional tilt sensor as described above.

Therefore not including a tilt sensor would lead to erroneous results when tracking head rotations. However, including a tilt sensor would further increase the price of the sensor.

3. Gyroscope Sensors

The gyroscope sensors make use of vibrating mechanical elements to detect head motion. All gyroscopes are based on the transfer of energy between two transfer modes of Coriolis acceleration which is proportional to the rate of rotation. The two most common approaches to realize a gyroscope are using vibrating beams and tuning forks. (Maenaka & Shiozawa, 1994) provides a great overview on gyroscopes based on beams. In the tuning fork method, the two tines of the fork are vibrated at their resonance frequency using electrostatic charge in x direction. When the sensor now rotates along the z axis an oscillation occurs in the y direction due to the Coriolis force, which when measured gives the rotation angle. For further details refer to (Yazdi et al., 1998).

The main advantage of these sensors is that they more accurately measure the rotation as compared to accelerometers and unlike the magnetometers do not require additional tilt sensors for compensation. The main limitation of these sensors is the drift caused due to temperature changes which needs to be compensated for in order to maintain accuracy.

A comparison of all the three sensors in head tracking for 3D audio applications is provided in (Christoph, 2007). The results from this review indicate that a head tracking device based on the Gyroscope sensor could be implemented in headphones to meet the size and power requirements to be implemented on a portable device such as a smartphone. Informal listening experiments conducted by the author suggested

comparable performance of headphones with gyroscopes to the Polhemus head tracker (Polhemus, 2012), a specialized and expensive dedicated head motion tracking system based on inertial sensors.

Even though gyroscope and headphone based head trackers can solve the size and power requirement for portable use, they are subject to drift due to temperature shift and age. These issues can be solved by recalibrating the sensor, which is a slow and difficult process for the end-user.

Computer Vision based Head Tracking

Computer vision based techniques have also been employed to track a user's head motion in order to deliver spatial audio. An approach to track head motion using four light sources and a web camera, implementing the POSIT algorithm (Dementhon & Davis, 1995) is described in (Mohan, Duraiswami, Zotkin, DeMenthon, & Davis, 2003). This approach led to a cheap methodology of generating head tracked spatial audio using a computer, web camera and inexpensive light sources. A survey of other computer vision techniques is presented in (Murphy-Chutorian & Trivedi, 2009).

While the computer vision techniques provide reasonable tracking of head movements, they suffer from a number of limitations. First, almost all of the computer vision based techniques require the user to face the camera. This requirement limits the use of these techniques on real world portable applications, as the user would have to always hold the camera in the device horizontally which may be difficult while moving in crowded spaces. An approach to eliminate this requirement is presented in (Ubilla, Domingo, & Cadiz, 2010). The authors used a Nintendo wiimote to augment the head motion

detection technique. Their approach detected and eliminated the camera inclination by using the wiimote's accelerometers to obtain user's head relative to earth. Second, the computer vision based approaches are sensitive to lighting conditions and thus have limited use in places with inadequate light. Third, these systems are computationally intensive which makes their use in portable devices limited. Finally, they are sensitive to the user's physical characteristics and are thus not completely identity invariant- an important requirement if the system has to be made commercially available.

Tracking Head Motion with Smartphones

Smartphones have become pervasive today and most of the high-end smartphone devices already come equipped with a microscopic vibrational gyroscope, 3 axis accelerometer, and magnetometer. The readings from these sensors can be combined to find the orientation of the phone.

The orientation information obtained from the smartphone can be used to track a user's head. In one such work described in (Naseh Hussaini, 2011), the smartphone was mounted on the user's head to obtain head tracked information. In a pilot study, the author found that the participants reported higher level of immersion when head tracking was enabled.

The smartphones' sensors have improved quite a bit in recent years but they still lack the required accuracy levels to be able to reliably provide orientation information to the user in indoor environments (Ogundipe, 2012; Ozcan, Fatih, Demirci, & Abul, 2012; Rodriguez, 2011; H. Wang, Elgohary, & Choudhury, 2012).

Although the above head-mounted approach may work, and is a viable solution for using spatialized audio in real-time portable systems as the accuracy of the smartphone sensor suite continues to increase, wearing smartphones on the head is not an ideal approach due to aesthetic issues. (Golledge et al., 2004) describes the findings of a survey where people rated the "cosmetic acceptability" of the navigational technology to be an important factor. Therefore while considering an accessible technology the visual aspect of the device cannot be ignored.

An alternate approach to solve the problem is to track the motion of the user's hand. Most blind and visually impaired users are already conversant in using a cane to detect obstacles. Thus sensors can be placed on canes to obtain hand tracking data. However, obtaining extra sensors for this purpose may be an expensive solution. Another approach could be using sensors already available in off the shelf smartphone devices and implementing the system on the smartphone itself. For the above propositions to be validated, people's ability to localize sound with hand movements needs to be tested. If localization performance with hand tracked spatial audio is found to be similar to head tracked spatial audio, we have good evidence of the efficacy of this approach and can build immersive systems around hand-tracking based on already ubiquitous smartphone devices. These systems would allow the users to localize sound while they are on the move using their hand movements.

This chapter describes a study which compares the target learning and spatial updating performance of participants with 3D audio generated by tracking user's hand motion and by tracking their head motion. The study also compares users' performance while

learning the object array with vision, which serves as a baseline control of optimal performance.

4.2 Method

The methodology of the second study compares spatial updating performance of the participants when learning an array of targets through spatial audio based on head motion, hand motion, and via visual inspection. The study was approved by the University of Maine's Institutional review Board (IRB). The experiment took about 1.5 hours to complete for each participant.

4.2.1 The Learning Modes

The learning modes for the study were:

1) Spatial audio with Head motion tracking

This mode emulated how objects would be heard with normal human hearing. The participants wore headphones with an inertia cube (Intersense Inc.) mounted as shown in Fig 4.1. The inertia cube is a head-tracking device developed by Intersense inc. and is based on nine miniature inertial sensing elements and uses Kalman Filters to provide head orientation with an accuracy of 1°.The inertia cube fed the Vizard virtual reality system ("WorldViz inc.," 2010), using the FMOD 3D library ("FMOD," 2011) with information on the current orientation, so that the sound could be modified accordingly.
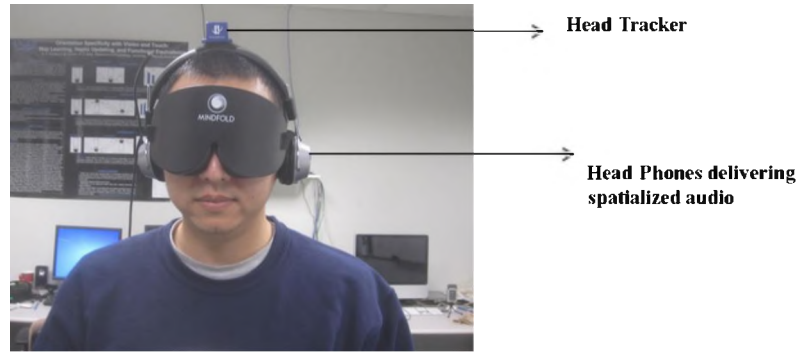
Figure 4.1 Spatial audio with Head Motion Tracking

The experimental trial began with the participant asked to learn the target array by orienting their head towards the left. They then slowly moved their head from a left to right direction. As the user's head came into a direct line with an object, they heard the name of the object along with the distance uttered continuously through their headphones generated through the rendering computer. Since the sound was coupled to the head motion of the user, it gave them an illusion that the sound source was located in the real space.

This sound was spatialized and the participants had the opportunity to move their head left and right to localize the sound completely. The sound played over a range of 30° left and right of the target. As the participant went past the object in the right direction, they heard the sound to be coming from the left ear. Similarly, if they went past the object in the left direction, the sound became louder in their right ear. The sound intensity of the object was equal in both ears as they were facing the target directly. Thus the participant learned the angular location of the target by localizing the sound with their head. As in the real world, the sound waves flow across the auditory field as the user moves, similar to optic flow with vision as we move our head.

While in the real world we hear 3D sound continuously emanating from a real object in all directions (e.g. an alarm clock), in the experiment the participant heard the 3D sound only within 30° to each side of the object's angular position. Though the angular range of 60° is somewhat limited as compared to the 360° range we have available in the real world, this restricted "auditory window" ensured that the participant heard only one target at a time while still providing a broad enough angular extent to readily localize the spatialized signal. This constraint ensured that the participants were able to localize the target with maximum accuracy as previous studies have shown that sound localization performance decreases substantially in the presence of interfering signals (Good, 1996; Langendijk, Kistler, & Wightman, 2001). Also, our pilot studies suggested that the 60° auditory window was enough for the participants to be able to accurately localize the sound by their head rotations.

2) Spatial audio with arm tracking

This mode was similar to the previous mode, except this time we placed the inertia cube on a stick as shown in Fig 4.2.

Figure 4.2 Spatial Audio with Arm Tracking

In this mode, the participants kept their head and body oriented in a fixed 0 degree position and they were only allowed to move their dominant arm to localize the sounds in the same sweeping fashion as they used when moving their head in the previous condition. The sound signals were modified by the Vizard software in response to the arm movement of the user. When the user pointed their arm directly towards the object they heard sound coming from both ears. When they moved their arm to the left of the target, they heard a higher intensity in their right ear, similar to what would happen if they moved their head while listening to a real target. In the same way when their arm went to the right of the object they heard a higher intensity in their left ear, again emulating real spatial hearing with head movements.

This mode is different than the haptic pointer interface described in (Loomis et al., 2005) or our hand pointing mode described in section 3.2. While the aforementioned modes triggered target names and distances only when the stick was in-line with a target, (that is based on proprioceptive information) the current mode ensured that the sounds were played continuously when the user's hand was within 30° left or right of the target. In other words, in the current mode, the user had the opportunity to move their hand left or right to localize the sound and hear its bearing in a spatialized manner, whereas in the previous modes, they only received a discrete non-spatialized sound when they were pointing directly to the target.

This mode aims to assess if the interaural spatial cues obtained by head movement can be replaced by arm movement. A finding of similar spatial updating performance with head and hand movement based spatial audio would mean that participants formed the same spatial image of the scene irrespective of whether their head or hand was tracked. This

would imply that we can replace head tracking (achieved with the expensive head trackers described in section 4.1) with hand tracking which can be achieved with handheld smartphone devices equipped with myriad sensors. This in turn would result in a more realistic and immersive 3D audio interface for indoor navigation for blind users than is currently available. However, this is a challenging task for the user as the spatial cues otherwise associated with movement of the head (where our hearing system is situated) would now be associated with movement of the hand. In other words, to be useful, there must be an accurate perceptual mapping of hand coordinates to head coordinates, as assessed on subsequent behavioral tasks.

3) Vision

In this mode, the participant stood at a fixed point and saw object images, illuminated by LED lights in a dark room (Fig. 4.3)
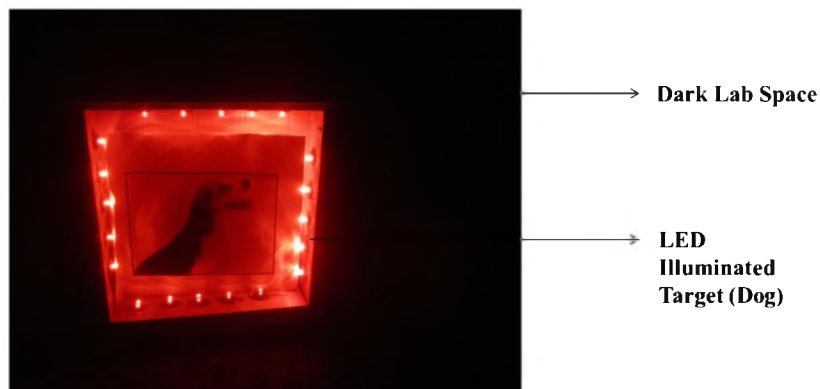
Figure 4.3 Vision Condition

The experimenter stood behind a closet and controlled the LED lights in the box in such a way that only one object was visible to the participant at a time. The exposure started

with the object placed at the target location on the left and slowly swept across to the right until the participant was exposed to all the targets.

The vision condition was not only matched in terms of information content (the target names and locations) with the spatial audio conditions (head motion or hand motion based), but also in terms of object encoding as the participants were able to see only one object at a time. The goal of limiting the access to the other two objects in the room was to match the information content requirements between the audio and vision conditions as much as possible.

In the current study, we evaluated spatial updating performance across the three modalities. Comparable performance with the three modes would mean:

a) Functional equivalence of spatial images that were generated when a target array is learned through vision or spatial audio.

b) Functional equivalence of spatial images that were generated with spatial audio – when user's head was tracked and when their hand was tracked.

4.2.2 Participants

Eighteen sighted University of Maine students (9 female, mean age= 24.9, SD= 4.08) participated voluntarily for the study and all provided signed informed consent forms. All the participants self-reported normal hearing and were monetarily compensated for their time and effort. The participants were screened using a simple spatial hearing test, where the participant was blindfolded and asked to point to the direction of a real sound source. All the participants passed this hearing test.

The study was conducted with only blindfolded sighted participants, rather than legally blind participants, as vision was one of the modalities being compared in the study. Also, evidence from previous studies suggests that there is little difference in learning between blindfolded-sighted and blind participants through non-visual modalities as spatial information is equally accessible to both groups (See Section 3.2.1 for further details).

4.2.3 Apparatus

This study was conducted in a lab room having dimensions 4.26 m by 5.71 m. The participants were blindfolded for the entire experiment (Mindfold, Inc. Tucson, AZ). The participants wore Creative HS-1200 (Creative Technology Ltd. USA) wireless headphones during the study to listen to instructions and stimuli. An inertia cube (Intersense , LLC Billerica, Massachusetts) was attached to the headphones/stick to determine orientation of the user's head/hand during the head motion and hand motion triggered conditions

A battery powered Light Emitting Diode (LED) light placed on the wireless headphones allowed us to track the precise position of the participant using an optical Precision Position Tracker (PPT) system (WorldViz inc., Santa Barbra, CA). This LED tracker also allowed us to measure the virtual positions of the targets in order to generate the Virtual auditory Environments (VAE).

For the visual condition, we used pictorial stimuli, which were placed in a frame, with LED lights around the frame (Fig. 4.3). A switch allowed the experimenter to turn the lights ON/OFF. In order to match the information content of audio modes with vision, we performed the study in a dark room for the visual condition. Whenever the experimenter

turned on the light of a particular frame to show the visual content of that target, other targets remained in the dark. This setup ensured that the vision operating in a "spatial domain" did not have any advantage over the other audio modes which operate in the "temporal domain".

The Virtual Auditory Environments were generated using Vizard 3.13(WorldViz inc., Santa Barbra, CA), using Python 2.4 (Python Software Foundation, 2012). The participants recorded their responses using a Nintendo Wii (Nintendo Inc.) remote ("Wiimote"). The A button of the wiimote was termed as "Start" and the B button was termed as "Stop" and was used to record the current state (Position and Orientation) of the participant. The Virtual Environment was generated and the study controlled through a desktop computer (Intel i7 2.65 GHz processor running on Windows XP with 2.5 GB RAM). The wiimote and the headphones were connected wirelessly to the controlling station with Bluetooth.

4.2.4 Stimuli

The target stimuli were names and pictures of objects found commonly in households and in office spaces. They were selected from the list of common stimuli as discussed in Table 3.1 (Snodgrass & Vanderwart, 1980).

In the Head motion triggered and Hand motion spatial audio conditions; the stimulus consisted of the target name followed by the distance. For example, Table 4 feet. In the visual condition, the participants learned the target name and distance by looking at the picture of the target. They learned the distance of the object through perception.

The audio stimuli were recorded as Wave files using the online AT&T Text to Speech Converter(AT & T Labs, Inc., 2010) using the US English Female voice Crystal. The stimuli were then edited in Audacity ("Audacity," 2010) to ensure consistent duration and waveform.

The polar coordinates of the targets locations were -75°/1.21 m, -45°/1.21 m, -15°/1.21 m, 75°/1.21 m, -45°/2.43 m, -15°/2.43 m, 15°/2.43 m, 45°/2.43 m and 75°/2.43 m. These locations were classified into three polygons (Fig. 4.4).
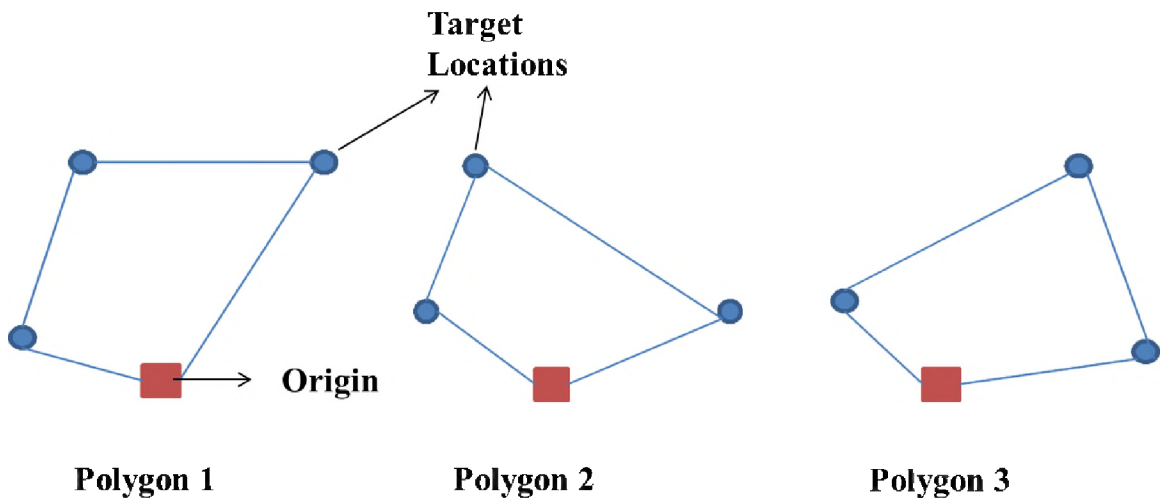
Figure 4.4 Target Locations for Experiment 2

4.2.5 Procedure

The design of the study was within subjects, with each participant being exposed to each of the three learning modes. The order of the three learning modes was counterbalanced. The overall procedure for the study consisted of five phases. The study began with familiarization of the equipment to the participants. The participants were given walking practice to walk 4 and 8 feet. Once the participants were comfortable in using the

equipment, and with blind distance perception, we started with the experimental trials of the study.

1) Learning Phase I

The first phase of the study was a multi-trial learning phase. The participants learnt a polygon consisting of three target names, through any of the three modes described above, while standing at the origin. As described earlier, in the Spatial audio with head motion mode, the participants heard the name of the target and its direction from its actual direction/position in 3D space when they positioned their head within 30° left or right of the target. The participants were able to localize the direction of the sound by moving their head from left to right and then right to left. Thus the participants received two exposures for the targets in a single trial.

The spatial audio with hand motion condition was similar where the participants first moved their arm in a sweeping fashion from left to right (instead of their head) and then reversed the process so they swept from right to left across the target array to localize the target name and location. They heard the name and distance of the target when their hand was within 30° left or right of the target. As with the head-tracked condition, upon hearing the auditory target signal, they had the opportunity to move their hand left or right to determine the exact location of the object. The exact angular location of the target was determined by the arm location where the participant heard the target in both ears.

Finally, in the vision condition, the participants learned the target polygon by seeing a stimulus one object at a time from left to right and then again from right to left as in previous conditions. Thus they were exposed to the targets two times as in the previous

conditions. The participants learned the objects by movement of their eyes. The duration of stimulus was set to 6 seconds and the duration between two stimulus presentations was 2 seconds. These were matched to the durations of spatial audio conditions through pilot studies.

2) Learning Criterion

To ensure that the participant learnt the array with sufficient accuracy, a learning criterion phase was introduced. To do this task, the participants stood at the origin and pressed the "Start" button on the wiimote to initiate the trial when they were ready. The computer randomly selected a target name from the three targets that the participants had learned in phase 1 using the current mode. Upon hearing the target name through their headphones, the participants were then asked to walk directly to this target. When they reached the location where they remembered the target to be, they pressed the "Stop" button. The experimenter then guided them back to the origin. The participants back translated to the origin each time. This process was repeated until the participants walked to all the three targets in the polygon for the current mode.

If the average "Target to Response" distance for the walks to the three targets was less than or equal to 0.739m, they passed the criterion, otherwise the participant had to learn the array again with the same mode. This criterion was chosen based on the average target to response distance described in our previous study (Chapter 3).

This process was repeated until either the participants passed the criterion successfully or they performed six learn-test criterion iterations, whichever was first.

3) Target to Target Walking Phase

Once the participant passed the learning criterion, they entered the next phase of the study which was the first walking phase known as the "Target to Target Walking Phase". In this phase, the experimenter led the participant directly to one of the target locations. The participants were reminded that they were aligned to this target with their back facing the origin. The participant then pressed the "Start" button and heard the instructions as "You are at the chair, walk to the couch". The participant was instructed to think about the location of the destination target from their current position and orientation and to only start walking to this target once it was instantiated in memory. When the participant believed that they had reached the location of the target, they pressed the finish button. At this point, the experimenter again brought them back to the origin via back-translation. The participants then re-oriented themselves to face laboratory north, with the help of the tactile foot rests at the origin.

We recorded "Time to think", as the time the participant took to recollect the location of the destination target. The participants were asked to not start walking from the source target until they had imagined the location of the destination target. We also recorded the response position of the participant, as defined by the location where they pressed the finish button. Finally, we recorded the total response time which was the time the participant took to walk from the source object location to the destination object location.

4) Learning Phase II (Re-exposure)

This phase was identical to Phase 1. The only difference between learning phases I and II was that in the latter, the participant was exposed to the target array only once as opposed to the double exposure allowed in Phase 1.

This phase allowed the participants to refresh the target locations in memory before making their responses in the Phase 5 testing, the polygon walking phase.

5) Polygon Walking Phase

In this phase, the participants walked to all the targets one by one, in a clockwise or counter-clockwise direction from the origin. To begin, the experimenter verbally reminded the participants of the sequence of the three targets in the array.

The participant stood at the origin and when ready, they pressed the start button. They then walked to the first target in the polygon. Once they believed they had reached the target location they pressed the finish button. They then continued to walk to the second target in the polygon and again pressed finish after reaching this location. They then walked to the third target and pressed finish again, at the remembered target location. Finally they oriented themselves to face the origin position, took three steps toward this position, and pressed finish again. We didn't ask the participant to walk fully back to the origin (and thus traverse the complete polygon), as our pilot studies suggested that the participants simply searched for the tactile foot-rest at the origin to mark their response, rather than deriving it via updating of the actual position.

The polygon walking phase tested whether the participants formed a global cognitive map of the scene, even though learning only occurred from a fixed position/orientation at the origin. If the participants walked to the targets successfully in this phase it indicates that they had formed a mental image of the whole scene in their mind, and were able to update it as they walked from one target to another. It is important for an individual to form these spatial images (independent of modality) to perform everyday spatial tasks

such as walking to a target after visiting another target. For example in the case of the following scenario with our friend Rita in the sandwich shop:

"Rita bought the sandwich and reaches out to the cash counter after visiting the soda fountain. At the cash counter she realizes that she would also like chips with her meal. So she pays for the chips at the cash counter and now walks back to the sandwich bar where the chips are kept."

This walk is different than the walk in the first scenario, where she walked to the sandwich bar from the door. In order to correctly walk from the cash counter she needs to update her location and have clear knowledge of the cognitive map of the global relations of the space and its constituent objects.

4.3 Results

We analyzed the data collected for the following three phases:

a) Phase 2. Learning Criterion Phase

b) Phase 3. Target to Target Walking Phase

c) Phase 5. Polygon Walking Phase

4.3.1 Learning Criterion Phase

A learning criterion ensured that the participants learned the target array before performing the walking phases 3 and 5. The participants passed the criterion test if their average walking error across the three target locations was less than or equal to 0.739m. The mean number of trials are given in table 4.7.

A within subjects Analysis of Variance (ANOVA) was conducted on the number of trials needed to achieve the learning criterion using the variable of modality. The effect of modality was significant $F(2,17)= 9.497$, $p<0.01$, $\eta^2_p =0.107$. Subsequent paired sample t-tests revealed that participants required fewer trials to reach criterion with vision (M= 1.0, SD=0.0) than with the Head motion based spatial audio condition (M=1.28, SD= 0.564), $t(53)= -3.622$, $p<0.01$. Participants also needed fewer trials to reach criterion with vision (M= 1.0, SD=0.0) than with the hand motion based spatial audio mode (M=1.33, SD=0.476), $t(53)= -5.148$, $p<0.01$. However no significant differences were found in reaching the learning criterion with the head motion based spatial audio mode (M=1.28, SD= 0.564) as compared to the hand motion based spatial audio mode (M=1.33, SD=0.476), $t(53)= -0.651$, $p=0.518$.

S. No.	Condition	Mean No of trials	Standard Deviation
1	Head Motion Based Spatial Audio	1.33	0.476
2	Hand Motion Based Spatial Audio	1.28	0.564
3	Vision	1.0	0.000

Table 4.7 Mean Number of Trials to Reach Criterion

4.3.2 Target to Target Walking Phase

In this phase the participants walked from one target location to another target location as described in section 4.2.5. We analyzed performance under the following

a) Time to imagine the destination target

b) Angle Errors

c) Distance Errors

d) Target to Response Distance

e) Response Time

Outliers greater than 2.5 standard deviations from the mean (n=4, 2.46% of total trials) were removed and were replaced with mean values prior to averaging

a) Time to imagine the destination target

The time to imagine destination target was calculated as the time taken by the participant to imagine the location of the destination target where they were supposed to walk. This was measured as the time elapsed from the "start" button press where the destination target utterance was first given, until when the participant took their first step to walk towards this target (when they moved 0.4 m). Higher imagine times exhibited for a condition would indicate that the participant required more time to recollect the location of the target learned from this modality and thus required more cognitive effort.

An Analysis of Variance on thinking times with the variable of modality showed no significant effect of mode $F(2,17)= 2.303$, $p= 0.103$, $\eta^2_p= 0.029$.
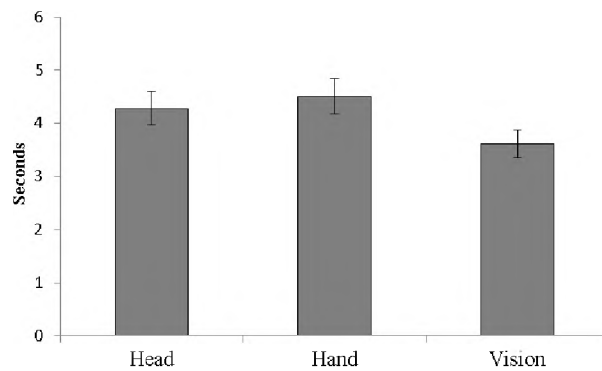
Figure 4.5 Mean Time to Imagine Target

b) Angle Errors

The angle between the source and destination target locations and the source and response target locations were calculated with the circular statistic method described in (Mahan, 1991). The difference in these angles generated the signed angle error. The absolute value of this signed angular error is referred to as absolute angle error.

1. Signed Angle Error

An Analysis of variance on signed angle error with the variable of modality showed no significant effect of the mode, $F(2,17)= 1.196$, $p= 0.305$, $\eta^2_p= 0.015$.

S. No.	Condition	Signed Angle Error	Standard Deviation
1	Head Motion Based Spatial Audio	1.439	14.717
2	Hand Motion Based Spatial Audio	6.482	18.504
3	Vision	4.040	16.644

Table 4.8 Mean Signed Errors in Target-Target Walking Phase

2. Absolute Angle Error

The absolute angle error for each trial was calculated as the absolute value of the signed angle error. ANOVA results showed no significant effect of modality on the absolute angle errors, $F(2,17)= 1.541$, $p= 0.217$, $\eta^2_p= 0.019$. The mean values of the absolute angle errors are depicted in Fig 4.6.
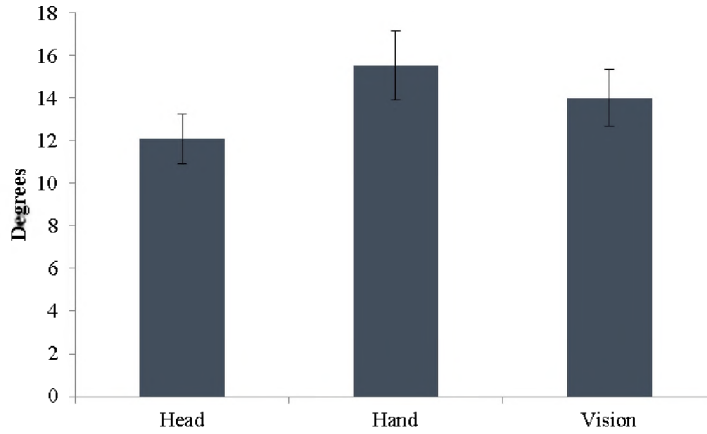
Figure 4.6 Absolute Angle Error for Target-Target Walking Phase

c) Distance Errors

The signed distance errors were calculated as the difference between the distance between the two target locations and the distance between the source target location and response. The absolute value of the signed distance errors gave the absolute distance error.

1) Signed Distance Error

ANOVA results showed no effect of modality on the signed distance errors, $F(2,17)= 0.938$, $p=0.394$, $\eta^2_p= 0.012$.

S. No.	Condition	Signed Distance Error(m)	Standard Deviation
1	Head Motion Based Spatial Audio	0.006	0.501
2	Hand Motion Based Spatial Audio	0.098	0.394
3	Vision	-0.014	0.451

Table 4.9 Mean Signed Distance Errors in Target-Target Walking Phase

2) Absolute Distance Error

ANOVA showed no effect of modality on the absolute distance errors, $F_{(2,17)}= 0.944$, $p=0.391$, $\eta^2_p= 0.012$.

S. No.	Condition	Abs Distance Error(m)	Standard Deviation
1	Head Motion Based Spatial Audio	0.396	0.301
2	Hand Motion Based Spatial Audio	0.323	0.243
3	Vision	0.350	0.280

Table 4.10 Absolute Distance Error in Target-Target Walking Phase

d) Target to Response Distance Errors

The target to response distance for the first walking phase was calculated as the distance between the target location and the response position of the participant for each trial. This measure gives us an estimate on how near the participants responses were to the targets. It is always positive as it is the Euclidian distance between the two points.

ANOVA results revealed a significant effect of modality on the target to response distance errors, $F_{(2,17)}= 3.641$, $p=0.029$, $\eta^2_p= 0.045$. Subsequent paired sample t-tests revealed that the response to target distance errors were significantly less in vision condition (M=0.596, SD= 0.330) than in the head motion spatial audio condition (M=0.776, SD= 0.397), $t(53)= 2.809$, $p=0.007$. There was no significant difference in means between the hand motion spatial audio condition (M=0.657, SD= 0.317) and vision condition (M=0.596, SD= 0.330), $t(53) = 0.992$, $p= 0.326$. There was also no significant difference in the means for target to response distance in head motion

(M=0.776, SD= 0.397) and Hand motion conditions (M=0.657, SD= 0.317), t(53)= 1.675, p=0.1.
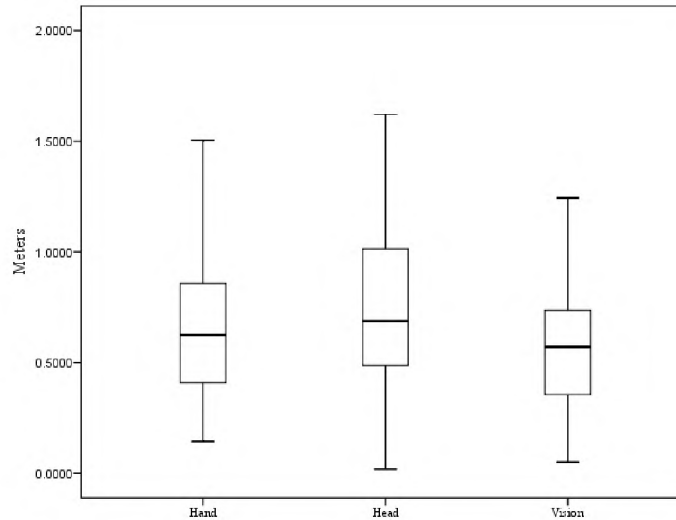
Figure 4.7 Target to Response Distance Error in Target-Target Walking Phase

e) Response Times

The response time in the target to target walking phase was calculated as the time taken by the participant to walk from one target location to the other. ANOVA results showed no significant effect of modality on the response times for the target-target walking phase, $F(2,17)= 2.451$, p=0.090, $\eta^2_p= 0.031$.
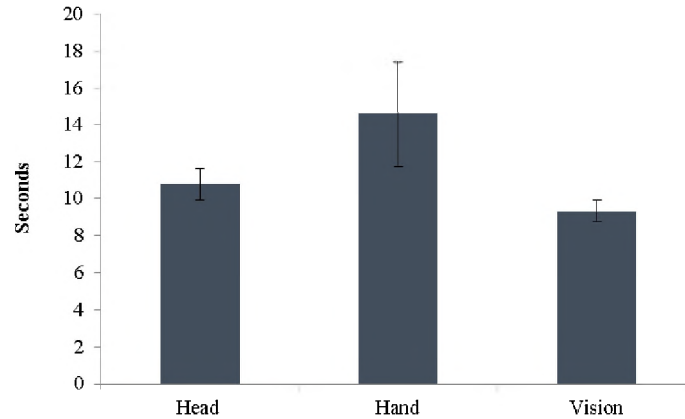
Figure 4.8 Response Times in Target-Target Walking Phase

4.3.3 Polygon Walking Phase

In the polygon walking phase the participants started walking from the origin and walked to the targets sequentially in a clock-wise or counter clock-wise direction. They marked their response each time they believed they had reached a target location. The details of this phase are described in 4.2.5.

We analyzed the performance of the participants in this phase according to the following measures:

a) Angle Errors

b) Distance Errors

c) Target to Response Distance

d) Response Time

Outliers greater than 2.5 standard deviations from the mean (n=9, 5.48% of total trials) were removed and were replaced with means prior to averaging

a) Angle errors

In the polygon walking task the angle error was defined as the difference between the angles generated by the origin and the target location and origin and response location. The absolute value of this signed angular error is referred to as absolute angle error.

1) Signed Angle Error

An Analysis of variance on signed angle error with the variable of modality showed no significant effect of the mode, $F(2,17)= 0.464$, $p= 0.630$, $\eta^2_p= 0.006$.

S. No.	Condition	Signed Angle Error	Standard Deviation
1	Head Motion Based Spatial Audio	-3.074	13.691
2	Hand Motion Based Spatial Audio	-5.567	15.467
3	Vision	-3.225	14.729

Table 4.11 Mean Signed Angle Error in Polygon Walking Phase

2) Absolute Angle Error

The absolute angle error for each trial was calculated as the absolute value of signed angle error. ANOVA showed no significant effect of modality on the absolute angle errors, $F(2,17)= 0.431$, $p= 0.651$, $\eta^2_p= 0.006$. The mean values of absolute angle errors are depicted in Fig 4.9.
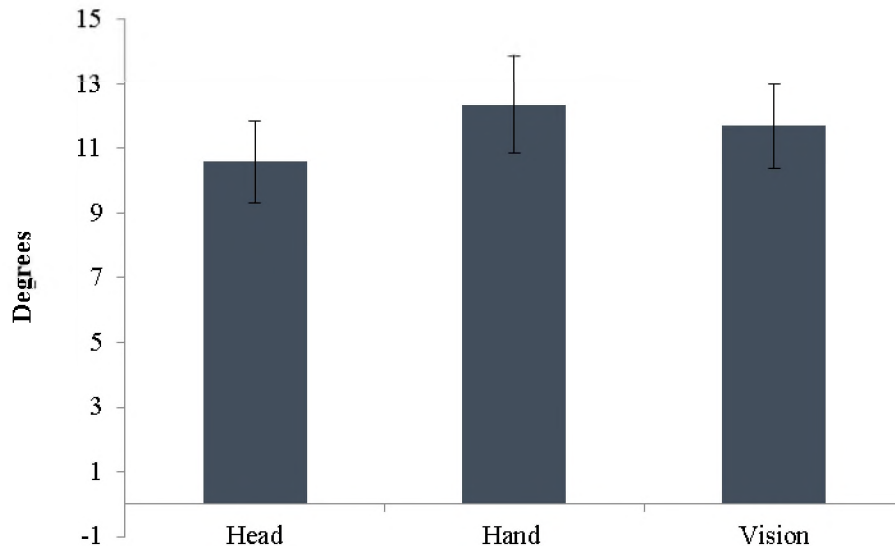
Figure 4.9 Mean Absolute Angle Errors in Polygon Walking Phase

The participants accumulated absolute angle error as they sequentially walked from target to target on the polygon walking task. This Significant accumulation of error was revealed by the ANOVA results, $F_{(2,17)}= 3.179$, $p=0.04$, $\eta 2p= 0.04$. Subsequent t-tests revealed that the participants walked the first leg of the polygon (M= 11.09, SD=7.32) with more greater angular accuracy than the third leg (M=18.87, SD= 24.67), t(53)= -2.161, p=0.036. Participants also walked the second leg of the polygon (M=11.65, SD= 12.28) with significantly less absolute angle errors than the third leg (M=18.87, SD= 24.67), t(53)= -2.232, p=0.03.

b) Distance Error

In the polygon walking phase the signed distance error was calculated as the difference between the target location and origin and response location and origin. The absolute value of the signed distance errors gave the absolute distance error.

1) Signed Distance Error

ANOVA showed no effect of modality on the signed distance errors, $F_{(2,17)}= 1.781$, p= 0.172, $\eta^2_p= 0.023$.

S. No.	Condition	Signed Distance Error(m)	Standard Deviation
1	Head Motion Based Spatial Audio	-0.004	0.433
2	Hand Motion Based Spatial Audio	-0.052	0.548
3	Vision	0.141	0.624

Table 4.12 Mean Signed Distance Error in Polygon Walking Phase

2) Absolute Distance Error

ANOVA results showed a significant effect of modality on the absolute distance errors, $F_{(2,17)}= 3.352$, p= .038, $\eta^2_p= 0.043$. Subsequent paired sample t-tests showed that the absolute distance error was significantly less in vision (M=0.321, SD= 0.287) than in the hand motion based spatial audio condition (M=0.501, SD= 0.391).

S. No.	Condition	Absolute Distance Error	Standard Deviation
1	Head Motion Based Spatial Audio	0.321	0.287
2	Hand Motion Based Spatial Audio	0.402	0.371
3	Vision	0.501	0.391

Table 4.13 Mean Absolute Distance Error in Polygon Walking Phase

The participants accumulated distance error as they walked the polygon. This was shown by the ANOVA results, $F(2,17)= 13.380$, $p<0.01$, $\eta^2_p= 0.149$. Subsequent t-tests revealed that the participants walked the first leg of the polygon (M= 0.295, SD=0.250) with less errors than the third leg (M=0.637, SD= 0.48), $t(53)= -4.322$, $p<0.01$. Participants also walked the second leg of the polygon (M=0.378, SD= 0.264) with significantly less target-target distance errors than the third leg (M=0.637, SD= 0.48), $t(53)= -3.516$, $p<0.01$

These findings are interpreted as showing that the ability to update target locations of the global array decays during this task. In other words, as there is no perceptual information to provide corrective feedback during polygon walking, these data show that path integration processes accumulate greater noise as people walk between the targets.

Target to Response Distances

The Target to Response distance for the polygon walking phase was calculated as the distance between the target location and the response position for each targets and response pair. This measure gives us an estimate on how near the participants responses were to the targets while traversing the polygon. It is always positive as it is the Euclidian distance between the two points.

ANOVA results showed no significant effect of modality on the Target to Response Distance Errors, $F(2,17)= 2.682$, $p=0.072$, $\eta^2_p= 0.035$.
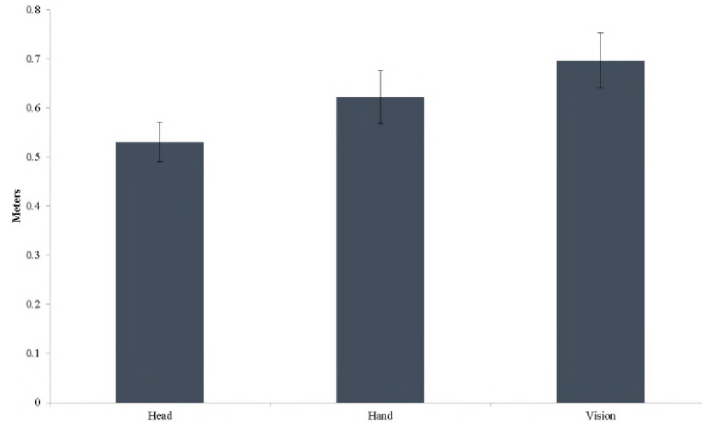
Figure 4.10 Mean Target to Response Distances

c) Response Times

The response times in the polygon walking phase were calculated as the time between consecutive "finish" button presses. ANOVA results revealed no significant effect of modality on the response times for Polygon walking phase, $F(2,17)= 1.874$, $p=0.157$, $\eta^2_p= 0.024$.
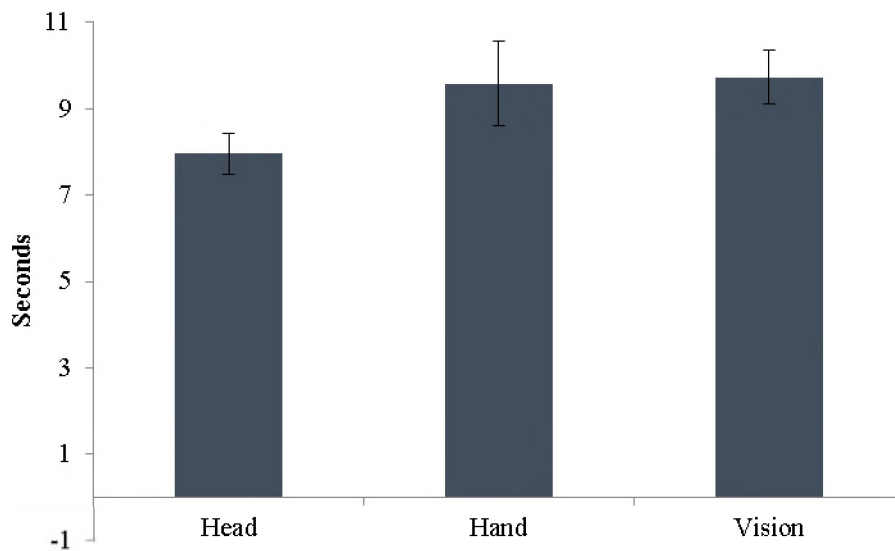
Fig. 4.11 Response Times in Polygon Walking Task

4.4 Discussion and Conclusion

I started this chapter by reviewing some of the previous research which demonstrates the necessity of head motion tracking in removing front-back confusions and in improving localization performance of the users. I then reviewed some of the techniques that can be used to track user's head motion. The traditional approaches included systems based on time of flight, spatial scan, mechanical linkage, phase difference, direct field sensing and hybrid methodology. Most of these systems were either too bulky or too expensive to be used in portable systems.

I then discussed whether headphones could be embedded with sensors to obtain head motion of users. Three different sensors namely, accelerometer, gyroscope and magnetometer were reviewed. The gyroscope sensor was found to fulfill the size and power requirement for use in portable systems. However, it is susceptible to its wear and tear related drifts and requires calibration frequently. Computer vision based techniques were studied next. These techniques are unsuitable for this research as they require the user to face the camera at all times. Since this is not feasible and desirable for a portable navigation system, I investigated smartphones as a potential candidate to provide head motion information as they are already sensor rich. However, these sensors are susceptible to noise when used indoors. We are optimistic about the sensors in smartphones to become resistant to indoor environmental noise in the near future.

To use smartphones as a replacement for head tracking for spatial audio, we need to investigate functional equivalence of spatial images formed through head motion and hand motion tracked spatial audio. To empirically evaluate the efficacy of hand motion

tracked audio with the "natural" head motion tracked audio and compare spatial updating performance of these modes with vision I designed and ran study 2.

The methods of the study are described in section 4.2. As expected the number of trials to reach criterion with vision was significantly less than with head motion based spatial audio or hand motion based audio. Analysis of data from the next phases reveals some interesting trends. Except for the target to response distance parameter in which vision was significantly better than hand motion based spatial audio, no significant differences were found in thinking times, absolute distance errors, absolute angle errors and response times. These results are exciting as they suggest that the spatial image formed through head motion based spatial audio, hand motion based spatial audio and vision are the same, at least for supporting the behaviors tested in this experiment.

The application of these findings in the real world means that we can replace the expensive head trackers described in this chapter with smartphone based trackers, as hand motion based spatial audio is functionally equivalent to head motion based audio and even vision. Thus it would help in the development of immersive 3D audio based spatial learning systems on off the shelf smartphone devices, without the need of expensive head-trackers. These systems would help in providing perceptual access to environmental information to visually impaired users such as Rita.

CHAPTER 5

COMPARING PERCEPTUAL AND NON-PERCEPTUAL AUDIO INTERFACES

IMPLEMENTED ON A SMARTPHONE

The previous studies explored the efficacy of audio based interfaces in helping individuals with low or no vision to learn spatial layouts. This chapter empirically investigates the efficacy of three smartphone based perceptual interfaces (Auditory Snapshot, SpeakOnTouch and Spatial SpeakOnTouch) with each other and with a non-perceptual interface (Spatial Language) in helping individuals form cognitive maps with the help of allocentric pointing. While our auditory snapshot and spatial language mode were identical to those described in Chapter 3, the latter two interfaces used kinesthesis and proprioceptive cues to aid spatial learning. In addition to the empirical study, participants were surveyed about their preferred choice amongst these interfaces.

The structure of this chapter is as follows: Section 5.1 provides an introduction and a description of previous work on this topic. Section 5.2 describes a study which was used to assess the efficacy of these audio interfaces on a smartphone. Section 5.3 provides the results from the current study. Finally in section 5.4, I discuss the implications of this research and provide some conclusions and broader contexts.

5.1 Introduction and Related Work

In this section, I will review some of the previous research in using kinesthetic information to deliver spatial information to the users.

5.1.1 Kinesthetic Cues as Perceptual Interfaces

Kinesthetic cues may be defined as the cues derived implicitly by humans by knowing the position of their body parts (arm, finger etc.) with respect to their body or with the surrounding environment. These cues are closely related to "proprioception" which is the sense of awareness of body parts and their movements. In our daily lives we utilize these cues unknowingly in many occasions. For example, imagine that you are in a dark room, and suddenly a mosquito comes and sits on your left arm. You would be able to brush it away with your right arm, even without being able to see its exact location. To perform this seemingly simple task, your brain has to construct an updated map of the body and its appendages in space and combine this information with the tactile information obtained through contact with the mosquito. Similarly, most proficient typists are able to type without looking at the keyboard or even with their eyes closed because the brain matches the knowledge of the key-map with the current location of the fingers. These cues are perceptual and are used by the brain intuitively. In the previous example when the mosquito bites us, our hand automatically reaches out to the point of contact. We do not require cognitive mediation to coordinate our hand to achieve this goal. Thus, kinesthetic interfaces can be used as effective non-visual modes of spatial learning by providing a map between the finger's positions in relation to the surrounding space.

5.1.2 Related Work

We already have seen some of the approaches to provide spatial information using non-visual interfaces in Chapter 2. Kinesthetic cues have also been used to deliver non-visual spatial content to blind and low-vision users in the past. The most important work in this

domain is reviewed below. I describe our methods (SpeakonTouch and Spatial SpeakonTouch) in section 5.2.1.

Daunys & Lauruska (2007) described a tablet or touchscreen computer based system which used sonification (non-speech audio tones) and speech audio to provide map based information. When the user touches a new region, its name is announced through speech. When the user is exploring a region, a constant tone is played. The volume of the tone increases as the user moves away from the boundary of the region.

Timbremap (Su, Rosenzweig, Goel, de Lara, & Truong, 2010) is another sonification based map learning system which allows blind users to explore indoor spatial layouts on off-the-shelf smartphone devices. The user explores the spatial layout by moving their finger on the screen. As they hit a path represented by a line on the map, they hear a tone. When they strayed left of this line, they heard a sound in the right ear hinting them to shift right. Similarly, when they move their finger on the right they would hear the sound in their left ear. In this way the users learned to follow spatial paths with the help of audio information. An empirical study evaluating the participants' performance in learning the layouts with the system found that the participants learned the map with 80% accuracy.

Several studies have combined audio and kinesthetic modalities to convey map information. One such system is TouchOver map (Poppinga, Magnusson, Pielot, & Rassmus-Gröhn, 2011). These authors investigated the use of vibration and speech audio to make maps on smartphones more accessible to the blind community. This system used OpenStreetMap as a platform to provide geographic content. As the user touched a physical feature on the map (e.g. a road), they heard continuous speech saying the name

of the road and the phone vibrated to indicate its spatial extent. In a study to evaluate the efficacy of this approach, participants learned maps non-visually (the visual access to the smartphone was blocked through a cardboard box) and then drew sketch maps. The analysis of these hand drawn maps indicated that the participants were generally successful in learning maps with the system. Raja, (2011) describes a novel non-visual spatial interface on a smartphone device using vibrotactile and audio cues known as "vibro-audio map". This system is used to convey indoor spatial layouts to the user. Whenever the user touches the rectangle representing corridors, the phone vibrates. In the room exploration mode, the users can tap on the map to learn about the room through speech. An empirical study found no significant differences in spatial learning through a comparison of Vibro-Audio Maps and traditional tactile maps, the gold standard in imparting spatial knowledge to people with low or no vision.

All the work described above used kinesthesis as a useful cue to deliver spatial information. We further explored the efficacy of this approach in helping people form mental spatial representations through the study described in this chapter. I was specifically interested to compare the usefulness of this perceptual interface with our previous perceptual interface (3 D audio). I also wanted to explore if spatialization of the kinesthetic audio cues would improve user's ability in forming cognitive maps. Moreover, I wished to explore how these perceptual interfaces fared against spatial language, the non-perceptual interface most commonly used to give non-visual spatial information to users.

In the next section I describe a study designed to evaluate spatial learning performance with four modes: 1) Auditory Snapshot 2) SpeakOnTouch 3) Spatial SpeakOnTouch 4) Spatial Language.

5.2 Method

This section, describes the methodology for the third empirical study which compared the spatial representations formed by the participants when learning a spatial layout consisting of four objects, through three perceptual modes and one non-perceptual mode, as implemented on a smartphone. The study was approved by the University of Maine's institutional review board (IRB) and took around 1.5 hours to finish.

5.2.1 The Learning Modes

a) Auditory Snapshot

This mode was implemented using spatial audio without head tracking, and was similar to the auditory snapshot mode described in section 3.1.2.1. The auditory snapshot starts with the object on the left most part of the scene playing first, followed by the next object and so on until all targets have been sounded from their virtual position in space. One snapshot is said to be completed when a person has heard all the object names along with their distances flowing from left to right across the target array. Each utterance of the object name is coupled with the distance of the object in the current implementation. The azimuth information of the object is provided to the user directly in the spatial audio signal. As an example: Suppose a scene (Fig. 5.1) consists of four objects, a table, a chair, a shelf, and a fan placed at 8 feet, 4 feet, 4 feet and 8 feet at angles: -90°, - 45° and 0° and +90°, respectively. The auditory snapshot of the scene would sound like: Table 8

feet, Chair 4 feet, Shelf 4 feet, and Fan 8 feet. Each sounds to the listener as if the objects were placed at the respective angles in the real world.
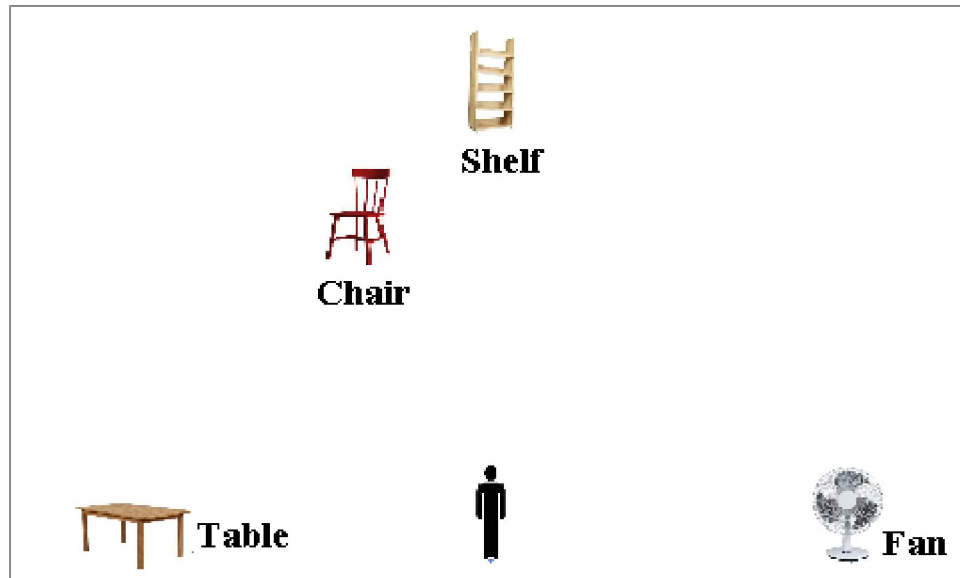
Figure 5.1 Sample Scene for Experiment 3

The participants listened to the snapshots unlimited number of times to help facilitate that they formed a spatial representation of the layout in their mind. They initiated each snapshot by pressing the "Menu" hardware key on the smartphone, located at the bottom left corner of the phone.

As described in chapter 3, this interface is useful in providing a global view of the scene to the user. The user is able to learn the objects and their spatial locations in the scene in a natural way through the virtual sound scene created by spatialized audio. Since spatial audio based interfaces are perceptual, this interface requires minimal thinking on the user's part to comprehend the object array.

As discussed in Sections 4.1 and 4.2, head tracking plays a vital role in minimizing front back confusions and allows a user to become completely immersed in the spatial sound scene by providing the interaural cues to the sound rendering engine. We also found through the study described in section 4.3, that spatial learning performance was not significantly affected when head motion based spatial audio was replaced by hand motion based spatial audio. This led us to the prospect of using the built-in orientation sensors, accelerometer, gyroscope and magnetometer in the smartphone device (See section 4.1.3.1 and 4.1.3.3 for details) to provide our application with the hand movement information. However, this version of the Auditory Snapshot interface did not utilize the sensors' capability as our pilot studies indicated noise in the readings obtained from the sensors. Previous research has also acknowledged this problem in indoor environments (Ogundipe, 2012; Ozcan et al., 2012; Rodriguez, 2011; H. Wang et al., 2012). The problem has generally been identified as the presence of various sources of noise such as metal file cabinets, electric wiring, cathode ray tube monitors, refrigerators, etc. The authors have proposed ways to counter this problem using sensor fusion techniques and smoothing algorithms. However, most of these techniques are computationally expensive and none of them have been tested for our purpose of tracking hand motion for spatial audio generation. Thus, we decided to implement the system as an initial proof-of-concept interface. Also, since all target objects were located in front of the participant there was no problem of front back confusion. Our future work would involve tracking user's hand motion by using the smoothed sensor readings from these devices to create an even more immersive experience for the user.

111

b) SpeakonTouch mode

The SpeakOnTouch mode is a perceptual mode which allows the user to learn the spatial layout primarily using kinesthetic cues. This mode is also implemented on the smartphone and allows the user to explore the spatial layout of the targets via kinesthetic information and audio labels as they move their finger around the touchscreen. As no vibration or other tactile indicators are given, other haptic cues beyond kinesthesis are limited.

In this mode, the blindfolded participants began the learning process by first finding the tactile landmark, a tactile "Loc-Dot" (Maxi-Aid Inc., 2012) placed at the origin position. They then moved their finger on the screen area to search for an object. As they touched any target they heard a tone (sine wave, 220Hz.). They then tapped on the screen to hear the name of the object and its distance from the tactile origin. Thus they got the angular information by swiping their finger at the location they heard the object name and tracing ack to the tactile origin, and received the object name and distance by speech which indicated the name of the object along with its distance from the origin. For example, the scene in Fig. 5.1 was available on the smartphone as Fig 5.2. Upon tapping the region indicated by the red circle on the lower right corner of the screen (Fig. 5.2), the participants heard "Fan 8 feet". They understood its angular location as 90° on their right, by swiping their finger back to the tactile landmark in the center of the screen. Thus the angular information about the target was available perceptually through the kinesthetic sense.

On touching this region the user hears "Fan" in both ears.

Smartphone
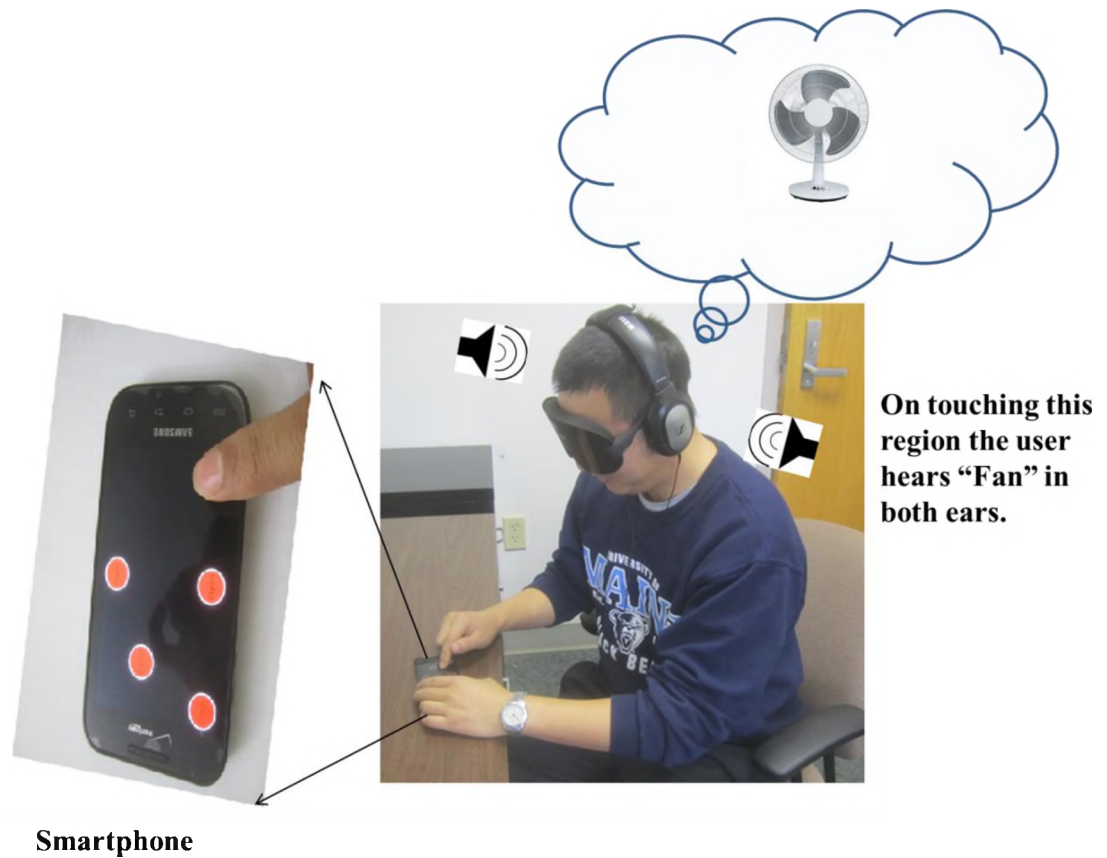
Figure 5.2 Illustration of SpeakOnTouch Mode

c) Spatialized SpeakOnTouch mode

The Spatialized SpeakOnTouch mode was similar to the SpeakOnTouch mode described in the previous section, except that this time the participants learned about the direction of the object with both kinesthetic and spatialized audio cues.

The blindfolded participants first searched for the tactile origin. They then explored the screen space for targets. As they touched any target they heard a tone (sine wave, 220 Hz) which was spatialized i.e., it appeared to the listener as if coming from the real direction of the target. Again when the listener tapped on the screen, they heard the name of the object along with its distance. This utterance of the object name and distance was also

spatialized. Therefore, the participant received information about the spatial layout through two perceptual cues: a) The kinesthetic cue, and b) The spatial audio cue.

As an example, when the user tapped on the red circle on the bottom right corner of the screen shown in Fig. 5.3, they heard "Fan 8 feet', only from their right ear.

On touching this region the user hears "Fan" only in right ear.
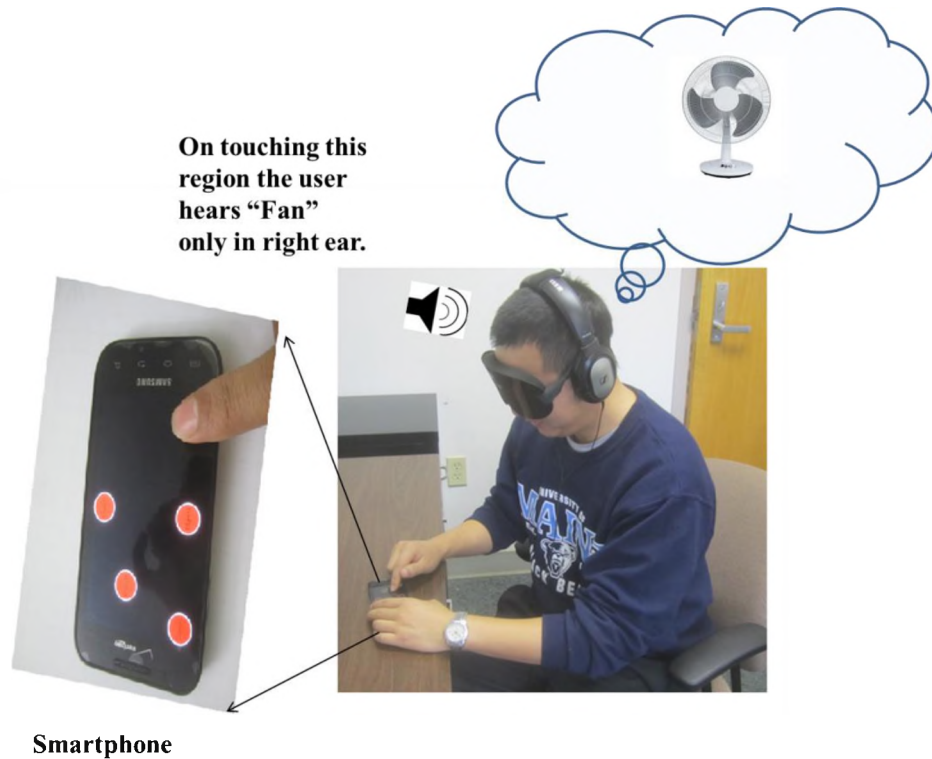
Smartphone

Figure 5.3 Illustration of Spatial SpeakOnTouch Mode

d) Spatial Language Mode

The spatial language mode was similar to the mode described in section 3.1.2.4. However, this time we implemented the mode on a smartphone. Another difference was that the participant had the ability to control the utterance of the target names and locations through gestures. I implemented two gestures, namely "forward" and "backward" gestures (Fig. 5.4). The forward gesture was registered when the user swiped

their finger from left to right. The backward gesture was registered when the user swiped their finger from right to left.

"Backward" Gesture   **"Forward" Gesture**
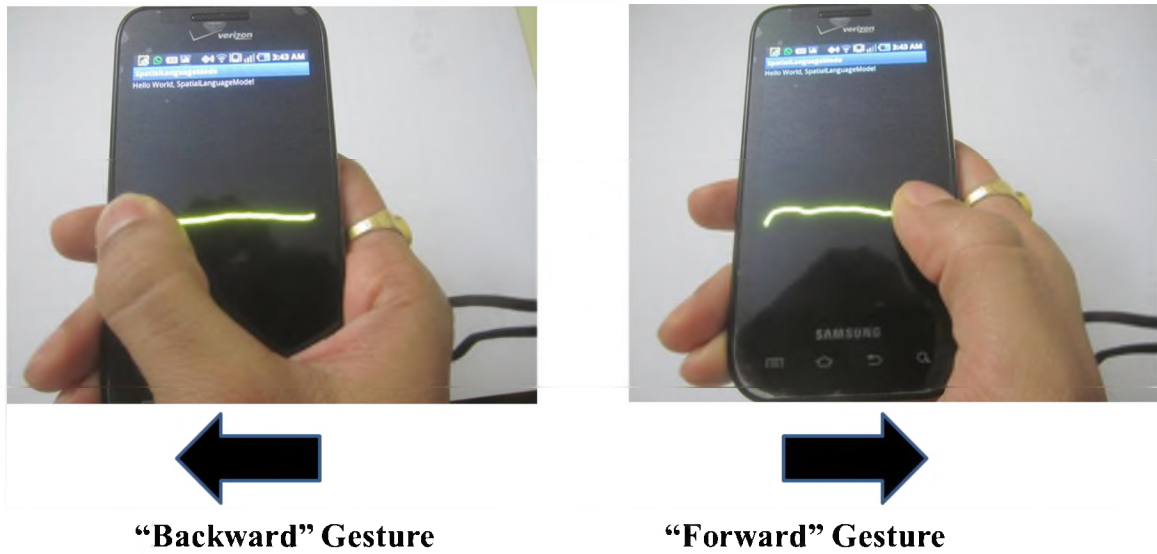
Figure 5.4 Gestures for Spatial Language Mode

The participants started learning the spatial layout with a forward gesture. The first forward swipe led them to learn the object located on the leftmost side of the scene. The next swipe led them to learn the next object located on the right of the first object. Subsequent forward swipes led them to learn the objects progressing to the right. Also, backward swipes led them to learn the object on the left of the current object. When the participant reached the rightmost object, any subsequent forward swipe would "cycle" such that they would hear the first object again. The participants were allowed to learn the spatial layout as many times as they wished.

As an example, the scene in Fig 5.1 could be learned with gestures described in Table 5.1

S.No	Gesture	Speech Output
1	Forward	Table 8 feet 90° Left
2	Forward	Chair 4 feet 45° Left
3	Backward	Table 8 feet 90° Left
4	Forward	Chair 4 feet 45° Left
5	Forward	Shelf 4 feet 0° Ahead
6	Forward	Fan 8 feet 90° Left

Table 5.1 Illustration of Spatial Language Mode

The spatial language mode is a non-perceptual mode and thus requires cognitive mediation on the part of the user. As we discussed in section 3.1.2.4, the efficacy of the spatial language mode in cognitive map development has been studied extensively in the past.

5.2.2 Participants

Sixteen sighted University of Maine students (9 female, mean age= 21.8 Years, SD= 2.63) participated voluntarily for the study and signed informed consent forms. All the participants reported normal hearing and were monetarily compensated for their time and effort.

This study was also conducted with blindfolded sighted participants, because of the ease of recruitment. Evidence from previous studies suggests that there is little difference in learning between blindfolded-sighted and blind participants through non-visual

modalities as the information conveyed is equally accessible to both groups (See Section 3.2.1 for further details).

5.2.3 Apparatus

This study was conducted in an office space. The participants were blindfolded for the duration of the entire experiment (Mindfold, Inc. Tucson, AZ, 2012). They wore Sennheiser HD 201 headphones (Sennheiser Electronic Corporation, Lyme, CT) during the study to listen to the stimuli. The smartphone used was a Samsung Galaxy S (Samsung Electronics, Suwon, South Korea) running on the Android Operating System version 2.2 (Fig. 5.4). The phone had a 1 GHz processor and 800 x 480 display (122.4 mmx64.2 mm). A custom experimental pointing device was fabricated using the Arduino Uno microcontroller (Arduino). The device (Fig. 5.5) was connected serially through the Universal Serial Bus (USB) interface to a Lenovo Computer with an intel i5 processor running Windows 7 at 2.4 GHz.

Figure 5.5 Experiment Setup for Study 3

5.2.4 Stimuli

The target stimuli were names and pictures of objects found commonly in office spaces. They were selected from the list of commonly found object items found by a survey conducted by pilot work in the lab (Kesavan and Giudice, unpublished pilot data). Table 5.2 depicts the names of the various targets used during the experiment.

Printer	Computer	Chair	Calendar
Board	Bag	Table	Coat
Fan	Mug	Shelf	Folder
Book	File	Trash	Pen

Table 5.2 Target Names for Study 3

The audio stimuli were recorded as Wave files using the online AT&T Text to Speech Converter (AT & T Labs, Inc., 2010) using the US English Female voice Crystal. The stimuli were then converted to MP3 format in Audacity (Audacity, 2010) to reduce the size of the files.

The targets consisted of four scenes with an additional scene used for introducing the interfaces. The polar coordinates of the target locations were +90°/1.21 m, +45°/1.21 m, 0°/1.21 m, +90°/2.43 m, +45°/2.43 m, and 0°/2.43 m.

5.2.5 Procedure

The design of the study was completely within subjects, with each participant being exposed to each of the four learning modes. The order of the four learning modes was counterbalanced. For every mode the study consisted of the following 5 phases:

1)   Interface Familiarization

2)   Learning Phase

3)   Learning Criterion

4)   Pairwise Pointing

5)   Task Load Test

Once the participants finished the six phases for all modes, they were asked to rate the interfaces in order of their preference. They were then given an opportunity to provide suggestions on improvement of these interfaces.

We will now explain each of the five phases of the study in more detail.

1) Interface Familiarization

The study began with allowing the participants to get familiar with the interface for that particular trial. The participants were allowed to go through the interface without any time limitations and were encouraged to ask questions while doing so.

For the interface familiarization phase for the first mode for each participant, all the experimental steps (1-5) were followed. This ensured that the participants were aware of

and understood the difference between the egocentric pointing (Phase 3) and allocentric pointing (Phase 4). This also gave the experimenter an opportunity to convey information about the subjective load test (Phase 5).

2) Learning Phase

In the learning phase, the blindfolded participants learned the spatial scenes with the help of each of the modes as described in section 5.2.1. For the auditory snapshot mode the participants were allowed to hear the 3D snapshots until they believed they had formed an accurate mental image of the scene in their mind. Similarly, for the spatial language mode the participants were free to swipe through the target scenes until they successfully built a spatial image of the scene. Also, for SpeakonTouch and Spatial SpeakonTouch modes the participants explored the scene with their fingers until they believed that they had a thorough acquisition of the spatial scene.

We did not enforce any time constraints in the learning phase as different modalities require different time periods to learn the same spatial scene. Klatzky et al., (2002) found that participants learned the targets slowly with language, a non-perceptual mode as compared to when learned with perceptual modes (3D audio and vision). However, subsequent investigation by Loomis et al., (2002) found that despite the disadvantage of slow learning through language, the mental spatial representations formed through language and through perceptual modes appear to be the same.

3) Learning Criterion

Once the participants encoded the spatial representation of the scene through the current mode, the experimenter probed them with the four target names in a random order. After hearing the name of the target, the participant rotated the pointer (Fig. 5.5) to the direction of the target. Before each pointing response by the participant, the experimenter aligned the arrow of the pointer to the participant's sagittal axis. After the participants pointed to the four targets, they were asked to rank the distances of the objects. If the participant's absolute pointing error across the four modes was less than 15° and their reported ranks achieved a correlation of 0.75, they passed the learning criterion, otherwise they were asked to learn the scene again with the same mode. Alternate learning and testing of the targets continued until the participants passed the criterion or they learned the scene six times, whichever was first.

4) Pairwise Pointing

After the learning criterion phase, the participants were probed with pairs of targets for that particular scene. For example for scene in figure 5.1, the participants were asked to point from one target to another (e.g. point from shelf to fan). Successful formation of a spatial image would allow the participants to compute target-target relations that are allocentric—not based on a coordinate system with them as the origin. There were four different objects and this resulted in 12 different pairs.

5) Task Load Test

To subjectively procure the workload estimates while performing the learning task, the NASA Task Load Index (NASA-TLX) was used (Hart & Staveland, 1988). The NASA-

TLX is a subjective and multidimensional tool that allows the experimenter to assess the workload on six different subscales, including: mental demand, physical demand, temporal demand, performance, effort and frustration. We administered the paper-pencil version of the test (appendix A).

After completion of the five phases for all the four modes, the participants ranked the four modes in order of their preference levels (1= most favorable, 4=least favorable). Finally, the participants were allowed to give subjective feedback and suggestions, some of which are discussed in Section 5.3.5.

5.3 Results

In this section we report the data analysis results for Phase 3 (Learning Criterion), Phase 4 (Pairwise Pointing) and Phase 5 (Task Load Test). We also report the preference ratings and subjective comments for the four modes.

5.3.1 Learning Criterion

In the learning criterion phase participants performed egocentric pointing to the four objects in a scene. If the average absolute pointing error was less than 15° and the distance ranking had a correlation of 0.75 the participant passed the test (Section 5.2.5). We analyzed

1. The number of trials needed by the participant to reach criterion.

2. Pointing Error for successful learning test

3. Response Latency for successful learning test

1. Number of trials to achieve criterion

No participant required more than 2 trials to achieve the criterion. A within subjects Analysis of Variance (ANOVA) was conducted on the number of trials needed to achieve the learning criterion using the variable of modality. The effect of modality did not reach significance $F(3,15)= 2.609$, $p=0.06$, $\eta^2_p =0.115$. However there was a trend for the participants to take more trials to achieve criterion for Auditory Snapshot (M=1.38, SD=0.5) and Spatial Language modes (M=1.25, SD=0.447) than SpeakonTouch (M=1.06, SD=0.250) and Spatial SpeakonTouch (M=1.06, SD=0.250). The mean number of trials needed to achieve learning criterion are shown in Table 5.3.

S. No.	Mode	Mean number of trials	Standard Deviation
1	Auditory Snapshot	1.38	0.500
2	SpeakOnTouch	1.06	0.250
3	Spatial SpeakOnTouch	1.06	0.250
4	Spatial Language	1.25	0.447

Table 5.3 Average Trials Needed to Achieve Criterion

2. Pointing Error for successful learning tests

We calculated the average pointing error for each successful learning test as the average of all absolute pointing errors for that test. An analysis of variance on average pointing

error with the variable of modality showed no significant effect of the mode, $F_{(3,15)} = 0.598$, $p = 0.619$, $\eta^2_p = 0.029$
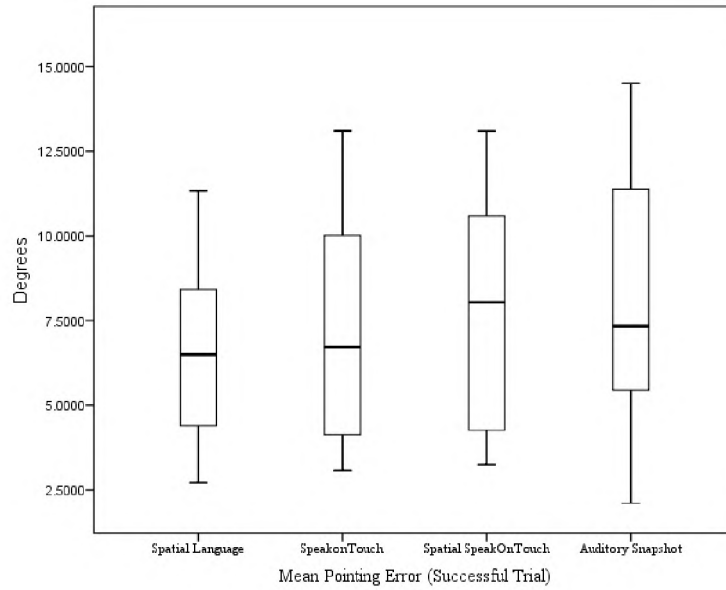

Figure 5.6 Mean Pointing Error for Successful Trial

3. Response Latency for the successful learning test

We also calculated the average time taken by the participant to perform egocentric pointing for successful tests. Again ANOVA showed no effect of modality on response times for successful pointing tests, $F_{(3,15)} = 1.919$, $p = 0.136$, $\eta^2_p = 0.088$.
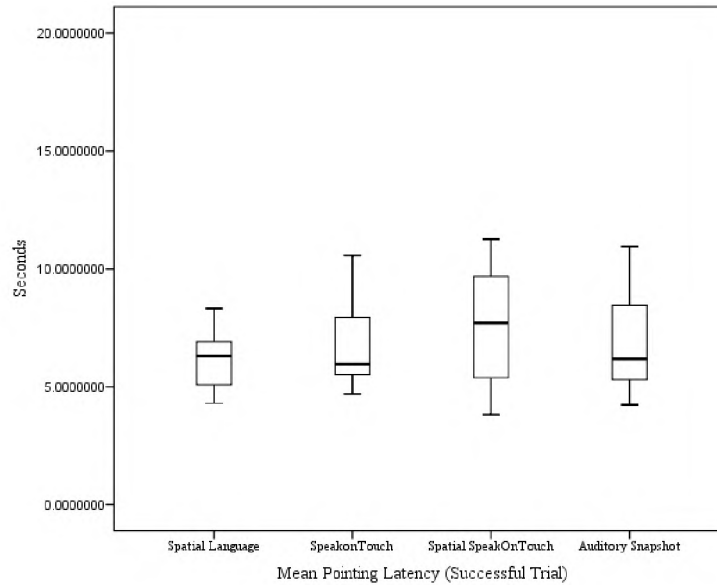
Figure 5.7 Mean Pointing Latency for Successful Trial

5.3.2 Pairwise Pointing

After successful completion of the learning criterion, the participants performed allocentric pointing as described in section 5.2.5. Two measures were relevant for data analysis: Absolute pointing errors and pointing latency.

1. Absolute Pointing Errors

We define signed pointing error as the difference between the pointing response by the participant for target- target angle and the actual angle between the two targets. We calculated the angle between the two targets with the help of a circular statistic method as described in (Mahan, 1991). We do not report ANOVA's on signed pointing errors as the means are subject to cancelling effects from target pairs with biases from different directions. We instead analyzed the absolute pointing errors.

125

The absolute pointing errors were calculated as the absolute value of the signed pointing errors discussed above. These pointing errors are an indicator of level of accuracy in the formation of a spatial image formed by the participant. Higher pointing errors indicate that the mental image was less accurately formed. ANOVA showed a significant effect of modality on absolute pointing errors, $F(3,15)= 2.781$, $p=0.04$, $\eta^2_p = 0.011$. Subsequent paired samples t-tests revealed that participants performed allocentric pointing task with less errors with auditory snapshot (M=24.357, SD= 38.541), a perceptual mode as compared to Spatial language, a non-perceptual mode (M=38.056, SD= 64.133). The mean absolute angle errors for the four modes are shown in Fig 5.6.
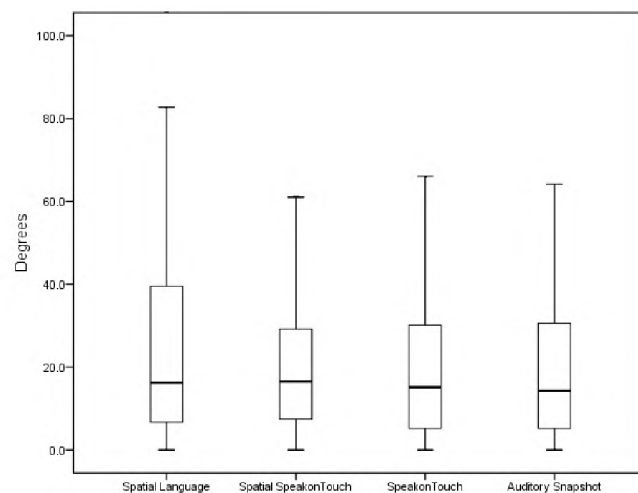

Figure 5.8 Absolute Pointing Error in Pairwise Pointing Task

4. Pointing Latency

The pointing latency for each trial was recorded by our pointing device. It was calculated as the time period between the experimenter's button press and the participant's response button press. The pointing latency measures the time taken by the participant to recollect

the spatial image before making allocentric judgments and is therefore an indicator of cognitive load required to recall the image. Higher response times mean the participant had difficulty in remembering the spatial image.

ANOVA results showed no significant effect of modality on the response times for target-target pointing performance in this phase, $F(3,15)= 0.755$, $p=0.520$, $\eta2p= 0.003$. This means participants took almost the same time across modalities to recollect the spatial image and perform the allocentric pointing task. The means for the response times for the four modes are shown in Fig 5.7
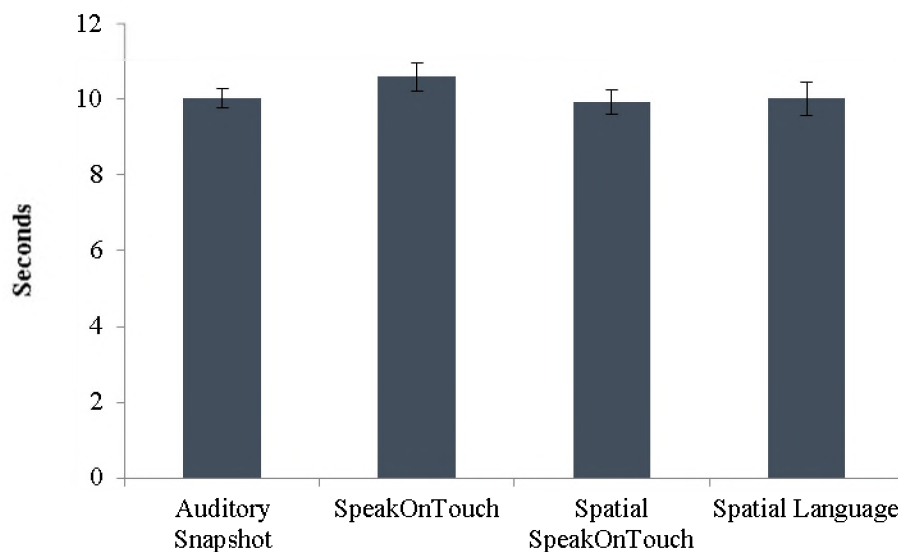

Figure 5.9 Average Latency for Pairwise Pointing Task

5.3.3 Task Load Test

As we discussed in section 5.2.5, we administered the NASA-Task Load Index Method (NASA-TLX) to evaluate workload for learning each interface, just after the pairwise pointing task. This index is a multidimensional rating procedure and provides an overall

score for six subscales: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort and Frustration level. Although it is possible to obtain a single work load index by assigning weights to each of the above six subscales, we preferred to evaluate ratings by the participants on each of these scales separately.

The NASA-TLX Load index (Appendix 1) allows the participants to rate the six subscales on a 20 point scale. We multiplied the responses with 5 to obtain a scale ranging from 0-100. The performance subscale ranges from Perfect (0) to Failure (20). All other subscales range from Very Low (0) to Very High (20). Thus the lower the score on all subscales, the lesser is the perceived workload. Fig 5.8 shows the average workload for each of the subscales for the four modes.
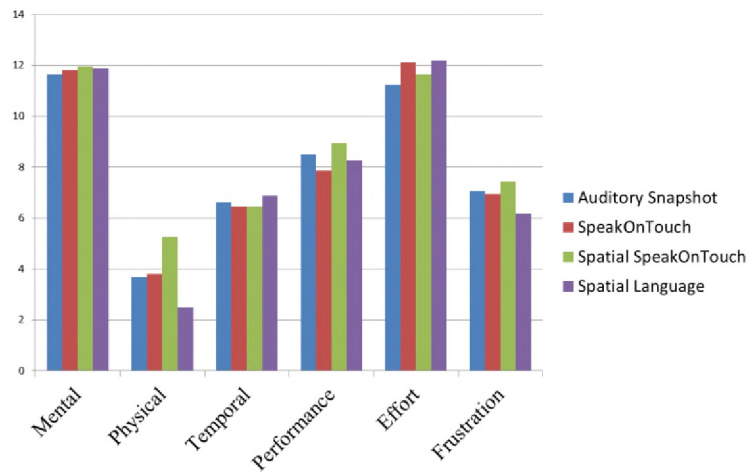
Figure 5.10 NASA TLX Mental Load Analysis

ANOVA results revealed no significant differences in perceived mental load for the four modes on the six subscales.

5.3.4 Participant Overall Preference

After the completion of the NASA-TLX for the last mode, we administered another survey where the participants were asked to rate the modes in order of preference with 1 being the most preferable and 4 being least preferable. Fig 5.9 shows the mean ratings for the four modes.
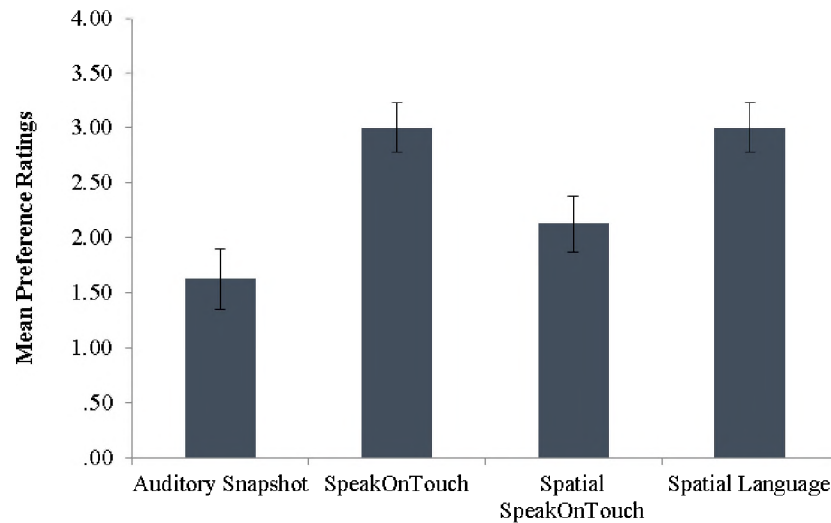
Figure 5.11 Mean Subjective Ratings

ANOVA showed a significant effect of modality on the preference levels for the four modes, $F(3,15)= 7.739$, $p<0.01$, $\eta^2_p = 0.279$. Subsequent paired samples t-tests revealed that the participants preferred Auditory Snapshot (M=1.63, SD= 1.088) over the Spatial Language mode (M= 3.00, SD=0.894), t(15)= -3.780, p= 0.002. Participants also preferred Auditory Snapshot (M=1.63, SD= 1.088) over SpeakonTouch (M= 3.00, SD=0.894), t(15)= -3.297, p= 0.005. The t-tests also revealed that Spatial SpeakonTouch (M= 2.13, SD= 1.025) was preferred over Spatial Language (M= 3.00, SD=0.894), t(15)= -2.267, p= 0.039 and over SpeakonTouch(M= 3.00, SD=0.894), t(15)= -2.573, p= 0.021.

5.3.5 Participant Comments

The participants were given an opportunity at the end of the experiments to convey their thoughts on the interfaces. Most participants utilized this chance and gave comments. Some of them are listed below:

P5: "I like the spatial audio mode as I don't have to do anything to learn about the targets."

P6: "I liked the mode in which I could touch with my finger to learn target and hear directional sound at the same time"

P7: "Once I learned how to use the touch modes [SpeakOnTouch and Spatial SpeakonTouch], learning became easy. I used my thumbs to better locate targets as sometimes finger did not pick the target. I like that I didn't have to do anything in 3D audio mode [auditory snapshot]."

P8: "The directional audio helps" [referring to spatial SpeakonTouch]

P10: "I had a difficult time finding objects using touch"

P13: "It was difficult to judge angles in 3D audio. It might be difficult to give information about objects at your back using degrees [referring to Spatial Language mode].

P14: "3D audio was very intuitive"

P15 "In the touch modes [SpeakOnTouch and Spatial SpeakonTouch] practice will help. I do not use degrees in everyday life, so I had difficulty in [spatial] language.

5.4 Discussion and Conclusion

The main aim of this chapter was to compare how the kinesthetic interface would fare against another perceptual interface (3 D audio) and a non-perceptual interface (spatial language). Second, I wanted to explore if providing 3D audio with proprioceptive information improves the ability of the participants to form spatial representations. Finally, I wished to implement these interfaces on an actual smartphone device.

Section 5.1 discussed how the kinesthetic cues are perceptual. We also reviewed some previous literature which utilizes these cues to impart spatial information through non-visual modalities. Through the literature, we found that touch screen based kinesthetic interfaces can provide spatial knowledge through vibration and audio.

In section 5.2 I introduced SpeakOnTouch and Spatial SpeakOnTouch modes. While the former provides spatial information solely through proprioceptive prompts, the latter combines proprioceptive information with spatialized directional audio cues. In this section I also discussed how the auditory snapshot and spatial language modes were implemented on the smartphone. Finally, I described the methodology for our study to test these perceptual interfaces with spatial language.

In section 5.3 I discussed the results of the study. For the number of trials needed to achieve criterion, no significant results were found. The auditory snapshot mode had a lower mean for the number of trials as compared to spatial language. For all participants who needed more than one trial for the learning criterion (except one), distance rank criterion was the reason for them failing the test. This means that while the participants were able to obtain the angular information in degrees in spatial language mode, they had

difficulties in remembering the distances. I also found that participants provided better allocentric judgments with 3D audio a perceptual mode as compared to spatial language. This might be because of the problem of recalling the spatial representation through a non-perceptual interface (spatial language). However, no significant differences were found in allocentric pointing between SpeakOnTouch and spatial SpeakOnTouch modes. No significant differences were found in pointing latencies indicating that the participants took almost the same time to recall the spatial image in all the four modalities.

To measure cognitive load in learning through these modes, we introduced the NASA-TLX multi-dimensional rating test. No significant differences in the six-subscales were observed, indicating similar work load across modalities. However analysis of subjective ratings by the participants demonstrated that the 3D audio mode was preferred over spatial language and SpeakOnTouch. Also Spatial SpeakOnTouch fared better in user ratings than vanilla SpeakOnTouch. The participants' comments also revealed that directional audio as a redundant cue helps in remembering the spatial layout.

Many participants liked the auditory snapshot mode as it provides a global view of the scene without any physical effort and with minimal cognitive effort. However, in the future we would like to provide the participant with a control on each object utterance. This can be implemented using gestures where each swipe would lead to playing of the next object.

Since in the auditory snapshot mode the audio signals were based on non-individualized HRTF's (Chapter 2), there were some issues in localization for some participants. A

future prospect would be to include some kind of head or hand tracking mechanism that will improve the sound localization.

One of the characteristics of the two new modes I described in this chapter (SpeakOnTouch and Spatial SpeakOnTouch) was that they were completely based on audio cues. Further research should test if adding vibration cues on touch would help the users better localize the location of the objects on the screen.

In sum, this study not only provided an opportunity to test the modes described in chapter 3, but also allowed us to implement and test two kinesthetic interfaces. Even though the results from this study point toward the superiority of perceptual interfaces over non-perceptual interfaces in helping individuals to form spatial representations through non-visual modalities, in general all the four smartphone based modes led to very high performance (Section 5.3). This result is very exciting as it means that we can overcome the problems of non-refreshable, expensive and inflexible assistive technology systems by further development of smartphone based interfaces that are inexpensive, portable, dynamic (that is they allow refreshable information) and support universal design principles.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

The ultimate goal of navigation, whether visual or non-visual, is the same- to travel safely and efficiently from one place to another. However non-visual navigation is considerably more difficult as it does not afford the cues offered by vision, which are critical for quick and easy travel (Giudice & Legge, 2008). Likewise, successful blind navigation requires two main components, (1) Access to the cognitive map of the space, and (2) Avoidance of obstacles while navigating. For centuries guide dogs and canes have been the principal assistive devices for the blind and low-vision community to help them navigate in both outdoor and indoor environments. While highly effective, both of these tools tackle only the second component required for effective navigation by helping blind people in avoiding obstacles in their path. To solve the problems in navigation associated with building up of an accurate cognitive map, several techniques have been used in the past. (Giudice & Legge, 2008; Thinus-Blanc & Gaunet, 1997) provide a good review of these techniques.

The problem with most of these techniques is that they run on custom hardware, are for a single purpose or use, or are comparatively expensive (see Section 2.1 and 2.3 for review). The convergence of smartphone technology offers a solution to these problems owing to their embedded sensors; customizable interfaces and relatively cheap cost (see Chapter 1 for a discussion). The aim of this thesis research was to evaluate the efficacy of audio based perceptual and non-perceptual interfaces in supporting spatial behaviors and in the development of cognitive maps without vision. These interfaces could then be

implemented on off the shelf smartphone devices to aid the blind and low-vision community in learning unfamiliar environments.

I summarize the main findings from this thesis and provide conclusions of this research in section 6.1. In section 6.2, I discuss some issues and future work to extend the current research.

6.1 Conclusions and General Discussion

I started this thesis by describing my five main research questions (see Section 1.1). In this concluding section I summarize how our research led to the answers of these questions

RQ1: Can audio be used as an alternative to vision to help blind individuals in forming accurate cognitive maps of indoor spatial layouts?

Yes. To answer this research question I performed a series of three behavioral experiments. Experiment 1 compared cognitive map development through four different audio modes rendered through a virtual reality system. Blindfolded Participants learned the scene from an origin via the four "audio only" modes and were asked to walk to the individual targets from a new drop off point. To perform this task correctly, participants needed to form a cognitive map of the scene and use accurate spatial updating of this representation. The overall data analysis showed that the participants performed this task with considerable accuracy taking nominal response times (see Section 3.3) which was interpreted as showing that they were successful in forming a cognitive map of the scene without vision and by the use of audio as the sole modality. In general, these findings are crucial for the development of accessible interfaces as they suggest that spatial

knowledge acquisition is possible by substituting visual information with audio information.

In experiment 2, we compared two spatial audio modes (based on head or hand motion) with vision as a control. In order to eliminate the spatial advantage for vision (in a way that all targets could be seen at the same time), we used a technique so as to offer visual access to only one target at a time. Also, instead of walking from a new drop off point, participants walked from one target to another. They also performed a polygon walking task in which they walked to each of the targets one after the other in clockwise or counter-clockwise direction. The goal of this task was to evaluate if the participants had learned the global spatial layout of the scene and not just individual locations. This is because in order to walk from one target to another the participants needed to know the relation between the two objects. Similarly to perform the polygon walking task, the participants needed access to the global cognitive map of the space. Except for the target to response distance parameter in which the visual condition was significantly better than audio conditions, we found no significant differences in spatial updating performance (in terms of response times, thinking times or absolute distance errors) of blindfolded participants between the visual and two audio conditions. These results are interpreted as showing that our 3D audio modes based on head or hand motion afforded similar spatial updating and cognitive map development as was built up from vision.

These results suggest development of functionally equivalent spatial representations for vision and "3D audio only" modes. This means that vision offered no significant advantage in terms of target angle and distance perception. No significant differences in the thinking times and response times mean that participants took equally long duration

in accessing the spatial representation independent of the modality of acquisition. This functional equivalence between the two perceptually driven interfaces is exciting as it means spatial audio based modes can substitute for vision and thus be used effectively in portable devices affording environmental access and navigation assistance.

These results are in agreement with previous research (Klatzky et al., 2002) where the authors found no inherent advantage of learning through vision as compared to spatial audio where vision was constrained to "content only" in order to eliminate modality specific cues as we did in our study.

Finally, in Experiment 3 the participants learned the spatial array with the help of three perceptual and one non-perceptual interface implemented on a smartphone. We found that blindfolded participants were able to perform self to object (egocentric) pointing judgments (Section 5.2.5.1 learning criterion) with considerable accuracy. They were also able to perform object to object (allocentric) pointing judgments by learning a spatial layout through audio modes implemented on a smartphone accurately. To perform these tasks participants need to build a cognitive map of the spatial representation. This again means that they were able to form a global spatial representation of the objects, while blindfolded and with the use of only audio as a modality.

Thus, the findings from the three experiments provide compelling evidence that it is possible for users to learn spatial layouts, update the targets in these layouts in both egocentric and allocentric tasks and form accurate cognitive maps of spaces with an "audio only" modality as the input. This result advocates the use of audio interfaces for spatial learning systems for the blind.

RQ2: Are there differences in audio based perceptual interfaces in terms of speed and accuracy of mental representations formation?

I evaluated different audio based perceptual interfaces in this thesis through the 1st and 3rd experiments. In my first experiment I compared 3D audio interfaces with hand motion and head motion triggered interfaces. We found that participants walked faster and more accurately to targets when their location was specified through 3D audio and hand motion triggered mode as compared to the head motion triggered mode. The former perceptual interfaces can be readily implemented on off-the shelf smartphone devices as compared to the head motion triggered interface which is difficult to implement on these devices without adding additional head trackers (See chapter 1). This is an important result as it favors the performance of perceptual interfaces which have the potential to be implemented on smartphone devices without the need of any extra setup, thus reducing the cost of these interfaces.

In experiment 3 we implemented two novel kinesthetic interfaces on a smartphone device for this thesis research. These interfaces are called SpeakOnTouch and Spatial SpeakOnTouch (section 5.2.1. SpeakonTouch conveyed speech based spatial information whenever the user touched the map on the device's screen. Spatial SpeakOnTouch also delivered speech based spatial information upon touching the map area, but it spatialized the speech output so that the sound appeared to come from the spatial location of the object in the real world. We compared these novel interfaces with auditory snapshot based on 3D audio and found no significant differences between the three perceptual

audio interfaces, which we interpreted as showing that the three modes led to formation of similar spatial images.

While the spatial SpeakOnTouch did not offer any added advantage over SpeakOnTouch in the pointing tasks, the subjective ratings of the two modes suggest the spatial SpeakOnTouch was preferred significantly more than vanilla SpeakOnTouch. Many participants reported that they enjoyed the redundancy of spatial cues offered in the latter mode. Further studies comparing the two modes needs to be done to establish the advantage (if any) of redundant cues in this mode.

RQ3: Are there differences in perceptual and non-perceptual interfaces in terms of speed and accuracy of mental representations formation?

Even though it is the easiest non-visual mode to generate and is the gold standard in providing directional information non-visually to blind as well as sighted users (e.g. in-car navigation systems, pedestrian navigation systems etc.), I argue against the use of spatial language because of the added cognitive demand it entails. The users of the spatial learning system may already have some inherent mental load (See Chapter 3 introduction for sample scenarios). The cognitive arbitration involved in this mode makes its use unfavorable for situations when a blind person is in constant interaction with the world and has additional cognitive load beyond interpreting the linguistic output from the interface.

In experiment 1, I compared spatial updating performance of spatial language with three other perceptual interfaces based on 3D audio, hand triggered and head triggered modes. To my surprise, I found that participants incurred significantly less absolute angle error

with spatial language as compared to the head motion triggered mode. We used spatial language as a control condition and therefore administered it at the last for each trial (section 3). In hindsight, we realize that this methodological decision may well have led to an artificially elevated level of spatial learning performance by the participants as compared to the head motion triggered interface. Also, since our experiment did not involve any additional cognitive load for the participants, the spatial updating performance may not have suffered due to cognitive arbitration of the non-perceptual mode. Future research should investigate if the results vary when an additional cognitive load is introduced.

In experiment 3, where we included spatial language as one of the conditions in a counterbalanced manner, we found that the participants performed more incurred more allocentric judgment errors when they learned through spatial language as compared to when they learned using 3D audio. These results from our third study indicate that the spatial audio based auditory snapshot mode, which is a perceptual interface, allowed the users to build a more accurate cognitive map. This advocates the use of a perceptual interface in non-visual spatial learning systems.

In sum, even though conventional spatial language based non-perceptual interfaces are easy to implement in a spatial learning system, I argue for the use of perceptual audio interfaces such as auditory snapshot as they support more accurate mental representations than the former non-perceptual interface based on the results of experiment 3.

Spatial language interfaces can be useful when the spatial information being conveyed not be precise. In certain instances, it is not necessary to impart exact information about the location of an object in the room. For example, suppose the user wants to learn the approximate location of a window in the room. The spatial information about the window can be imparted to the user in the following manner: "window located on the right". Even though this language based approach underspecifies the location of the target, it is successful in providing a coarse description of the space, which in this case is sufficient. The user requires minimal effort to comprehend the space. Therefore the spatial language approach is suitable when precise locations of the objects are not required.

RQ4: Can head tracking be replaced with hand tracking to generate more immersive spatial audio?

Yes. This research question relates to the implementation of 3D audio interfaces on smartphone device. We described why head tracking is essential for 3D audio generation and how it can be substituted with hand tracking if the two techniques lead to development of similar cognitive maps in Chapter 4.

To answer this question we compared spatial updating performance of the participants when they learned a spatial scene through either of two modes in Experiment 2. We found no significant differences in the spatial learning and updating performances by the participants in the two modes (Section 4.3). This means the participants formed comparable spatial images through the two modalities. This result is extremely useful as it means we can develop immersive spatial audio applications on off-the shelf

smartphone devices by replacing the use of expensive and aesthetically unpleasant head trackers with hand tracking obtained through smartphone sensors.

RQ5: Are there users' preference differences with respect to effectiveness and usability of interfaces tested in this thesis?

To evaluate the interfaces discussed in this thesis, I administered two surveys of participants to better understand their preferences in Experiment 1 and Experiment 3. Although these results from blindfolded participants may not be as powerful as they would have been if the participants were blind or visually impaired, we argue the validity of these results on the premise that the blindfolded participants had access to the same perceptual cues as a blind person would have had in the same conditions (see section 3.2.1 for further explanation of this design choice). While experiment 1 yielded no significant differences in the ratings for perceptual and non-perceptual interfaces, results from experiment 3 showed that the participants favored the 3D audio mode, a perceptual interface, over spatial language, a non- perceptual interface. Also subjective comments described in section 5.3.5 suggest that participants liked the intuitiveness of the 3D audio mode and thus gave it considerably higher ratings than spatial language. This further gives weight to our hypotheses that since perceptual interfaces exert no additional cognitive load, they offer means to learn a spatial layout in a more natural way. Another reason for the higher subjective ratings for 3D audio might be that it does not require any additional thought to understand the interface. Spatial sound is a natural formula observed in our daily lives and is experienced by all people with normal hearing irrespective of their experiences (We discussed human sound localization in section 2.2)

as compared to spatial language which is an artificial mechanism devised by humans as a mean to convey spatial information.

6.2 Some Issues and Future Directions

As discussed in this thesis, the main motivation of the research was to investigate the usefulness of various audio interfaces in the development of cognitive maps in blind and low vision users. There are a number of issues which should be further investigated in future research. We discuss these issues in this section.

1.   Using Speech based Interfaces

This research focused on imparting spatial information through the use of speech based audio. Speech based interfaces are already pervasive in outdoor navigation systems. For example the GPS based in car navigation systems use speech as primary modality in conveying spatial information. Also many blind users are already proficient in speech based output modalities through their use of computer screen reading software such as Jaws (Freedom Scientific, St. Petersburg, FL).

However there are two potential disadvantages of using speech as the primary audio modality

a)   Speech intelligibility can be drastically reduced in the presence of noise in the environment (Miller, 1947; Rhebergen & Versfeld, 2005). We assume that the spatial learning system would be used in noisy environments such as shopping malls, office spaces etc. Therefore using only speech based audio may be a problem in these high ambient noise environments.

b) Speech signals are known to be harder to localize than non-speech signals (Tran, Letowski, & Abouchacra, 2000). This is because most of the speech signals lie in a low frequency range (less than 6 kHz) and thus they lack the high frequency components which are important for sound localization (Gilkey & Anderson, 1995). This also makes the determination of the elevation of the sound source difficult. One of our proposed interfaces, namely, the auditory snapshot depends on 3D audio, and since we rely on speech to convey target information it might be difficult for some users to localize the objects correctly.

Future research should try to investigate methods to counter these problems. Methods to increase speech intelligibility should be explored as in (Brungart & Simpson, 2005; MacDonald, Balakrishnan, Orosz, & Karplus, 2002). To increase sound localizability high frequency non-speech sounds such as bursts of pink noise may be padded before the speech signal as they have been found to be easy to localize (Walker & Lindsay, 2006).

2.   Distance Perception

Two of our modes, namely, Auditory snapshot and Spatial SpeakOnTouch rely on 3D audio. In this research we provided distance information for all the modes with the use of speech, for example 8 feet, 4 feet etc. Past research has shown distance perception in virtual audio is a difficult feat to achieve and is quite inaccurate, mostly due to distance compression (Zahorik, 2002). However, to provide more immersive experience, future systems should embed distance information in the sound signal eliminating the need of explicit distance information in 3D audio. Therefore there is a need for future research to find ways for more accurate virtual audio distance perception.

3.   Smartphone Sensors

To be able to provide a more immersive experience to the user, the hand of the user must be tracked (see Chapter 4) in order to provide important interaural cues. Even though the smartphones sensors have improved quite a bit in recent years, they still lack the required accuracy levels to be able to reliably provide orientation information to the user in indoor environments (Ogundipe, 2012; Ozcan et al., 2012; Rodriguez, 2011; H. Wang et al., 2012). However, with improvement in electronics the accuracy of these sensors is also predicted to be improved. Future studies should further investigate the applicability of hand motion tracked spatial audio on real devices.

4.   Headphones

All our experiments were conducted using over the ear headphones. These headphones provided the participants with high sound quality during the experiments. However, the use of these headphones is not advisable in the real world, as they tend to block extremely important environmental sounds which are vital for blind navigation. Also, they are visually obtrusive making their use in the real world even more limited. In a survey on the preferred components for blind navigation described in (Golledge et al., 2004) found that most blind individuals preferred collar or shoulder mounted speakers and rated over the ear headphones amongst the least desirable components.

One Solution to this problem can be the use of Bone-conduction head phones which provide conduction of sound to the inner ear through the bones of the skull, thereby leaving the outer ear open to environmental sounds. A study described in (Walker & Lindsay, 2005) found that the participants were able to navigate with good efficiencies

with the use of these bonephones in the SWAN (Wilson et al., 2007) system, though the performance was not as efficient as with over the ear headphones. Future research should empirically evaluate the spatial learning performance with audio interfaces using bonephones instead of conventional headphones.

We expect that the research described in this thesis is an important step in the direction of the development of smartphone based spatial information systems, based on perceptually driven audio interfaces. We expect these systems to be most beneficial to the blind and low-vision community. However, the spatial audio modes described throughout this report could also be extremely useful to the sighted community in learning new spaces when visual attention to the screen is not possible or desirable, for example 1) When the user is navigating in a museum and wants to know the points of interest around them, 2) When the user is in a car and cannot (should not) take their eyes off the road, 3) Soldiers walking in the dark, 4) Firefighters in a building full of smoke, etc.

BIBLIOGRAPHY

ABTIM, Wuppertal Germany. (2012). Retrieved from
http://www.abtim.com/home__e_/home__e_.html

Algazi, V. R., Duda, R. O., Thompson, D. M., & Avendano, C. (2001). The CIPIC HRTF database. *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)* (pp. 99–102). IEEE. doi:10.1109/ASPAA.2001.969552

Arduino. (2012.). Retrieved from http://arduino.cc/

Ashmead, D. H., Davis, D. L., & Northington, A. (1995). Contribution of listeners' approaching motion to auditory distance perception. *Journal of Experimental Psychology: Human Perception and Performance, 21*(2), 239–256. doi:10.1037/0096-1523.21.2.239

At & T Labs, Inc. (2010). Retrieved from
http://www2.research.att.com/~ttsweb/tts/demo.php

Audacity. (2010). Retrieved from http://audacity.sourceforge.net/

Azuma, R. T. (1995). *Predictive tracking for augmented reality*. Chapel Hill North Carolina.

Begault, D. R. (1994). *3-D sound for virtual reality and multimedia*. New York, New York, USA: Academic Press. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.20.8443&rep=rep1&type=pdf

Behringer, R., Tam, C., McGee, J., Sundareswaran, S., & Vassiliou, M. (2000). A wearable augmented reality testbed for navigation and control, built solely with commercial-off-the-shelf (COTS) hardware. *Proceedings IEEE and ACM International Symposium on Augmented Reality (ISAR 2000)*, (Isar), 12–19. doi:10.1109/ISAR.2000.880918

Bronkhorst, A. W. (1995). Localization of real and virtual sound sources. *The Journal of the Acoustical Society of America, 98*(5), 2542. doi:10.1121/1.413219

Brungart, D. S., & Simpson, B. D. (2005). Optimizing the spatial configuration of a seven-talker speech display. *ACM Transactions on Applied Perception, 2*(4), 430–436. doi:10.1145/1101530.1101538

Caruso, M., & Bratland, T. (1998). A new perspective on magnetic field sensing. *Sensors Expo Proceedings* (pp. 195–213). Retrieved from http://www51.honeywell.com/aero/common/documents/myaerospacecatalog-documents/Defense_Brochures-documents/Magnetic__Literature_Technical_Article-documents/A_New_Perspective_on_Magnetic_Field_Sensing.pdf

Christoph, P. (2007). 3-D Audio in Mobile Communication Devices : Methods for Mobile Head-Tracking, *4*(13).

Daunys, G., & Lauruska, V. (2007). Sonification system of maps for blind. *Universal Access in Human-Computer Interaction. Ambient Interaction, 4555/2007*, 349–352. doi:10.1007/978-3-540-73281-5_37

Dedes, G., & Dempster, A. G. (2005). Indoor GPS Positioning Challenges and Opportunities. *Vehicular Technology Conference, Vol 1* (pp. 412–415).

Dementhon, D. F., & Davis, L. S. (1995). Model-based object pose in 25 lines of code. *International Journal of Computer Vision, 15*(1-2), 123–141. doi:10.1007/BF01450852

Easton, R. D., & Sholl, M. J. (1995). Object-array structure, frames of reference, and retrieval of spatial knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(2), 483–500. doi:10.1037/0278-7393.21.2.483

El-Natour, H. A., Escher, A.-C., Macabiau, C., & Boucheret, M.-L. (2005). Impact of Multipath and Cross-Correlation on GPS Acquisition in Indoor Environments. *Proceedings of the 2005 National Technical Meeting of The Institute of Navigation, San Diego, CA* (pp. 1062–1070). San Diego, CA. Retrieved from http://luci.ics.uci.edu/predeployment/websiteContent/weAreLuci/biographies/faculty/djp3/LocalCopy/098_E2-1.pdf

FMOD. (2011). Melbourne: Firelight Technologies. Retrieved from http://www.fmod.org/

Ferguson, E. L., & Hegarty, M. (1994). Properties of cognitive maps constructed from texts. *Memory & Cognition, 22*(4), 455–473. doi:10.3758/BF03200870

Freedom Scientific, St. Petersburg, FL. (2012). Retrieved from http://www.freedomscientific.com/about/about.asp

Gilkey, R. H., & Anderson, T. R. (1995). The accuracy of absolute localization judgments for speech stimuli. *Journal of vestibular research : equilibrium & orientation, 5*(6), 487–97. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8589858

Giudice, N. A., Bakdash, J. Z., Legge, G. E., & Roy, R. (2010). Spatial learning and navigation using a virtual verbal display. *ACM Transactions on Applied Perception, 7*(1), 1–22. doi:10.1145/1658349.1658352

Giudice, N. A., Betty, M. R., & Loomis, J. M. (2011). Functional equivalence of spatial images from touch and vision: evidence from spatial updating in blind and sighted individuals. *Journal of experimental psychology. Learning, memory, and cognition, 37*(3), 621–34. doi:10.1037/a0022331

Giudice, N. A., & Legge, G. E. (2008). *The Engineering Handbook of Smart Technology for Aging, Disability, and Independence.* (A. S. Helal, M. Mokhtari, & B. Abdulrazak, Eds.)... *Handbook of Smart Technology for* ... (pp. 479–500). Hoboken, NJ, USA: John Wiley & Sons, Inc. doi:10.1002/9780470379424

Giudice, N. A., Walton, L. A., & Worboys, M. (2010). The informatics of indoor and outdoor space. *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness - ISA '10* (p. 47). New York, New York, USA: ACM Press. doi:10.1145/1865885.1865897

Golledge, R. G. (1991). Tactual strip maps as navigational aids. *Journal of Visual Impairment & Blindness, 85*(7), 296–301.

Golledge, R. G., Marston, J. R., Loomis, J. M., & Klatzky, R. L. (2004). Stated Preferences for Components of a Personal Guidance System for Nonvisual Navigation. *The Journal of Visual Impairments and Blindness, 98*(3), 135–147.

Good, M. D. (1996). Sound localization in noise: The effect of signal-to-noise ratio. *The Journal of the Acoustical Society of America, 99*(2), 1108. doi:10.1121/1.415233

Google Maps. (2012). Retrieved from http://www.maps.google.com

Griesinger, D. (1999). Objective Measures of Spatiousness and Envelopment. *16th International Conference: Spatial Sound Reproduction.*

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In P. A. Hancock & N. Meshkati (Eds.), *Human Mental Workload* (pp. 139–183). Amsterdam: Elsevier. Retrieved from http://humansystems.arc.nasa.gov/groups/TLX/downloads/NASA-TLXChapter.pdf

Hasser, C. J., & Roark, M. R. (1998). Tactile Graphics Display.

Heinroth, T., & Buhler, D. (2008). Arrigator — evaluation of a speech-based pedestrian navigation system. *Intelligent Environments, 2008 IET 4th International Conference* (pp. 1–4).

Ho, C., & Spence, C. (2005). Assessing the effectiveness of various auditory cues in capturing a driver's visual attention. *Journal of experimental psychology. Applied, 11*(3), 157–74. doi:10.1037/1076-898X.11.3.157

Holllns, M., & Kelley, E. K. (1988). Spatial updating in blind and sighted people. *Perception & Psychophysics, 43*(4), 380–388. doi:10.3758/BF03208809

Intersense , LLC Billerica, Massachusetts. (2012).

Kahana, Y., Nelson, P. A., Petyt, M., & Choi, S. (1999). Numerical Modelling of the Transfer Functions of a Dummy-Head and of the External Ear. *16th International Conference: Spatial Sound Reproduction.*

Kammermeier, P., Buss, M., & Schmidt, G. (2000). Dynamic display of distributed tactile shape information by a prototypical actuator array. *Proceedings. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2000) (Cat. No.00CH37113)* (Vol. 2, pp. 1119–1124). IEEE. doi:10.1109/IROS.2000.893169

Klatzky, R. L., Lippa, Y., Loomis, J. M., & Golledge, R. G. (2002). Learning directions of objects specified by vision, spatial audition, or auditory spatial language. *Learning & memory (Cold Spring Harbor, N.Y.), 9*(6), 364–7. doi:10.1101/lm.51702

Klatzky, R. L., Marston, J. R., Giudice, N. A., Golledge, R. G., & Loomis, J. M. (2006). Cognitive load of navigating without vision when guided by virtual sound versus spatial language. *Journal of experimental psychology. Applied, 12*(4), 223–32. doi:10.1037/1076-898X.12.4.223

Klepeis, N. E., Nelson, W. C., Ott, W. R., Robinson, J. P., Tsang, a M., Switzer, P., Behar, J. V., et al. (2001). The National Human Activity Pattern Survey (NHAPS): a resource for assessing exposure to environmental pollutants. *Journal of exposure analysis and environmental epidemiology, 11*(3), 231–52. doi:10.1038/sj.jea.7500165

Kramer, G., Walker, B., Coordinator, P., Bonebright, T., Cook, P., Flowers, J., Miner, N., et al. (1999). *Sonification Report : Status of the Field and Research Agenda.*

Kulhavy, R. W., Schwartz, N. H., & Shaha, S. H. (1983). Spatial Representation of Maps. *The American Journal of Psychology, 96*(3), 337–351. Retrieved from http://www.jstor.org/stable/1422316

Langendijk, E. H. A., Kistler, D. J., & Wightman, F. L. (2001). Sound localization in the presence of one or two distracters. *The Journal of the Acoustical Society of America, 109*(5), 2123. doi:10.1121/1.1356025

Linvill, J. G., & Bliss, J. C. (1966). A direct translation reading aid for the blind. *Proceedings of the IEEE, 54*(1), 40–51. doi:10.1109/PROC.1966.4572

Logitech, Morges, Switzerland. (2012). Retrieved from http://www.logitech.com/en-us/home

Loomis, J. M. (1985). *Digital map and navigation system for the visually impaired (white Paper)*. Santa Barbara.

Loomis, J. M., Golledge, R. G., & Klatzky, R. L. (1998). Navigation System for the Blind: Auditory Display Modes and Guidance. *Presence: Teleoperators and Virtual Environments, 7*(2), 193–203. doi:10.1162/105474698565677

Loomis, J. M., Lippa, Y., Golledge, R. G., & Klatzky, R. L. (2002). Spatial updating of locations specified by 3-d sound and spatial language. *Journal of experimental psychology. Learning, memory, and cognition, 28*(2), 335–45. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11911388

Loomis, J. M., Marston, J. R., Golledge, R. G., & Klatzky, R. L. (2005). Personal Guidance System for People with Visual Impairment: A Comparison of Spatial Displays for Route Guidance. *Journal of visual impairment & blindness, 99*(4), 219–232. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2801896&tool=pmcentrez&rendertype=abstract

MacDonald, J. A., Balakrishnan, J. D., Orosz, M. D., & Karplus, W. J. (2002). Intelligibility of Speech in a Virtual 3-D Environment. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 44*(2), 272–286. doi:10.1518/0018720024497934

Mackensen, P. (2004). *Auditive Localization. Head movements, an additional cue in Localization*. Technical University of Berlin.

Maenaka, K., & Shiozawa, T. (1994). A study of silicon angular rate sensors using anisotropic etching technology. *Sensors and Actuators A: Physical, 43*(1-3), 72–77. doi:10.1016/0924-4247(93)00668-T

Mahan, R. P. (1991). *Circular Statistical Methods: Applications in Spatial and Temporal Performance Analysis. United States Army Research Institute for the Behavioral and Social Sciences*. Georgia.

Makous, J. C., & Middlebrooks, J. C. (1990). Two-dimensional sound localization by human listeners. *The Journal of the Acoustical Society of America, 87*(5), 2188–200. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2348023

MapQuest. (2012). Retrieved from http://www.mapquest.com/

Maxi-Aid Inc. (2012). Loc-Dots. Farmingdale, NY: Maxi-Aids Inc. Retrieved from http://www.maxiaids.com/products/1265/Loc-Dots-Keyboard-Key-Location-Dots-Clear.html

McClendon, B. (2011). A new frontier for Google Maps: mapping the indoors. Retrieved from http://googleblog.blogspot.com/2011/11/new-frontier-for-google-maps-mapping.html

Melanson, M. (2010, September 9). Google Maps for Android Gets Turn-By-Turn Walking Directions, Satellite Imagery. *Read Write Web*. Retrieved from http://www.readwriteweb.com/archives/google_maps_for_android_gets_turn-by-turn_walking.php

Micello, Sunnyvale CA. (2012.). Retrieved from http://www.micello.com/

Miller, G. A. (1947). The masking of speech. *Psychological Bulletin, 44*(2), 105.

Mindfold, Inc. Tucson, AZ. (2012). Retrieved from http://www.mindfold.com/

Mohan, A., Duraiswami, R., Zotkin, D. N., DeMenthon, D., & Davis, L. S. (2003). Using computer vision to generate customized spatial audio. *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)* (p. III–57). IEEE. doi:10.1109/ICME.2003.1221247

Murakami, K. (2006, September 4). Too many people getting lost in new downtown library. *Seattle Post Intelligencer*. Seattle. Retrieved from http://www.seattlepi.com/local/article/Too-many-people-getting-lost-in-new-downtown-1213582.php

Murphy-Chutorian, E., & Trivedi, M. M. (2009). Head pose estimation in computer vision: a survey. *IEEE transactions on pattern analysis and machine intelligence, 31*(4), 607–26. doi:10.1109/TPAMI.2008.106

Naseh Hussaini, S. (2011). Mobile SoundAR. *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA '11* (p. 1777). New York, New York, USA: ACM Press. doi:10.1145/1979742.1979844

Nielson. (2012). America's new Mobile Majority: a look at Smartphone Owners in the US.

Nintendo Inc. (2012.). Retrieved from http://www.nintendo.com

Nuckols, B. (2009). Fewer than 10 percent of blind Americans can read Braille. *Missourian*.

Ogundipe, O. (2012, May). Assisting visually impaired using smart-phone sensors. *Coordinates Magzine*. Retrieved from http://mycoordinates.org/assisting-visually-impaired-using-smart-phone-sensors/

Ozcan, R., Fatih, O., Demirci, M. F., & Abul, O. (2012). An Adaptive Smoothing Method for Sensor Noise in Augmented Reality Applications on Smartphones. *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, 2012, 93*(6), 209–218. doi:10.1007/978-3-642-30607-5_19

Perrett, S., & Noble, W. (1997). The contribution of head motion cues to localization of low-pass noise. *Perception & psychophysics, 59*(7), 1018–26. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9360475

Point Inside Bellevue, WA. (2012).

Polhemus. (2012). Fastrak. Retrieved from http://www.polhemus.com/?page=Motion_Fastrak

Poppinga, B., Magnusson, C., Pielot, M., & Rassmus-Gröhn, K. (2011). TouchOver map. *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services - MobileHCI '11* (p. 545). New York, New York, USA: ACM Press. doi:10.1145/2037373.2037458

Pralong, D. (1996). The role of individualized headphone calibration for the generation of high fidelity virtual auditory space. *The Journal of the Acoustical Society of America, 100*(6), 3785. doi:10.1121/1.417337

Raaflaub, K. A., & Talbert, R. J. A. (2009). *Geography and Ethnography: Perceptions of the World in Pre-Modern Societies* (p. 147). John Wiley & Sons.

Raja, M. (2011). The Development and Validation of a New Smartphone Based Non-Visual Spatial Interface for Learning Indoor Layouts. *The University of Maine*, (Master's Thesis). Retrieved from http://130.111.64.156/theses/pdf/RajaM2011.pdf

Rayleigh, Lord. (1907). XII. On our perception of sound direction. *Philosophical Magazine Series 6, 13*(74), 214–232. doi:10.1080/14786440709463595

Rhebergen, K. S., & Versfeld, N. J. (2005). A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. *The Journal of the Acoustical Society of America, 117*(4), 2181. doi:10.1121/1.1861713

Rice, M. T., Jones, D., Golledge, R. G., & Jacobson, R. D. (2003). Progress in multimodal cartographic interfaces: Symbolization and examples. *Yearbook of the Association of Pacific Coast Geographers, 65*.

Rieser, J. J., Guth, D. A., & Hill, E. W. (1986). Sensitivity to perspective structure while walking without vision. *Perception, 15*(2), 173–188. doi:10.1068/p150173

Rodriguez, A. (2011). Indoor Positioning using Sensor-fusion in Android Devices, (September).

Roffler, S. K. (1968). Factors That Influence the Localization of Sound in the Vertical Plane. *The Journal of the Acoustical Society of America, 43*(6), 1255. doi:10.1121/1.1910976

Rolland, J., Baillot, Y., & Davis, L. (2001). A survey of tracking technology for virtual environments. *Fundamentals of wearable ...*, 1–48. Retrieved from http://books.google.com/books?hl=en&lr=&id=gLbWtOLersUC&oi=fnd&pg=PA67&dq=A+SURVEY+OF+TRACKING+TECHNOLOGY+FOR+VIRTUAL+ENVIRONMENTS&ots=lQQ7T6eobn&sig=NX3eQBuIG3Y_VqkmAjJYtyHgfuM

Samsung Electronics, Suwon, South Korea. (2012.). Retrieved from http://www.samsung.com

Sandvad, J. (1996). Dynamic Aspects of Auditory Virtual Environments. *Audio Engineering Society Convention 100*. Retrieved from http://www.aes.org/e-lib/browse.cfm?elib=7547

Schiller, J. S., Lucas, J. W., Ward, B. W., & Peregoy, J. A. (2012). Summary health statistics for U.S. adults: National Health Interview Survey, 2010. *Vital and health statistics. Series 10, Data from the National Health Survey*, (252), 1–207. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22834228

Schweikhardt, W., & Klöper. (1984). *Rechnerunterstützte aufbereitung von bildshirmtext-grafiken in eine tastbare darstellung Institut* (pp. 1–16).

Sennheiser Electronic Corporation. Lyme, CT. (2012). Retrieved from http://www.sennheiserusa.com/home

Sensable, Woburn, MA. (2012). Retrieved from http://www.sensable.com/haptic-phantom-desktop.htm

Shaw, E. A. (1974). The External Ear. *Handbook of sensory physiology, 5*(1), 450–490.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of experimental psychology. Human learning and memory, 6*(2), 174–215. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7373248

Strachan, S., Eslambolchilar, P., & Murray-Smith, R. (2005). GpsTunes: controlling navigation via audio feedback. *Proceedings of the 7th international conference on*

Human computer interaction with mobile devices services (2005), *111*(September), 275–278.

Strothotte, T., Fritz, S., Michel, R., Raab, A., Petrie, H., Johnson, V., Reichert, L., et al. (1996). Development of dialogue systems for a mobility aid for blind people. *Proceedings of the second annual ACM conference on Assistive technologies - Assets '96* (pp. 139–144). New York, New York, USA: ACM Press. doi:10.1145/228347.228369

Su, J., Rosenzweig, A., Goel, A., de Lara, E., & Truong, K. N. (2010). Timbremap. *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services - MobileHCI '10* (p. 17). New York, New York, USA: ACM Press. doi:10.1145/1851600.1851606

Sundareswaran, V., Wang, K., Chen, S., Behringer, R., McGee, J., Tam, C., & Zahorik, P. (2003). 3D audio augmented reality: implementation and experiments. *The Second IEEE and ACM International Symposium on Mixed and Augmented Reality, 2003. Proceedings.* (pp. 296–297). IEEE Comput. Soc. doi:10.1109/ISMAR.2003.1240728

Tatham, A. F. (1991). The design of tactile maps: theoretical and practical considerations. In K. Rybaczak & M. Blakemore (Eds.), *International Cartographic Association: Mapping the Nations* (pp. 157–166). London.

Tatham, A. F., & Dodds, A. (1988). *Proceedings of the Second International Symposium on Maps and Graphics for Visually Handicapped People: King's College, London.* King's College, London.

Thinus-Blanc, C., & Gaunet, F. (1997). Representation of space in blind persons: vision as a spatial sense? *Psychological bulletin, 121*(1), 20–42. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9064698

Thurlow, W. R., Mangels, J. W., & Runge, P. S. (1967). Head movements during sound localization. *The Journal of the Acoustical Society of America, 42*(2), 489–93. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6075942

Tran, T., Letowski, T., & Abouchacra, K. (2000). Evaluation of acoustic beacon characteristics for navigation tasks. *Ergonomics, 43*(6), 807–827. doi:10.1080/001401300404760

Ubilla, M., Domingo, M., & Cadiz, R. (2010). Head Tracking for 3d Audio using the Nintendo WII Remote. *International Computer Music Conference (ICM2010)* (pp. 1–8). Retrieved from http://web.ing.puc.cl/~dmery/Prints/Conferences/International/2010-ICMC2010-Ubilla.pdf

Vidal-Verdú, F., & Hafez, M. (2007). Graphical tactile displays for visually-impaired people. *IEEE transactions on neural systems and rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society, 15*(1), 119–30. doi:10.1109/TNSRE.2007.891375

WHO. (2012). Visual impairment and blindness. Retrieved July 22, 2012, from http://www.who.int/mediacentre/factsheets/fs282/en/index.html

Walker, B. N., & Lindsay, J. (2005). Navigation performance in a virtual environment with bonephones. *In Proc. of the Int'ernational Conference on Auditory Display (ICAD2005)* (pp. 260–263).

Walker, B. N., & Lindsay, J. (2006). Navigation performance with a virtual auditory display: effects of beacon sound, capture radius, and practice. *Human factors, 48*(2), 265–78. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16884048

Walker, B. N., & Mauney, L. M. (2010). Universal Design of Auditory Graphs. *ACM Transactions on Accessible Computing, 2*(3), 1–16. doi:10.1145/1714458.1714459

Wallach, H. (1940). The role of head movements and vestibular and visual cues in sound localization. *Journal of Experimental Psychology, 27*(4), 339–368. Retrieved from http://psycnet.apa.org/journals/xge/27/4/339/

Wang, H., Elgohary, A., & Choudhury, R. R. (2012). No Need to War-Drive : Unsupervised Indoor Localization, 197–210.

Wang, R. F. (2004). Between reality and imagination: when is spatial updating automatic? *Perception & psychophysics, 66*(1), 68–76. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15095941

Warren, N., Jones, M., Jones, S., & Bainbridge, D. (2005). Navigation via continuously adapted music. *CHI '05 extended abstracts on Human factors in computing systems - CHI '05*, 1849. doi:10.1145/1056808.1057038

Warusfel, O., & Eckel, G. (2004). LISTEN-Augmenting everyday environments through interactive soundscapes. *Virtual Reality for Public Consumption, IEEE Virtual Reality 2004 Workshop* (p. Vol 27). Chicago, IL.

Wenzel, E. M. (1993). Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America, 94*(1), 111. doi:10.1121/1.407089

Wenzel, E. M. (1996). What Perception Implies About Implementation of Interactive Virtual Acoustic Environments. *Audio Engineering Society Convention 101*. Retrieved from http://www.aes.org/e-lib/browse.cfm?elib=7426

Wightman, F.L., & Kistler, D. J. (1999). Resolution of front–back ambiguity in spatial hearing by listener and source movement. *The Journal of the Acoustical Society of America, 105*(5), 2841–2853. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Resolution+of+front-back+ambiguity+in+spatial+hearing+by+listener+and+source+movement#0

Wightman, Frederic L. (1989). Headphone simulation of free-field listening. I: Stimulus synthesis. *The Journal of the Acoustical Society of America, 85*(2), 858. doi:10.1121/1.397557

Wilson, J., Walker, B., Lindsay, J., Cambias, C., & Frank, D. (2007). Swan: System for wearable audio navigation. *Wearable Computers, 2007 11th IEEE International Symposium on* (pp. 91– 98).

WorldViz inc. (2010). Santa Barbra, CA. Retrieved from http://www.worldviz.com/

Wu, J.-R., Duh, C.-D., Ouhyoung, M., & Wu, J.-T. (1997). Head motion and latency compensation on localization of 3D sound in virtual reality. *Proceedings of the ACM symposium on Virtual reality software and technology - VRST '97*, 15–20. doi:10.1145/261135.261140

Yang, J. (2012). Smartphones in Use Surpass 1 Billion, Will Double by 2015. Retrieved from http://www.bloomberg.com/news/2012-10-17/smartphones-in-use-surpass-1-billion-will-double-by-2015.html

Yazdi, N., Ayazi, F., & Najafi, K. (1998). Micromachined inertial sensors. *Proceedings of the IEEE, 86*(8), 1640–1659. doi:10.1109/5.704269

Zahorik, P. (2002). Assessing auditory distance perception using virtual acoustics. *The Journal of the Acoustical Society of America, 111*(4), 1832. doi:10.1121/1.1458027

Zickuhr, K. (2012). Three-quarters of smartphone owners use location-based services their location with friends. Retrieved from http://www.pewinternet.org/~/media//Files/Reports/2012/PIP_Location_based_services_2012_Report.pdf
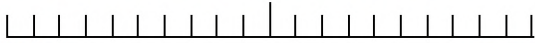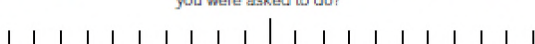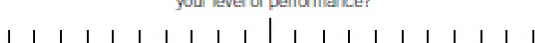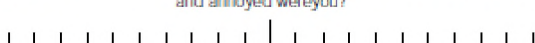
APPENDIX

NASA TLX

Adapted From

:Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In P. A. Hancock & N. Meshkati (Eds.), *Human Mental Workload* (pp. 139–183). Amsterdam: Elsevier. Retrieved from http://humansystems.arc.nasa.gov/groups/TLX/downloads/NASA-TLXChapter.pdf

Figure 8.6

NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name	Task	Date

Mental Demand How mentally demanding was the task?

Very Low ———————————— Very High

Physical Demand How physically demanding was the task?

Very Low ———————————— Very High

Temporal Demand How hurried or rushed was the pace of the task?

Very Low ———————————— Very High

Performance How successful were you in accomplishing what you were asked to do?

Perfect ———————————— Failure

Effort How hard did you have to work to accomplish your level of performance?

Very Low ———————————— Very High

Frustration How insecure, discouraged, irritated, stressed, and annoyed were you?

Very Low ———————————— Very High

BIOGRAPHY OF AUTHOR

Shreyans Jain was born in Ujjain (Madhya Pradesh), India on 2nd October 1988. He was raised in Ujjain and he graduated from St Mary's Convent School, Ujjain in 2006. He then attended Institute of Engineering and Science, Indore affiliated to Rajiv Gandhi Technological University, Bhopal, India and graduated in 2010 with a Bachelor of Engineering (Honors) in Computer Science and Engineering. He worked as a student intern at Infobeans Systems Pvt Ltd in Indore in summer of 2009. Shreyans enrolled as a Master of Science graduate student in the Department of Spatial Information Science and Engineering at The University of Maine in Fall 2010. Shreyans Jain is a candidate for the Master of Science Degree in Spatial Information Science and Engineering from The University of Maine in December 2012.