Enhancing Point Cloud Processing using Audio Cues

Thabo Ntsoko

Supervised by: Dr. George Sithole

Thesis submitted for degree of Master of Science in Geomatics

School of Architecture, Planning and Geomatics, Division of Geomatics

University of Cape Town

Acknowledgements

I would like to thank Prof. Heinz Ruther and the Zamani team for allowing me to use their data in doing this research. I'm also grateful to the South African National Space Agency (SANSA), for they have helped me financially this past year.

I'm grateful to the staff at the Geomatics Department, both core and support staff. The support shown by them to me has helped me through this journey. Special thanks also goes to Janice McMillan, Janet Small and Prof. Crain Soudien for all the support they have given me since day one. All the colleagues I have and had also deserve to be thanked, particularly Emmanuel Akrofi and Kevin Musungu.

I've thought about how to thank my supervisor, Dr. George Sithole, from the minute I started this project. Even at this stage I don't have the perfect words. What he has done for me goes beyond funding me and taking me under his wing to do this research. He believed in me and has seen me grow and work hard to get this done. Even when I felt I couldn't do this he encouraged me and was patient with me. He is the reason why I put in so much into this, for I was and still am inspired by him. I'll be forever grateful for what he has done for me and all the lessons he has taught me.

The decision to continue with my studies was not an easy one, as it affected my family. I'm grateful for the family that I have as they supported my decision. My mother has been with me from day one, with my father, brother and sister supporting me as well. I thank them and like to also thank my relatives as well. My love for them will always be there.

Acronyms and Technical Terms

1. *2D* – Two-dimensional.
2. *3D* – Three-dimensional.
3. *Anechoic* – Environment producing no echoes/reflections.
4. *API* – Application Programming Interface.
5. *Audio Context* – The rendering of audio through created sound sources.
6. *Aural* – Relating to hearing.
7. *Buffer Object* – An object for storing sample audio data.
8. *Free Field* – An environment where audio is listened without headphones.
9. *HF* – High frequency.
10. *LF* – Low frequency.
11. *Listener Object* – An object for receiving emitted sound.
12. *Octree* – A recursive and regular subdivision of 3D space.
13. *OpenAL* – Open Audio Library.
14. *Pinna* – Visible and external part of the ear.
15. *RHS* – Right Handed System.
16. *Sound Source Object* – An object for emitting sound.

Abstract

Today many airborne and terrestrial acquisitions capture point clouds of scenes or objects to be modelled. But before modelling can be done point clouds need to be taken through processing steps such as registration, cleaning, simplification, etc. These point clouds are usually manually processed before being processed automatically. Manual processing of point clouds depends on the visual interaction the user has with the point cloud provided by the visual cues.

This research investigated enhancing the level of interaction the user has with the point cloud when processing it. The proposed method augments audio in point clouds to enhance its processing where visual cues are limited. This investigated finding objects/points of interest in the point cloud while processing it by estimating the position (azimuth and elevation) and depth of audio objects associated with these point cloud objects. The occupancy of space of audio objects was also investigated to determine the unseen events around objects of interest in the point cloud.

For example, in a scan registration problem, audio could be augmented to a misaligned scan. As this scan is manually rotated and translated into alignment, various audio cues can be used to inform the user of the state of this alignment. An outlier separated from a surface in a point cloud could be identified and removed by augmenting audio to a volumetric brush that does the point cloud cleaning. Associating audio cues of the audio object with the depth of the outlier to the surface could help the user identify this outlier. Similar implementation could be adopted in point cloud simplification tasks.

Various audio cues exist which allow a listener to discern particular information about a sound source. This is done by the human auditory system, using cues such as intensity, pitch, reverberation and HRTFs to discern this information. However, limitations exist in retrieving this information.

Literature supports the use of the auditory interface in applications commonly built for the visual interface. The addition of the auditory interface is seen as a way of increasing the interaction users have with applications and therefore improving the experience. An auditory interface was built to help undertake this research. The test subject was immersed in the auditory environment by wearing headphones. This meant that the subject and the virtual listener were merged, allowing the subject to receive emitted audio. The perception of the audio was with respect to the virtual listener.

An auditory interface was created using OpenAL. OpenAL has a listener object which receives audio and audio objects which emit the audio. Audio data for each audio object is stored by a buffer object. Objects from an octree partitioned point clouds were associated with an audio object.

Through this interface, a sound emitting source was made to change its location and the tester required to estimate its position at each time. To do this position estimation the test subject had to click (with a mouse) on a response location displayed on the computer screen. A response location represented a location where the sound emitting source could be. If the sound emitting source was not where the clicked response location was then the estimation made would be incorrect. Position estimation was tested when the number of response locations increased and also when noise sources were introduced. The test results showed position estimation being affected by introduction of noise sources and the spatial distribution and the increase of the number of response locations. In the situation where there were no noise sources, absolute azimuth and elevation accuracies were 0.6° and 1°, respectively. Considering the distance of 1 m between the screen and the tester, these translate to x and y screen values of about 0.01 and 0.02 m, respectively. The case of two noise sources (both azimuth and elevation of 2°) translates to x and y values of about 0.04 and 0.04 m, respectively.

The binaural cues (inter-aural intensity difference (IID) and inter-aural time difference (ITD)) appeared to contribute the most in estimating the position of an emitted sound. The IID was the dominant cue in this regard. This means that to estimate the position of an audio augmented object in a point cloud, the user needs to pay attention to the intensity of the audio object. The differences in these intensities received by each ear could help estimate the position of the area of interest better.

A source was made to emit sound at various locations with the depth changing. Two sets of depths, each with three depth values were used. Tests were done for each set of depths where the depth of the sound source changed between the three depths of a set. The tester used the number keys of the keyboard to determine the depths in the tests. When the depth changed the intensity and the frequency of the emitted sound also changed, giving an indication to the tester that the depth has changed. The tester made depth estimations assisted by intensity and frequency cues.

A 100% depth estimation accuracy was obtained for tests in all instances.

However, this does not mean that depth estimations can be made error free. The reason for these accurate estimations is because the depth changes were coarse and were made so because making depth estimations is a crude exercise. To estimate the depths of audio augmented objects in point clouds, intensity and frequency can be relied on as depth cues. This is for depth changes that are coarse, however.

The occupancy of space was determined by putting a sound emitting source in different reverberant environments. In the open environment that was tested, mountainous environment, the accuracy of estimating if the virtual user was in this environment was 52.7%. In closed environments, cave, hallway and room environments, the accuracies were 33.3%, 35.3% and 35.3%, respectively. The manner in which the sound was reflected in each environment signalled the occupancy of space through the nature of the reflection. Occupancy of space was determined relatively well in some instances. The implications of this are that the type of reflections perceived, could alert the user of the existence of unseen events and their surroundings while processing a point cloud.

The findings in this study indicate that point cloud processing can be enhanced by augmenting the point cloud with audio. This would help the user find objects/points of interest in the point cloud where the visual cues are limited. This audio augmentation opens up other possibilities when processing point clouds, for example, implementing haptic interfaces in processing tasks.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Background to Study

The need to remotely capture spatial information of physical environments and objects in these environments using various sensors is common. The resulting product from sensors comes in the form of point clouds. Held (2012, pg. 7) defined a point cloud as "a set of vertices in a 3D coordinate system." To capture a surface in its entirety and represent it as a point cloud, numerous range measurements need to be made. Using the known direction of a range measurement, a point cloud such as the one in figure 1.1 can be obtained. In this example, a colour gradient indicates varying distance of points from a particular reference point.

Point clouds are useful in applications such as manufacturing, medicine, obstacle avoidance, geography, design, surveying, mobile mapping, cultural heritage, navigation and so on (Bosse *et al.*, 2012; Fabio, 2003; Linsen, 2001; Mahmoudi and Sapiro, 2009; Pauly *et al.*, 2002; Vosselman and Maas, 2010). Geometric models are usually produced from point clouds. These models are useful in a number of problem solving tasks, including applications such as rendering and simulation (Kolluri *et al.*, 2004).

Using point clouds for any problem solving task, often requires some processing of the raw point clouds (Vosselman *et al.*, 2004). According to Bucksch and Lindenbergh (2008), Kolluri *et al.* (2004), Linsen (2001), Mederos *et al.* (2003), Pauly *et al.* (2002), Song and Feng (2008) and Woo *et al.* (2002), the following are some of the processing that can be performed on point clouds to improve the effectiveness of their use:

– Removal of noise arising from sensor's measuring errors.

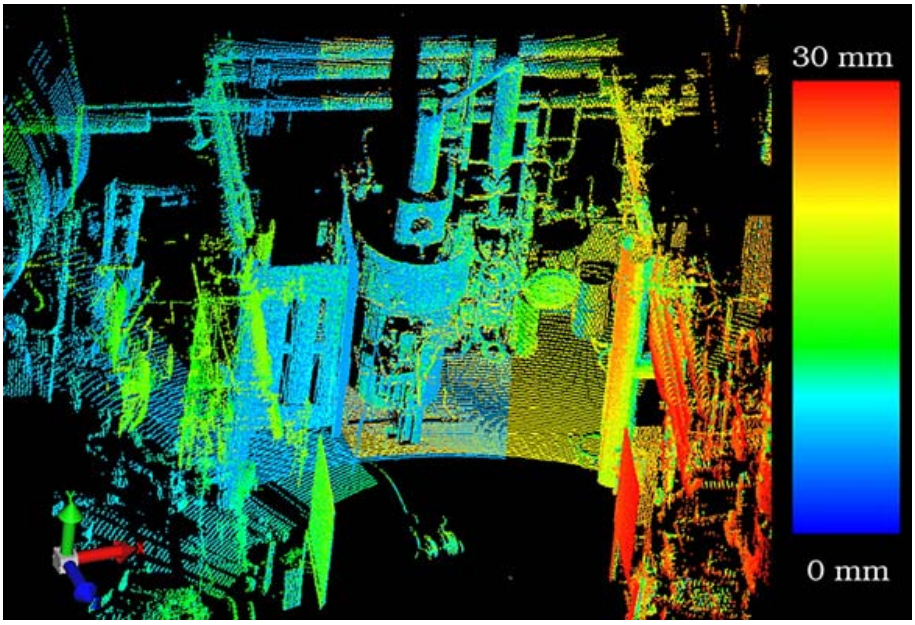


Figure 1.1: Captured point cloud (Rabbani *et al.*, 2007).

- Distortion removal by smoothing of the point cloud. This also arises from scanner's measuring errors.
- Simplification – this entails reducing the amount of data, i.e., down-sampling the point cloud.
- Modification and editing of represented objects.
- Scan registration

For processing and general interpretation tasks, the user needs to extract spatial information from the point cloud. In this context, extraction of spatial information refers to the determination of spatial locations of points/objects of interest in the point cloud.

For example, the user might want to locate areas in the point cloud which need to be smoothed out or down-sampled. The extraction of such information can be automatic, semi-automatic or manual. Due to the large size of point clouds, most processing will be automatic. However, manual or semi-automatic interventions are often be required to refine results from automatic processing.

Manual processing requires interaction with the point cloud and for now this interaction is visual. However, visual interaction has limitations. These limitations exist where the user can not see the necessary detail to carry out the point cloud processing. An example of this would be in aligning scans together when the user can not see how the scans are aligned at one time.

One way to overcome some of the limitations associated with visual processing is to augment the visual interaction with audio. This project investigates the augmentation of visual point cloud processing with audio and explores the limitations of this augmentation. The auditory system can help in acquiring detailed information about the surroundings, assisting in providing information in cases where the visual system is limited (Kapralos *et al.*, 2003).

The auditory system uses audio cues that allow humans to determine properties such as the direction and the depth to a sound emitting object. This is analogous to how the visual system uses visual cues to determine depth of environments and positions of objects within those environments.

As Kapralos *et al.* (2003) noted, the direction and depth perception of objects in the environment can be determined very accurately at times by the auditory system. Additionally, audio cues can help the auditory system determine the type of environment where sound is emitted and also properties such as the size, shape and texture of the sound source (Handel, 1989).

Auditory augmentation of a visual interface attempts to provide audio cues that inform the viewer/listener with hints of the location of objects. For audio cues to be useful they should reinforce the visual experience. Mereu and Kazman (1996) found that having an auditory interface can improve the experience for users of 3D computer applications.

1.2 Research Objective

The objective of this project is to investigate the augmentation of visual interfaces using audio cues and the determination of the limitations of interfaces augmented in this way. To support the objectives of the research the following questions are posed:

1. In audio augmented point cloud processing what are the limitations of estimating positions of points?

2. In audio augmented point cloud processing what are the limitations of estimating depths of points?

3. What are the limitations of judging the occupancy of space using audio cues?

4. How can an audio augmentation be implemented for point cloud processing?

Method of Research

The stated research objective is approached by first investigating various manual point cloud processing techniques and instances where the visual cues are limited in these techniques. The augmentation of audio in the processing is then suggested in those instances where the visual cues are limited. Specific audio cues that could potentially enhance the processing are discussed.

The relevant audio cues and the human auditory system are studied to gain an understanding of how audio can be augmented in point cloud processing. These cues are stated and discussed in terms of the spatial information they can help retrieve and how this is achieved. The audio cues are then tested in the context of augmenting audio in point cloud processing.

The results from the tests are stated and put into context of the research objective. Finally conclusions are drawn from the results and the implication of the results for audio augmented point cloud processing discussed.

1.3 Significance of the Study

Currently, point cloud processing software are not supported by audio augmentation. Therefore, building audio into visual interfaces has the potential to radically change the way operators interact with point clouds and the way operators experience and work with applications.

Incorporation of an auditory interface could potentially assist in extracting information efficiently and timeously. This process could help the user extract information that cannot be extracted by relying on the visual system alone.

Successfully using the auditory interface to extract spatial information of objects in point clouds could also lead to new developments in point cloud processing. For example, point cloud processing could be further augmented with haptic interfaces. There exist research projects where investigations focus on improving the interaction with the data to better understand it. Such can be seen in projects like the one by Lee *et al.* (2011), where tangible interfaces are created where users can "see, feel and control computations"(Lee *et al.* (2011, pg. 327)).

This study and its outcomes could also be beneficial to researchers and professionals in the photogrammetry and remote sensing fields. These users are constantly interacting with and processing point clouds and could benefit from a system that allows them to do so better.

1.4 Scope of the Research

The study focuses mainly on information extraction from point clouds. In particular, the study investigates the use of audio cues to determine the directions and distances/depths to objects from the user in a virtual environment. Occupancy of space of the objects is also be studied. Only spatial information of objects is investigated, other properties such as colour, size and texture of objects are not investigated.

The operation of the human auditory system is discussed, particularly the part that is relevant to achieving the goals of this study. Auditory cues that are responsible for extracting information as specified are outlined. The manners in which these cues extract this information are discussed, together with the accuracies and limitations that are associated with these cues.

This dissertation outlines how an auditory interface can be added to aid the visual interface in spatial information extraction from point clouds. This is outlined in the context of adding the auditory interface as an additional interface through computer software.

Augmentation of audio in various point cloud processing techniques is provided. Augmentation in these techniques is given to provide context and example of how an audio interface can be added in point cloud processing. This augmentation is therefore not only limited to the techniques discussed.

This study requires some knowledge of acoustics, defined by Lewis (2013) as "the study of physical properties of sound." The study therefore only focuses on acoustical properties and aspects which are relevant to the project.

1.5 Plan of Development

Chapter 2 focuses on audio augmentation in various point cloud processing problems. Literature of the point cloud processing techniques is given in this chapter. Following that, augmentation of audio in each processing technique is given. This includes equations and diagrams that demonstrate how the augmentation can be carried out.

In Chapter 3, the functionalities of the human auditory system are discussed. The manner in which human beings perceive sound/audio to make sense of the sound source is explained here. The type of information that different audio cues can help the listener extract about the sound source are outlined and discussed. Also in this chapter, the use of headphones as a manner of delivering sound to the user is reported.

Chapter 4 outlines how an audio augmented system can be implemented for workstations. The resources used in this implementation are mentioned. This includes hardware and software used.

Chapter 5 outlines the limitations of the audio augmented point cloud processing. The tests carried out to address the research questions are mentioned. Following that are explanations of how data was analysed.

Chapter 6 reports on the results obtained from the tests explained in Chapter 5. These results are analysed with the discussion focused on what they mean in the context of this research.

Chapter 7 brings this research to a conclusion. Here, conclusions are drawn and this is based on the findings from Chapter 6 and put into context of reviewed literature (Chapters 2 and 3). Recommendations are then be made based on these conclusions.

Chapter 2

Audio Augmentation in Point Cloud Processing

Point clouds are fed through various processing steps before 3D models are created out of them. These processing steps include, coarse (and fine) scan registration, data cleaning, simplification, surface fitting, smoothing and hole filling.

Presently, manual processing of point clouds is done via a visual interface. Visual interfaces are not without limitations. Limitations to visual interaction can be experienced where the user can not see necessary detail in the point cloud to do the processing. Augmenting visual interfaces with audio cues offers one way of overcoming the limitations of visual interfaces.

The audio augmentation entails associating objects of interest in the point cloud with audio objects. Using audio cues of audio emitting objects, spatial information such as the position, depth and occupancy of space of the audio objects can be estimated. This could in turn allow the user to retrieve spatial information of point cloud objects necessary when processing the point cloud. General audio augmentation in point cloud processing is explained with the aid of figure 2.1. Augmentation of audio for different processing techniques is discussed from section 2.1 to 2.3.

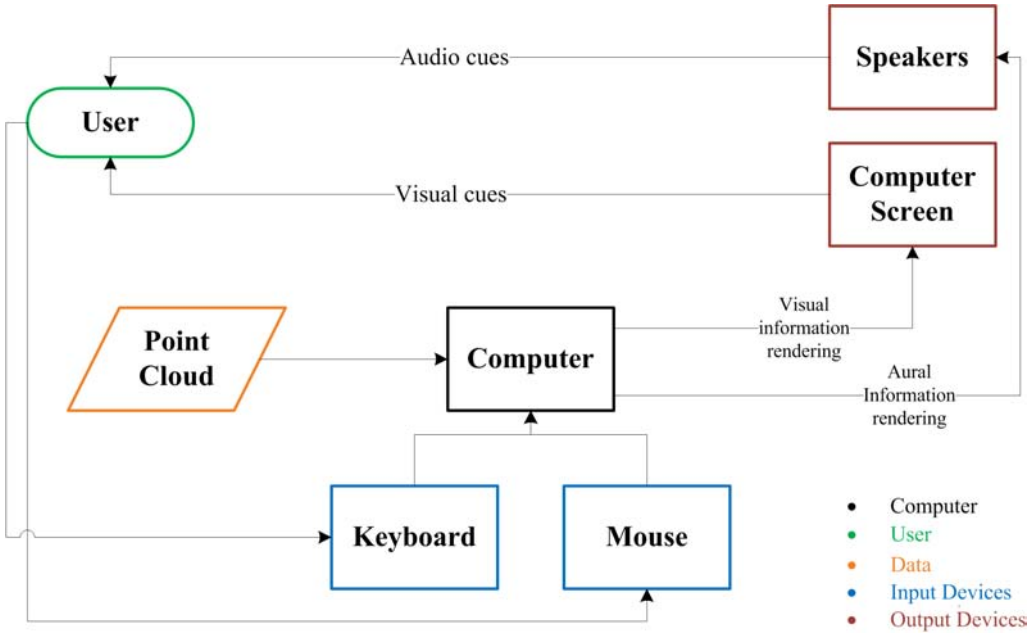


Figure 2.1: Audio augmentation in point cloud processing.

User

The *user* controls the entire manual processing task. During processing, the *user* interacts with the *computer* through the *input devices*. This interaction allows the *user* to receive processing output through the *output devices*. The *computer*, *input devices* and *output devices* entities are explained below.

Computer

The *computer* stores *point cloud* which needs to be processed. Using the *computer*, an **audio context** is created. The creation of an audio context entails connecting with the *computer's* audio device and opening it to render audio. The objects (e.g., point clouds) are rendered and displayed on a computer screen. Simultaneously the audio cues are also rendered and presented to the user through speakers. It is important that the audio cues generated complement the render imagery so that the *user* experience is enhanced and seamless.

Rendering the audio cues requires the objects (e.g., point clouds) be treated as emitters of sound. Audio objects emanate audio as required by the *user*.

The emanated audio is received by the *user*.

The next critical event in the *computer* is **audio augmentation**. Audio augmentation entails 'attaching' an audio object to an area of interest in the *point cloud*. Audio augmentation is dependent on the type of processing the *user* needs to perform and the specific augmentations will be explained in sections to follow.

Input Devices

There are two *input devices*, the **keyboard** and the **mouse**. These devices allow the *user* to do the manual processing by manipulating the data. The typical actions that can be done with these devices include, rotating and translating data, deleting and adding primitives to the data, etc. The data manipulation commands are sent directly to the *computer*. Interactions with *input devices* begin a feedback loop in which visual and audio renderings are updated by the *computer* and presented by the computer screen and speakers until a steady state has been achieved, i.e., interaction with the *input devices* ceases.

Output Devices

The *user* receives information of data manipulation through *output devices*. There are two types of *output devices*, the **computer screen** and the **speakers**. The computer screen provides information to the user in the form of **visual cues**, whereas the speakers provide it in the form of **audio cues**. In general, speakers refer to any audio output device. Headphones are used in this work. Providing information through these devices is triggered by the *input devices* when they send manipulation commands to the *computer*, which in turn communicates this manipulation through *output devices*. This also depends on the processing carried out.

2.1 Coarse Registration

It is often impossible to capture all the data from one perspective using sensors. As a result, the data is acquired from multiple viewpoints (Rabbani *et al.*, 2007; Xie *et al.*, 2010). Held (2012, pg. 16) observed that "as terrestrial laser-scanners are based on emitting light and receiving its reflection,

they can only record points which are in direct line-of-sight to the instrument. Hence, to fully capture the object and obtain a complete 3D digital reproduction, it is necessary to scan the object from different perspectives."

Scans captured from different viewpoints need to be aligned together, where the scans are translated and rotated in a process commonly referred to as registration. According to Xie *et al.* (2010, pg. 563), "the registration for two point clouds is to determine the best geometric transformation that brings one cloud into alignment with the other in a common coordinate system." There are two forms of registration, fine and coarse registration (Xie *et al.*, 2010). The focus here is on the latter. Coarse registration is the rough alignment of multiple point clouds in preparation for an automatic fine alignment. Its purpose is to facilitate a fast convergence of automatic fine alignment algorithms.

To roughly align two scans, the reference and the target scans are related by a rigid Euclidean transformation (Brenner *et al.*, 2007). This transformation is given in matrix form by equation 2.1,

$$\begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix} = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{pmatrix} \begin{pmatrix} x_2 \\ y_2 \\ z_2 \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix} \tag{2.1}$$

where, x_1, y_1, z_1 and x_2, y_2, z_2 are the three-dimensional positions of a corresponding area of interest in the reference and the target scans, respectively. $r_{11}, r_{12}, r_{13}, r_{21}, r_{22}, r_{23}, r_{31}, r_{32}$ and r_{33} are the parameters of the 9×9 rotation matrix that aligns the target scan with the reference scan. t_x, t_y and t_z are the x, y and z translation parameters that need to be applied to the target scan to register it with the reference scan.

Using these transformation parameters, misaligned scans can be coarsely aligned. However, coarse registration is generally a poor exercise (Xie *et al.*, 2010). The following text proposes how this can be improved when audio is augmented in the process.

Coarse Registration with Audio Augmentation

The aim of registration is to align scans. Therefore, an audio object is 'attached' to the misaligned scan which needs to be aligned with the reference scan. The 'attachment' means assigning a point's 3D position and orientation to the audio object. The 'attached' audio object will be transformed

with the misaligned scan in 3D space.

To receive audio cues, the audio object 'attached' to the misaligned scan has to be emitting audio. As the audio object is transformed in space with the misaligned scan, change in its 3D position and orientation can be aurally communicated using intensity and Head Related Transfer Function (HRTF) (vide subsection 3.1.2) audio cues. The reverberation cue can inform the *user* about spatial changes of the points around the point in the misaligned scan which is being transformed.

Audio cues can potentially provide alignment information to the *user* that the visual cues are not able to communicate. Consider a scanned building shown in figure 2.2, where two separate scans (blue and red) were done from separate perspectives. To align these two scans the transformation process begins with the rotation.

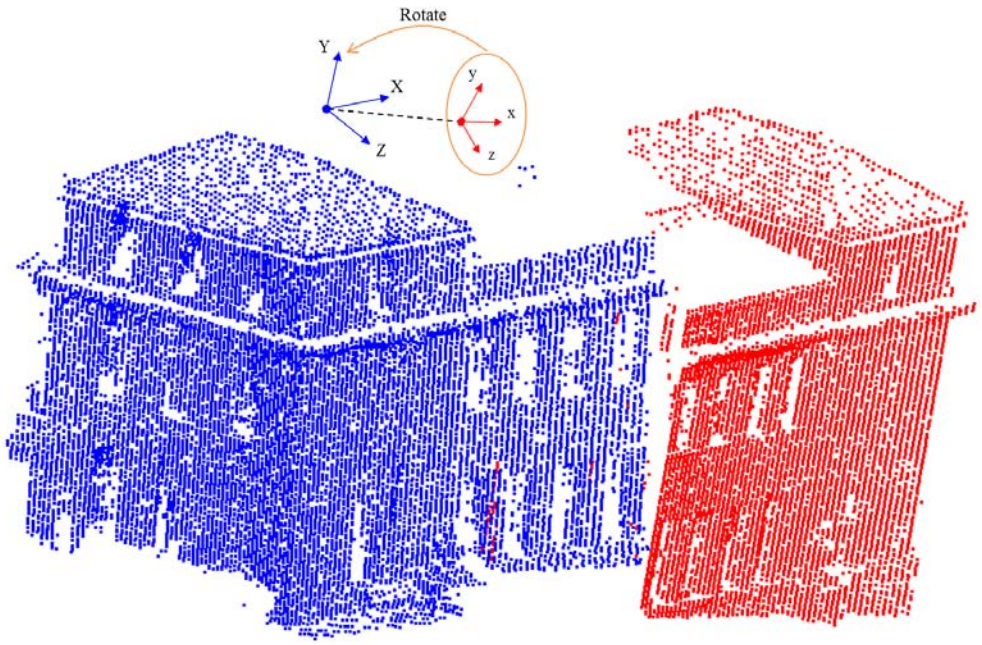


Figure 2.2: Scanned building with separate scans with the misaligned scan needing to be rotated.

Coarse Rotation with Audio Augmentation

The rotation is performed using the rotation matrix of equation 2.1. All the points of the right scan are rotated and mapped into the coordinate system of the left scan. The resulting rotation aligns the scans.

The orientation of the misaligned scan can be tied to the orientation of an audio object attached to it. As a result, as the scan is rotated in space, the orientation of the audio object will change as well. This could provide aural information about the scan's orientation. Mathematically this link is expressed in equation 2.2,

$$R_{object}(R_x, R_y, R_z) = \alpha, \beta, \kappa \tag{2.2}$$

where, R_x, R_y and R_z are the orientations of the scan in X, Y and Z-axes with respect to the reference scan. α, β and κ are the orientations of the audio object in the reference system in units of degrees, giving the orientation of the audio object R_{object}. With reference to figure 2.2, audio augmentation can potentially lead to the red scan being rotated to align with the blue scan.

Coarse Translation with Audio Augmentation

Manual scan translation is done post rotation. The aim of the translation is to minimise the separation (depth) between surfaces. In figure 2.3 the two surfaces are shown separated by depth d. Using coordinates of a common target scanned when each of the scans were attained, this depth can be determined using equation 2.3. In this equation, x_1, y_1 and z_1 are the coordinates of target T in the blue scan and x_2, y_2 and z_2 are the coordinates of the same target (T') in the red scan.

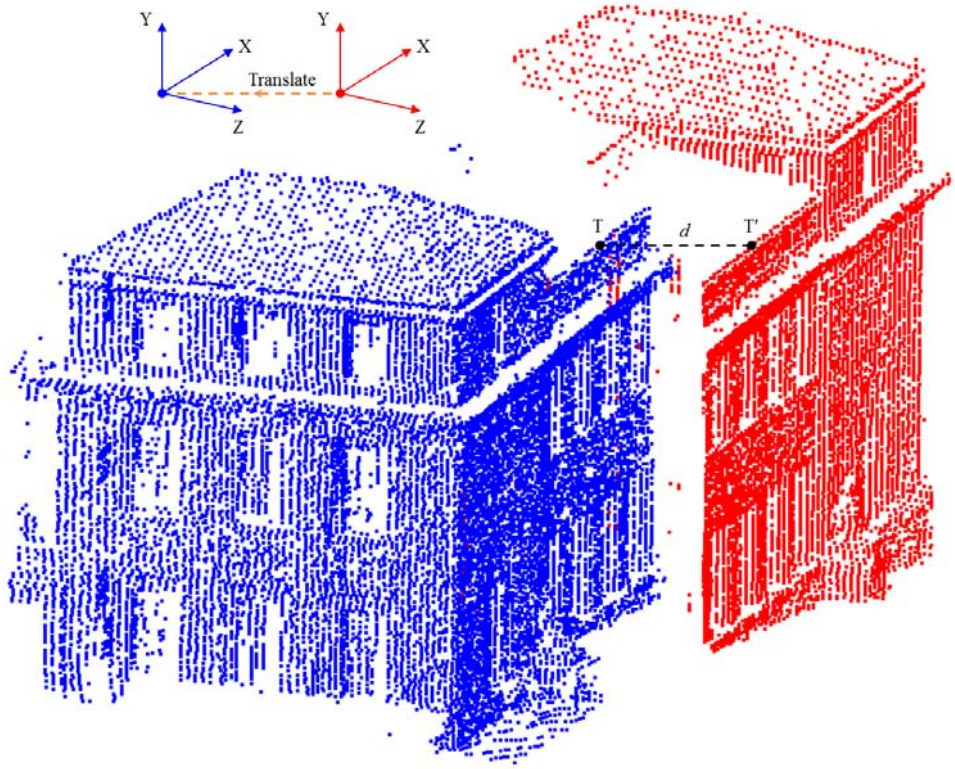


Figure 2.3: Scanned building with separate scans with the misaligned scan needing to be translated through depth d.

$$\begin{aligned} d &= \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2} \\ &= \sqrt{(t_x)^2 + (t_y)^2 + (t_z)^2} \end{aligned} \tag{2.3}$$

Augmenting audio can potentially alert the *user* of this translation where visual cues are limited. The movement of an audio object attached to the red scan can be used to inform the *user* of this translation where the visual cues are limited. This can be done by tying spatial location of the audio object to the translation vector $\mathbf{t}$ of equation 2.1. Spatial location of an audio object is in terms of azimuth, elevation and depth. This relationship is expressed in equation 2.4,

$$L_{object}(t_x, t_y, t_z) = a, e, d \tag{2.4}$$

where, L_{object} is location of the audio object as a function of translation parameters t_x, t_y and t_z, which result in azimuth, elevation and depth, expressed as a, e and d. The intensity and the HRTF cues can inform the *user* of this location in space.

The audio augmented scan alignment could lead to the scans being aligned as illustrated in figure 2.4.

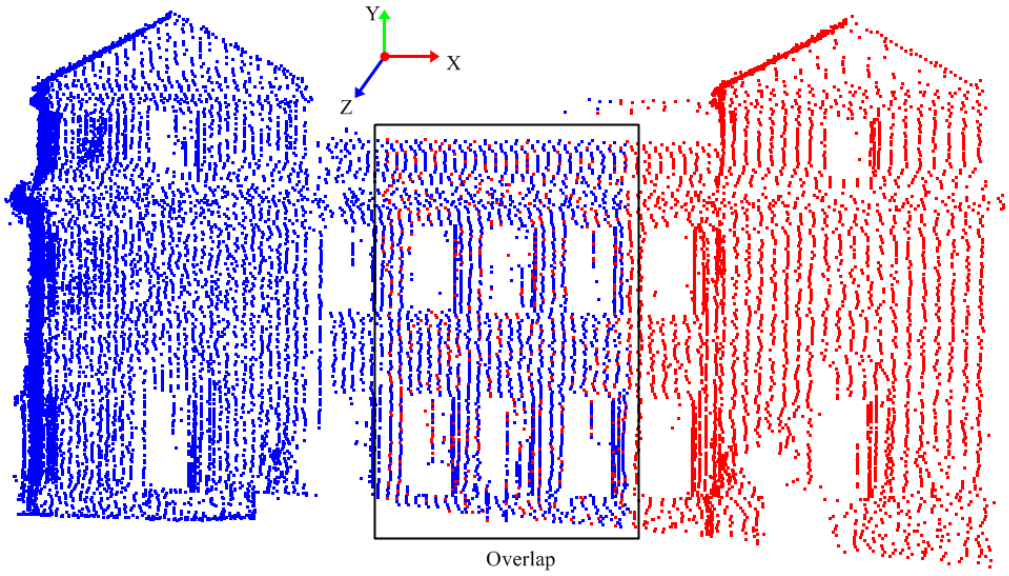


Figure 2.4: Scanned building with separate scans aligned to form a complete scan of the building.

Reverberation (vide subsection 3.1.3) of an audio object can be used in situations where the *user*, immersed in the point cloud attempts to move a surface in relation to another. In this situation, the *user* might need to know how the surfaces are moving in relation to each other in areas which are out of view. This is illustrated in figure 2.5. The *user* is immersed in the point cloud via the camera which is viewing points only within the view frustum, as shown in the figure.

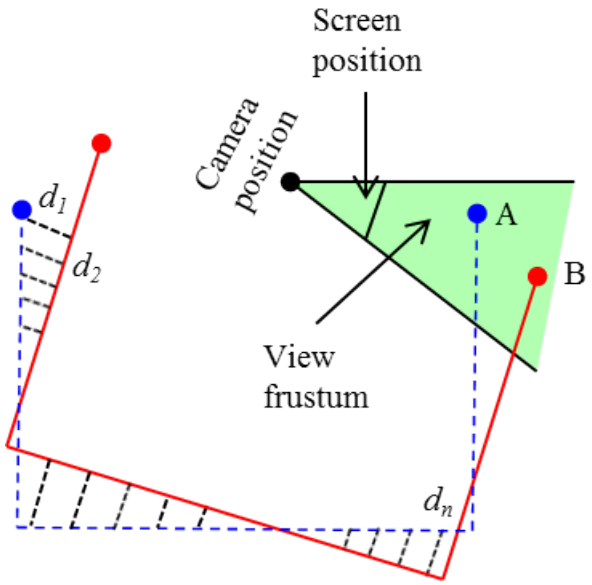


Figure 2.5: User surface B towards surface A while immersed in the point cloud.

In the figure, red lines represent a misaligned scan/surface, while blue lines represent an aligned one or reference scan. While the *user* attempts to move surface B to align with surface A, there could be undesirable movement elsewhere. Or, the *user* might only be interested in the rate at which these surfaces are converging towards each other. A reverberant audio object could aid in alerting the *user* of this movement.

The reverberations of an audio object attached to surface A, could be tied to the average distance between the points of the two surfaces. This average distance can be determined using equation 2.5, where,

$$D = \frac{\sum_{k=1}^{n} d_k}{n} \tag{2.5}$$

D is the average distance, d_k is the k^{th} shortest distance between points of surface A and B and n is the number these calculated distances.

The reverberation of the audio object attached to surface A can then be a function of average distance D. This relationship could be defined using equation 2.6, where, τ is the total experienced reverberation, measured in decibels (dB).

$$\tau(D) = R \tag{2.6}$$

2.2 Data Cleaning

Sensors often produce defective point clouds containing small amplitude noise and/or numerous outliers (Schall *et al.*, 2005; Xie *et al.*, 2004). These defects lead to point clouds similar to the one illustrated in figure 2.6. To successfully model surfaces, these defects need to be removed. Automatic procedures of removing defects exist, however, this is still a largely manual procedure (Schall *et al.*, 2005). Weyrich *et al.* (2004) designed a point-based cleaning system that uses a volumetric brush to remove defects (see figure 2.7).

Figure 2.6: Defective point cloud with noise and outliers (Schall *et al.*, 2005).

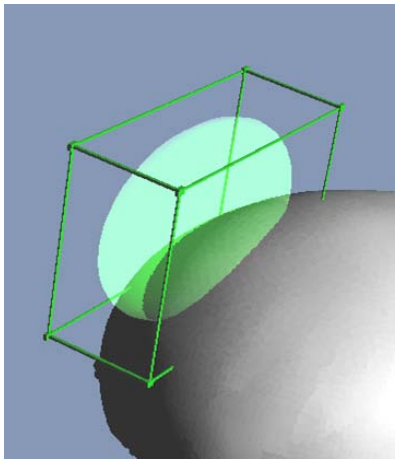


Figure 2.7: Volumetric brush for removing defects in point clouds depicted by an ellipsoid (Weyrich *et al.*, 2004).

The volumetric brush can be moved freely in space around the point cloud to clean it. Weyrich *et al.* (2004, pg. 5) commented that "outlier classification can be confined to the volumetric brush." The volumetric brush can also be made to move in accordance to the object surface. Noise and outliers are therefore removed using such a brush. The following section proposes augmenting audio in this cleaning processes.

Data Cleaning with Audio Augmentation

The aim of audio augmentation in data cleaning is to associate sound with the separation of points from their parent surfaces. Here, outlying points will have distinctly different sounds from those points that lie on a surface. As discussed before, outliers in a point cloud can be removed using a volumetric brush. The audio augmentation in this process therefore entails 'attaching' an audio object to the volumetric brush. This means that the audio object will be assigned the 3D coordinates of the volumetric brush and made to move with it in space.

In the first instance, the volumetric brush isolates the region of the point cloud that is to be cleaned. On a visual interface it may appear as a cube or a sphere. The volumetric brush also isolates the points to be aurally augmented. The first step in the processing is to determine the surfaces (defined by the points) contained within the volumetric brush. This can be done using a method of fitting a plane to the neighbours of the point

of interest, as suggested by Weyrich *et al.* (2004) (see figure 2.8). Using a method of least squares, a best fit plane is fitted to the neighbouring points. Neighbours of a point are those that fall within a pre-set radius of the point of interest. A point inside a volumetric brush can be classified as an outlier if it satisfies a condition such that its distance (depth) from the plane, is greater than the average distance of its neighbours from the plane. This condition is expressed in equation 2.7,

$$d > \frac{\sum_{k=1}^{n} d_k}{n} \tag{2.7}$$

where, d is the depth of the point of interest from the plane, d_k is the distance of the k^{th} point from the plane and n is the number of neighbours within the pre-set radius.

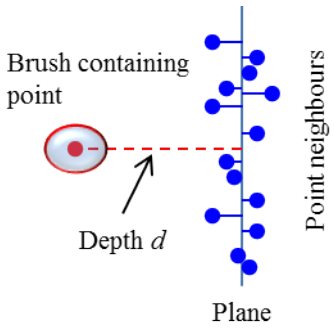


Figure 2.8: 2D view of fitting a plane to the neighbours of a point inside a volumetric brush to determine the point's depth to help in classifying it an outlier or not.

As illustrated using figures 2.8 and 2.9, a point can be classified an outlier using the plane fitting technique described above. By associating sound with the depth d, the *user's* appreciation of the separation of the outlying point from the surface can be enhanced.

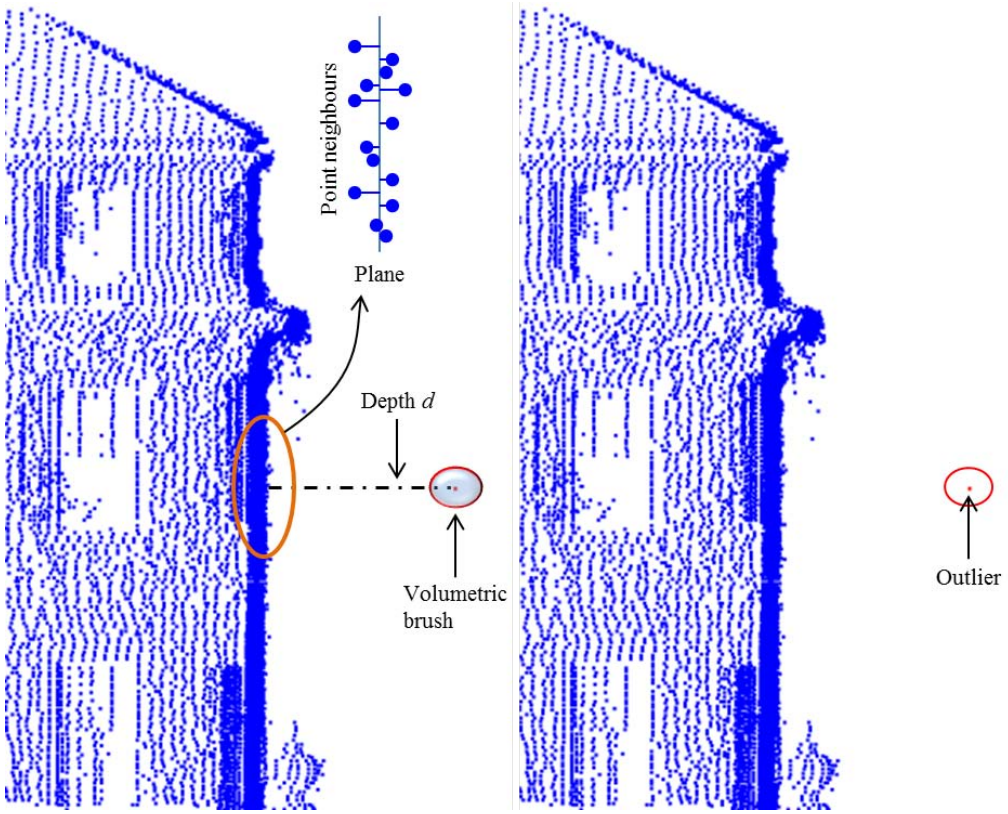


Figure 2.9: Detecting and removing outliers from a defective point cloud using a volumetric brush augmented with audio.

The benefit of this audio enhancement can be appreciated considering that in highly dense point clouds visual editing can be cumbersome because a point cloud has to be viewed from various angles to identify outliers, especially those outliers that are near to their parent surface. Examples of this are shown in figure 2.9.

Here, an example of audio enhancement is to associate the depth of points with the intensity or pitch of a sound. The user can then isolate those points that have a high pitch/intensity and remove them. Equation 2.8 shows one possible way of associating depth with sound intensity.

$$\begin{aligned} \phi(d) &= M - F \times V \\ &= M - F \times 20log_{10}(\frac{s_d}{s_0}) \end{aligned} \tag{2.8}$$

In equation 2.8, $\phi(d)$ is the sound intensity as a function of depth d. V is the loss of intensity in decibels (dB) determined by the initial audio object distance s_0 and the current distance s_d from the listener. (V is the commonly used intensity attenuation model in 3D audio simulation applications and was taken from Vorländer (2008).) M and F are user-defined values, where M is the desired initial audio intensity in decibels and F is a factor which controls the rate at which the intensity drops with depth.

Figure 2.10 demonstrates how F affects the change in intensity as the depth changes. In this case, M was chosen to be 35 dB. Depth is from the plane to where the brush is currently, as shown in figure 2.11. Information such as the 3D position of the brush and the number of points contained in it can be communicated using other cues such as the HRTFs and reverberation. This method of audio augmentation can potentially aid in manual point cloud cleaning processes where the visuals are limited.

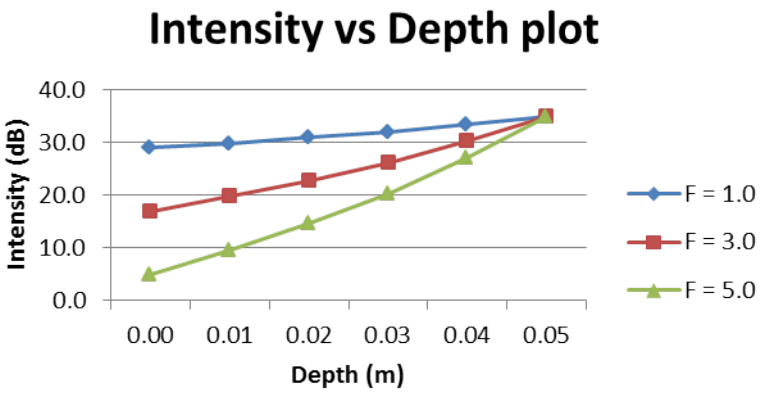


Figure 2.10: Intensity depth plot where the intensity of the audio object attached to the volumetric brush changes with depth and the attenuation depends on factor F.

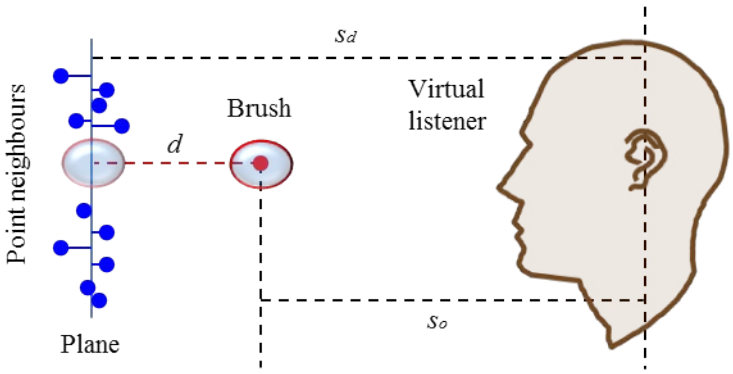


Figure 2.11: Removing an outlier with the volumetric brush augmented with an audio object.

2.3 Simplification

It is often useful to have simpler versions of complex models, as the computational cost of using a model is directly related to its complexity (Garland and Heckbert, 1997). This simplification of models entails reducing the number of polygons making the model by applying a simplification algorithm. Figure 2.12 shows an example of simplification of a 3D model.

Figure 2.12: The bunny model on the left has 69,451 mesh triangles, whereas the simplified version on the right has only a 1,000 mesh triangles (Garland and Heckbert, 1997).

Point clouds can be simplified too, where the quality of surface representation is preserved (Pauly *et al.*, 2002). Point cloud reduction depends on the needs of the user. Alexa *et al.* (2003) applied varying reduction to the same data set with the quality of surface representation preserved.

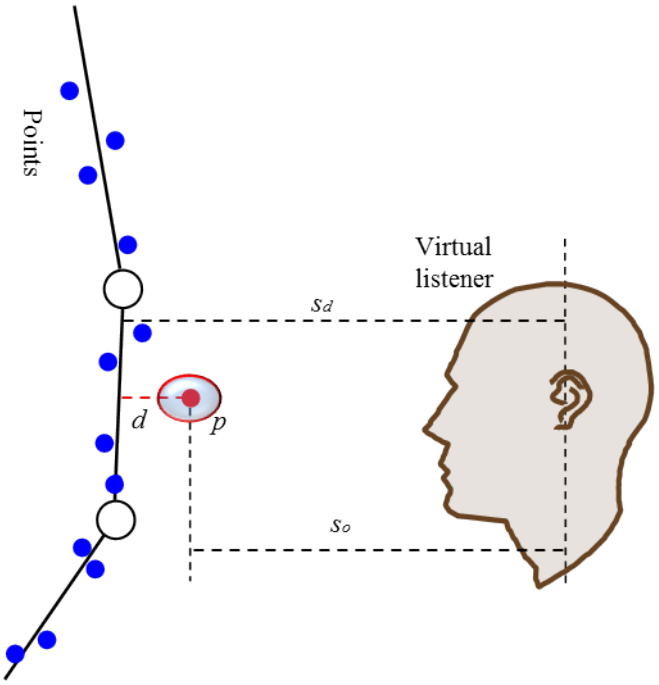


Figure 2.13: A cross-section of a simplified mesh. Improving the quality of the mesh using audio cues.

Simplification with Audio Augmentation

As mentioned already, simplification algorithms attempt to reduce the size of a point cloud while preserving the fidelity of the representation of surfaces. For example, the form and curvature of surfaces are preserved. After simplification, surfaces are visually inspected to seek out places where there has been over or under simplification.

The degree of simplification can be estimated as a function of local surface curvature and the point density. Ideally the point density should be directly proportional to the local surface curvature. The stronger the local surface curvature, the higher the point density. Here audio enhancement is achieved by associating sound with the function of the local surface curvature and the point density.

Improving the quality of the created mesh entails running simplification algorithms to create more triangles and make the mesh more detailed. Audio augmentation in this process can be used to identify those areas where more detail needs to be created. Visual cues could be limited in informing the *user*

of this, depending on the perspective of the camera.

Figure 2.13 shows a cross-section of a simplified area in the mesh. Point p, highlighted in red, is a considerable distance from the mesh triangle, leading to considerable loss in detail. The *user* might want to change this and improve the detail. Determination of the depth d of point p from the surface of the triangle can be done using an audio object attached to a tool similar to the volumetric brush mentioned in subsection 2.2.

As the *user* moves this tool, audio cues can be used to inform the *user* of this depth. Audio intensity is a good candidate to act as the cue that can inform the *user* of this depth. This problem is similar to that discussed in subsection 2.2, and the depth, using audio intensity can be determined using equation 2.8. Additionally, the HRTF and pitch cues can be used to determine the location of the tool and/or refine depth information emitted to the *user*.

2.4 Discussion

The presented examples are not exhaustive. Other areas of processing such as hole-filling, surface fitting, smoothing, etc., could also benefit from audio augmentation. Point cloud processing techniques have some overlapping attributes and the augmentation mentioned here can be used in a similar manner in other processing techniques.

In this project the audio enhancements are used to present to a user a sense of the location of phenomena, the depth of phenomena and the prevalence of phenomena (ambience). Chapter 3, provides in detail the functionality of the human auditory system and various audio cues that could be used in the augmentation.

In Chapter 5 the limitations of auralisation will be tested. From the tests the limits of the application of audio enhancements for point cloud processing will be determined and discussed.

Chapter 3

Literature Review on Audio

This chapter will provide literature on the human auditory system and acoustics. Different audio cues will be explained, with emphasis put on their importance to the listener. This chapter will also provide insight into how sound is perceived using headphones and how this is different to sound perceived in the free sound field, i.e., without headphones.

3.1 The Human Auditory System

Handel (1989, pg. xi) noted that, "listening is centripetal; it pulls you into the world." In agreement with this statement, Kapralos *et al.* (2003, pg. 1) observed that "hearing can serve to guide the visual attention and therefore eases the burden off the visual system." These observations emphasise the importance of the auditory system to humans in navigating their environments.

Position and depth estimation of audio sources around the listener are done with respect to a coordinate system whose origin is the centre of the listener's head. Throughout this dissertation, position estimation will refer to estimating the direction from which sound is emitted, in terms of azimuth and elevation. Depth will refer to how far the sound source is from the listener. Figure 3.1 by Kapralos *et al.* (2003) and Kendall (1995), graphically illustrates this coordinate system. The coordinate system has an axis (interaural axis) and three planes (median, frontal, and horizontal planes) which will be used in defining the position of an audio source with respect to the listener.

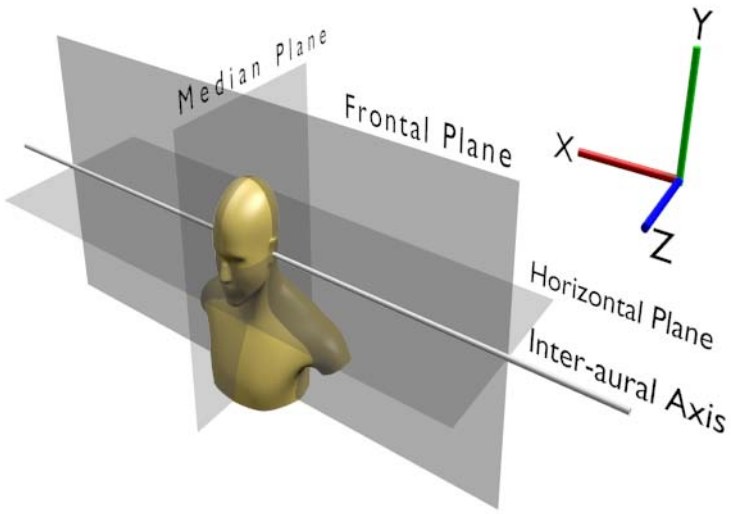


Figure 3.1: Coordinate system of the human auditory system (Kapralos *et al.*, 2003 and Kendall, 1995).

The inter-aural axis is the axis that passes through a person's left and right ears. The median plane is a vertical plane, intercepting the inter-aural axis at right angles midway between the ears. The horizontal plane is a plane containing the inter-aural axis and intercepting the median plane at right angles midway between the ears. The frontal plane is a vertical plane containing the inter-aural axis and intercepting the median plane at right angles midway between the ears.

The point where the planes and the inter-aural axis intersect is the origin of the coordinate system. The idea of depth and position (azimuth and elevation) estimation is illustrated in figure 3.2. This figure illustrates the idea of a listener estimating both the depth and position from which the sound source (red sphere in the figure) is located. For example, in this figure, the correct estimation of the source's position would be at depth X m, azimuth of -30° and elevation of 10°.

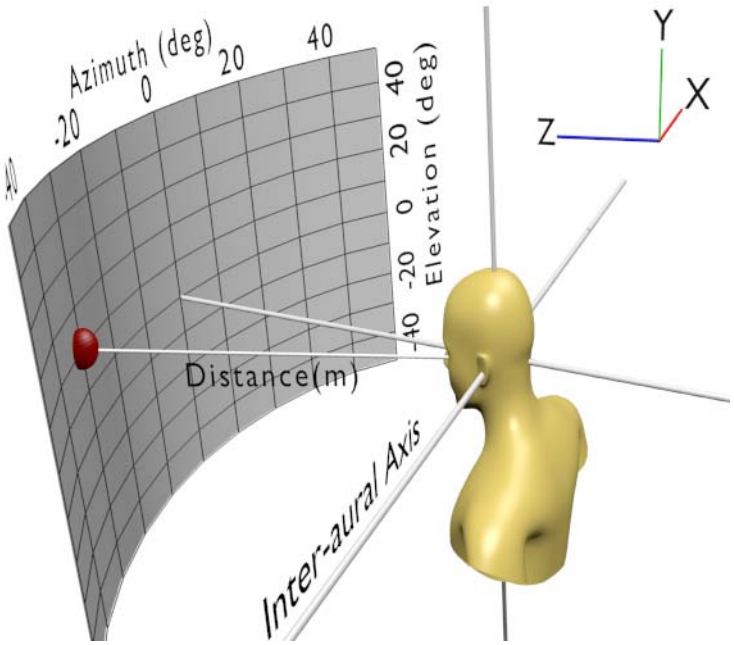


Figure 3.2: Depth (in meters) and Positional (azimuth and elevation in degrees) Estimation. The position of the sound source can be estimated relative to the listener's head at the origin in the system illustrated by the coordinate system shown. In this figure, the red sphere represents the sound source whose spatial information is to be extracted.

The auditory system uses various audio cues to discern spatial information. The auditory cues that will be discussed in this study are:

1. Binaural Cues
2. Head Related Transfer Functions (HRTF)
3. Reverberation

Information extraction using these audio cues is more accurate when these cues operate in a collaborative manner. This is usually the case when audio is emitted from a source in the free field, in which case the auditory system combines these cues to get the information. These audio cues are the ones suitable for the purposes of this study and will be discussed in the following subsections.

3.1.1 Binaural Cues

Binaural cues are audio cues that exist due to binaural hearing – hearing that employs both ears. Binaural hearing affords humans the ability to estimate the position of sound (Vorländer, 2008). There are two cues in binaural hearing, namely, inter-aural time difference (ITD) and inter-aural intensity difference (IID). Rayleigh (1875) formulated the Duplex theory which underlines how binaural cues operate.

Due to the distance separation between the ears, audio emitted from a source reaches the ear on the side of the audio source before reaching the ear shadowed away from the audio source by the head. The time difference between these two events is called inter-aural time difference. Kapralos *et al.* (2003, pg. 11) observed that audio intensity gets attenuated before reaching the ear shadowed by the head. This attenuated intensity of the sound is the inter-aural intensity difference.

The maximum ITD value occurs when the sound source is located to the side of the ear directly along the inter-aural axis, i.e., at azimuth of $\pm 90^{\circ}$ (Handel, 1989; Kapralos *et al.*, 2003). The minimum detectable ITD is 0.01 ms and ITD values are always less than 1 ms (Handel, 1989).

In theory, ITD and IID values of sound sources in the median plane are equal to zero. The shape of the head and the existence of complexly shaped ears, amongst other things, makes this false and as a result, these cues will be near zero for audio sources in the median plane. For sources in the median plane, the precision of position estimation in terms of determining whether the source is at the front or the back or up or down, is less (Razavi *et al.*, 2005). This leads to a position estimation confusion, which can be resolved with the help of other audio cues (Vorländer, 2008).

Audio augmented point cloud processing would require that positions of audio objects be estimated, depending on the processing. The ITD and IID cues could be influential here, especially in situations where they are more prevalent. As noted, these cues are more prevalent towards the inter-aural axis and could assist in processing if areas of interest are in this region.

3.1.2 Head Related Transfer Functions

Begault (1994, pg. 52) defined Head Related Transfer Function (HRTF) as

"the spectral filtering of a sound source before it reaches the eardrum that is caused primarily by the outer ear/pinna." Pinnae are asymmetrically and complexly shaped and as a result have notches and grooves. The presence of these notches and grooves give rise to sound reaching the eardrum having experienced time delays of a range of 0 – 300 μs (micro-seconds), depending on the sound source location with respect to the listener (Begault, 1994; Kapralos *et al.*, 2003). These time delays result in spectral content of the sound reaching the eardrum being different to that emitted by the source. Begault (1994) noted that amplitude differences contribute to differences in spectral content in addition to time delays.

3.1.3 Reverberation

When sound is emitted by a source in the free sound field, the sound waves will be reflected by surfaces and objects with which they come into contact with. The reflected waves will be less intense than the direct waves that travel unhindered from the source to the listener. According to Kapralos *et al.* (2003), these surfaces and objects have the potential of absorbing some of the sound wave while the other part gets reflected. This phenomenon gives rise to *reverberation*. Begault (1994, pg. 100) defined reverberation as "the energy of a sound source that reaches the listener indirectly after reflecting off surfaces within the surrounding space occupied by the sound source and the listener."

Reverberation is said to be irregular and complex and dependent on environment geometry, material of objects in the environment and the spectrum of the sound (Kapralos *et al.*, 2003). Begault (1994) noted that sound energy builds and decays quicker in smaller rooms (environments) than in larger rooms, providing useful information about the size of the environment. According to Begault (1994) and Kapralos *et al.* (2003), reverberation can also act as a depth cue (this will be discussed in sections to come) in addition to providing information about the type of environment. There are two categories of reverberations/reflections that a sound wave can experience, early reverberations/reflections and late reverberations/reflections. Whether a reflection is early or late, depends on the time it takes to reach the listener after the direct sound has done so.

3.2 Positioning Estimation of Sound Sources

Position estimation of an audio source entails estimating the direction from which sound is coming from after being emitted by the source. Position estimation has two components to it, lateral estimation and vertical estimation. In point cloud processing there could arise situations where the lateral and vertical positions of objects (points) are needed, for example, in coarse alignment of scans. This could be achieved with estimating the position of the augmented audio object.

Lateral estimation entails estimating the azimuth from which sound is emitted, this is done with respect to the median plane which is at 0° azimuth (see figure 3.2). Vertical estimation on the other hand entails estimating the elevation from which sound is emitted and this is done with respect to the horizontal plane which is at 0° elevation (see figure 3.2).

3.2.1 Positioning Estimation Errors

Even with the use of all position estimation cues, which will later be discussed, error free estimation cannot be achieved. These errors will be shown in the following order: angular errors, azimuth errors and then elevation errors. The discussion around these errors is done based on the experiments done by Brungart *et al.* (1999).

Angular Error

Angular error includes both azimuth and elevation errors. According to Brungart *et al.* (1999, pg. 1960), angular error "corresponds to the angle between the 3D vector from the centre of the head to the source location and the 3D vector from the centre of the head to the response (estimated) location." It is a combination of azimuth and elevation errors in position estimation of a sound source. (See figure 3.3.)

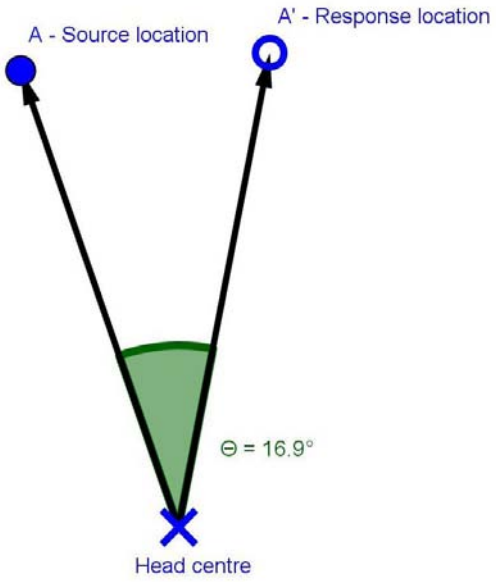


Figure 3.3: Angular Error: angle between the vector from the centre of the head to the source actual location and the vector from the centre of the head to the estimated location. This only shows a 2D view of the situation.

Angular errors are not the same for all locations around the listener. Brungart *et al.* (1999) found that the largest errors are behind and above the listener and that the smallest are for sources placed relatively far in front of and to the side of the listener. Furthermore, it was found that angular error increased as the sound source approached the listener to within $25cm$, especially for front and above sources. Additionally, angular errors were also shown to increase slightly with elevation. The overall mean angular error was reported to be 16.9°. The results are based on experiments done with four subjects where stimulus was presented to them from 27 locations.

For augmented audio, it should be expected that the angular error will vary in the manner explained above. The user must take necessary steps to ensure where possible that th listener and an audio object of interest are not in a spatial relation that would lead to high angular errors.

Azimuth Error

Azimuth error is the error which occurs when the auditory system determines lateral directions to sound sources. Brungart *et al.* (1999) found that there

exist systematic biases for each subject (listener) in indicated source regions when estimating azimuths (and elevations). The biases are said to change systematically with source location and therefore need to be accounted for. As a result, Brungart *et al.* (1999) had to calculate a special measure of variability response called the bias-corrected root-mean-square (BCRMS) error. BCRMS excludes the systematic response bias and was calculated for each listener for all source locations (nine elevation and depth locations in the case of azimuths).

Instances may arise where the user might need to use the azimuth of an audio object to infer some information related to the point cloud being processed. According to the above findings, each user will experience different azimuth accuracies from augmented audio. This suggests that the point cloud processing experience will vary from user to user. This does not suggest that the experience will not be enhanced by the augmentation, however.

Elevation Error

Elevation error is the error which occurs when the auditory system determines vertical directions to sound sources. Brungart *et al.* (1999) found elevations to be estimated with less error than azimuths. The mean elevation error was found to be 11.3°, compared to the mean azimuth error which was found to be 12.6°.

Brungart *et al.* (1999) compared obtained results with the results obtained from studies done by Wightman and Kistler (1989) and Makous and Middlebrooks (1990). Numerical differences in the results were cited and these were a result of different conditions used in doing the experiments. However, Brungart *et al.* (1999, pg. 1963) observed that "all three studies indicate that directional position estimations are least accurate when the source is located above and behind the head."

Similarly, the elevation of an audio object could be useful to the user in audio augmented point cloud processing. Based on the literature, while using elevation of a sound source, the user must expect that errors will be more above and beyond the listener. The user must avoid such instances to maximise the potential of audio elevation in processing.

To be able to estimate the position an audio source, the auditory system makes use of various audio cues. The cues responsible and the manner in

which they are used by the auditory system to estimate the position of sound sources is discussed below.

3.2.2 Binaural Cues as Position Estimation Cues

The Duplex theory is built around the notion that with ITD and IID cues, the direction from which an audio source is located can be estimated. Audio sources that are not in the median plane, will have ITD and IID values discernible by the auditory system. As previously mentioned, in the median plane, ITD and IID are virtually zero.

According to Kapralos *et al.* (2003), the Duplex theory alone is incomplete for sound position estimation, as listeners with only one hearing ear are able to estimate the position of sound too. There is a shortcoming inherent in ITD and IID cues: the front-back confusion problem (Brungart and Rabinowitz, 1999; Brungart *et al.*, 1999; Kapralos *et al.*, 2003).

Front-back confusion occurs for audio sources located in the median plane. A source directly in front of the listener will have the same ITD and IID values as that which is directly behind the listener. This gives rise ambiguity, i.e., being able to tell which source is in front and which is behind. Listeners can get rid of these ambiguities by using head movements when in the free field (Brungart *et al.*, 1999; Kapralos *et al.*, 2003; Thurlow *et al.*, 2005). Head movements allow for the relative position between the listener and the source to change and therefore give rise to ITD and IID values to allow for better position estimation (Handel, 1989). (See figure 3.4 after Kapralos *et al.* (2003) for front source head movements illustration. The same principle applies for disambiguating locations of sources located behind the listener, this illustration can be found on Kapralos *et al.* (2003, pg. 23)).

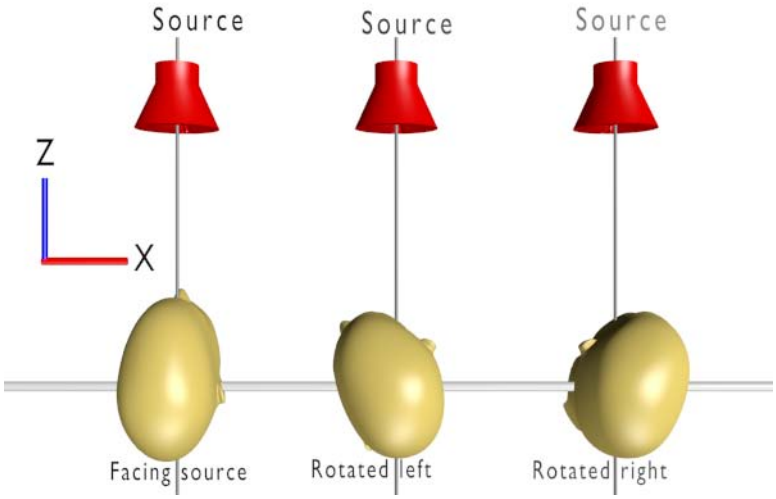


Figure 3.4: Front source head movements. The listener rotating his/her head to make better position estimations (Kapralos *et al.*, 2003).

Head movements (and other shortcomings) have led many researchers to consider Rayleigh's Duplex theory to be incomplete and that head-related transfer functions (HRTFs) are essential in complimenting binaural cues in sound source position estimation. Akeroyd (2006) does credit these binaural cues in sound position estimation, however.

To estimate the position of an augmented audio in point clouds would rely on the binaural cues as they play an important role in this regard. As observed, the binaural cues are prevalent away from the median plane. It should be expected then that in audio augmented point cloud processing, the placing of an audio object with respect to the median plane will play a role in estimating the position of this audio. If the audio object is along the median plane then, finding a way for the listener to do head movements could help enhance the position estimation of the audio object and therefore enhance the processing.

3.2.3 Head-related Transfer Functions as Position Estimation Cues

Kapralos *et al.* (2003, pg. 13) defined HRTF as "the filtering of the sound source spectrum caused by the complex interactions of the sound waves with the head, shoulders, torso and the outer ear (pinna) prior to reaching the ear drum."

Bronkhorst (1995) noted that HRTFs act as position estimation cues for the auditory system. HRTF is a monaural cue and Kapralos *et al.* (2003, pg. 13) defined how it operates by stating that "the HRTF modifies the spectrum and timing of a sound signal reaching ears in a location dependent manner which is recognized by the listener and used as a position estimation cue." As previously stated, the HRTF cue helps estimate the position of sound sources in cases where there is ambiguity for binaural cues. HRTF cue is therefore a very important cue, as Brungart *et al.* (1999) discovered that it allows for estimation of elevation and azimuth of a sound source. In support of this, Kapralos *et al.* (2003, pg. 14) observed that "HRTFs can provide information used to estimate vertical directions (elevations) and to remove ambiguities due to front-back confusions." Handel (1989) also made a similar observation.

Different HRTFs exist for different locations around the listeners head. Brungart and Rabinowitz (1999, pg. 1465) pointed out that "HRTF varies substantially with depth for nearby sources (less than 1m) and HRTF is virtually independent of depth for sources beyond $1m$." HRTFs are also frequency dependent as Brungart *et al.* (1999, pg. 1957) found that "the magnitude of the HRTF is relatively greater at low frequencies than at high frequencies when the source is near the head." The auditory system uses such differences to do azimuth and elevation estimations. However, Brungart *et al.* (1999) found that position estimation accuracy degraded as sources got closer to the listener. Brungart and Rabinowitz (1999) discovered that elevation estimation could be independent on depth from the sound source, whereas azimuth estimation could be dependent on it.

As suggested, HRTFs can be useful where the binaural cues are limited and can assist in the processing requiring position estimation in this case. The above stated limitations of the HRTFs should be noted too so the augmentation is such that these limitations affect the point cloud processing.

3.3 Depth Estimation of Sound Sources

The estimation of depth(s) to audio source(s) by the human auditory system has not been studied as much as position estimation of audio sources. It has been found by many, however, that depths are poorly estimated in comparison to position estimation. The potential use of audio depth in point cloud

processing techniques such as cleaning and simplification was mentioned in Chapter 2. The cues allowing for depth perception are observed.

3.3.1 Intensity as a Depth Cue

According to Zahorik *et al.* (2005) and Väljamäe (2005), intensity can be used effectively by the auditory system as a depth cue. However, Mershon and King (1975) observed that as physical auditory depths increase, estimated depths increasingly become underestimates of physical auditory depths in situations where intensity is the only depth cue available. This observation was supported by Zahorik *et al.* (2005). Audio intensity gets attenuated when sound travels directly from the source to the listener. Assuming that the power of the audio from the source is kept constant, the intensity of the sound perceived by the listener will decrease with the increase in depth and increase with the decrease in depth. Kapralos *et al.* (2003) provided the following model to show the attenuation of audio intensity with depth:

$$L_{loss} = 20 \times log_{10}(\frac{s_d}{s_0}) \tag{3.1}$$

where, L_{loss} is the loss in intensity measured in decibels (dB), s_0 is the initial audio source depth and s_d is the current depth between the listener and the audio source.

This model given by equation 3.1, follows the inverse square law of audio intensity attenuation. Begault (1994), Mershon and King (1975), Zahorik *et al.* (2005) and Shinn-Cunningham (2000) observed that for every doubling of depth of the sound source from the listener, a 6 dB loss in source intensity is experienced. The 6 dB loss however, is only experienced for sources in anechoic (non-reverberant) environments and for omnidirectional sources only. For sources which are not omnidirectional, Begault (1994) and Kapralos *et al.* (2003) found that sound intensity in this case drops by 3 dB for doubling of depth. The model given by equation 3.1 is said by Kapralos *et al.* (2003) to be incomplete. Furthermore, Begault (1994) and Kapralos *et al.* (2003) noted that the model with the 6 dB drop is commonly used in 3D audio simulation applications.

For nearby sources (within about 1m) of the listener, audio intensity is not as effective a cue for depth perceptions. Binaural cues play a role in depth estimations for nearby sources, this will be explained later in sub-subsection 3.3.3. Mershon and King (1975) observed that it is particularly intensity

differences over a range of 20 dB which contribute to intensity being useful as a relative depth cue.

Normally, other depth cues exist, acting as partners to intensity in the task of estimating depths. Zahorik *et al.* (2005, pg. 412) observed that "perceived depth can bear little relationship to the physical depth" when intensity is the only available depth cue. This suggests that other cues exist in attaining auditory depth accuracies. In sub-subsections that follow, the roles played by these other cues in auditory depth estimations will be outlined.

Determining depths of audio objects in point clouds using intensity as a depth cue could enhance the processing in cases where the stated limitations do not exist. As noted, it should be expected that as the physical depth increases, depth underestimates should be expected. For example, in those situations where points are significantly far from the surface of interest, intensity alone should not be expected to be the best indicator of the depth.

3.3.2 Direct-to-reverberant Energy Ratio as a Depth Cue

Sound emitted by a source gets reflected off surfaces and objects it interacts with before the listener can hear the sound (see subsection 3.1.3 on reverberation/reflections). The ratio of energy of the direct sound to energy of the reflections (direct-to-reverberant energy ratio) acts as a depth cue, according to Handel (1989), Mershon and King (1975), Shinn-Cunningham (2000), Bronkhorst (1995), Bronkhorst and Houtgast (1999) and Zahorik *et al.* (2005).

Bronkhorst and Houtgast (1999, pg. 517) stated that "perceived depth depends on ratio of direct-to-reverberant energy ratio." Furthermore, Bronkhorst and Houtgast (1999, pg. 518) noted that perceived depth was estimated mainly by "the number of reflections and the relative level of these reflections." Zahorik *et al.* (2005) found that reverberant energy depends particularly on two factors, namely, the size of the room and the acoustic properties of the reflecting surfaces.

Zahorik *et al.* (2005, pg. 413) noted that "increments of 5 to 6 dB in direct-to-reverberant energy ratio were found to be just-noticeable, over a range of energy ratios (0–20 dB)." The use of direct-to-reverberant energy ratio as

a depth cue in 3D audio simulation applications was supported by Zahorik *et al.* (2005), by reporting that the number of simulated reflections can be increased, in which case this will mean that the apparent depth has increased. Similarly, simulated reflections can be decreased, implying a decrease in the apparent source depth.

Simulated audio reflections could assist in depth perception of objects in point clouds. By controlling the amount of simulated reflections, the user could improve the depth perception of augmented audio and therefore points of interested in point cloud processing tasks.

3.3.3 Binaural Cues as Depth Cues

It was previously pointed out that binaural cues play a role in making audio source depth estimations. Binaural cues play a role particularly for nearby audio sources. Shinn-Cunningham (2000, pg. 227) observed that "for nearby sources, changes in depth depend on source direction" – this implies that binaural cues are used. According to Zahorik *et al.* (2005), binaural cues are almost independent on depth for sources located more than 1m away, therefore making the use of binaural cues for far away audio sources unlikely.

According to Mershon and King (1975) and Handel (1989), the use of binaural cues in making depth estimations is rather contradictory and unclear. The confusion arises in knowing which binaural cue, whether ITD or IID, is useful in making depth estimations. ITD and IID are already known to play an important role in position estimations of sound sources, Zahorik *et al.* (2005) however, believes that it is the IID which is responsible for depth estimations where binaural cues are involved.

The use of binaural cues in making depth estimations is said to be an area that requires more research. Handel (1989) believes that in this area, binaural cues are inferior and are most probably dominated by other cues like the intensity cue and direct-to-reverberation energy ratio cue. This convinced Handel (1989) that including binaural cues as depth cues in 3D audio applications is rather unnecessary.

As depth cues in audio augmented point cloud processing, the binaural cues might not be obviously useful. The reviewed literature does not fully support using binaural cues for depth perception, as contradictions exist.

3.3.4 Frequency Changes as Depth Cues

Frequency of audio can be a useful depth cue, according to Handel (1989) and Kapralos *et al.* (2003). Kapralos *et al.* (2003) furthermore noted that spectral changes can provide relative depth information, unless the listener has prior knowledge of the source, in which case absolute depth information can be provided. The frequency spectrum of an audio source changes with depth because of the interaction of the sound wave with the atmosphere.

It is high frequency sounds which are said to be affected more by atmospheric conditions and leading to spectral changes. Greater attenuation is experienced for higher frequency components as the depth between the source and the listener increases, as highlighted by Handel (1989) and Kapralos *et al.* (2003).

The complexity of spectral changes of sound, make this notion of frequency being used as depth cues quiet contradictory, especially in auditory depth simulations (Handel, 1989). Handel (1989) and Kapralos *et al.* (2003) both agree that spectral changes can be used as depth cues, particularly for high frequency sounds, but they emphasise that familiarity with the sound can be very useful and lead to better auditory depth estimations.

Changes in the frequency of an audio object could provide depth information of objects in point clouds for processing. The reviewed literature suggests that audio frequency does change with depth, as already observed with audio intensity. This cue is worth exploiting in point cloud processing where depths of objects are required.

3.3.5 Familiarity with the Emitted Sound

The depth cues discussed above might not all be available to be used by the auditory system in making depth estimations. It depends significantly on the conditions from which sound is emitted and where the listener is located. For example, Shinn-Cunningham (2000) found that in reverberant environments, the IID cue is irrelevant in making depth estimations, it is under anechoic conditions where it contributes. Nonetheless, what is popular amongst many researchers is the fact that the more the cues available, the better the accuracy in making depth estimations.

Intensity and direct-to-reverberant energy ratio cues are said to be more

critical in depth estimations. However, Handel (1989, pg. 108) observed that, "perception of depth is relatively crude and susceptible to experience, whereby experience tends to dominate physical variables."

Listeners are said to make better depth estimations of audio sources emitting sounds that they are familiar with. An example of such a sound is speech. Speech has on average a sound pressure level of 70 dB (Vorländer, 2008). Humans are familiar with speech more than any other sound. As a result, auditory depth estimations to speeches are more accurate than for other types of sounds because of this element of familiarity. Handel (1989) demonstrated that for sources located at 0° azimuth in an anechoic chamber, listeners overestimated depths from where shouts were emitted and underestimated whispers in reference to normal speech. These results are shown in figure 3.5.

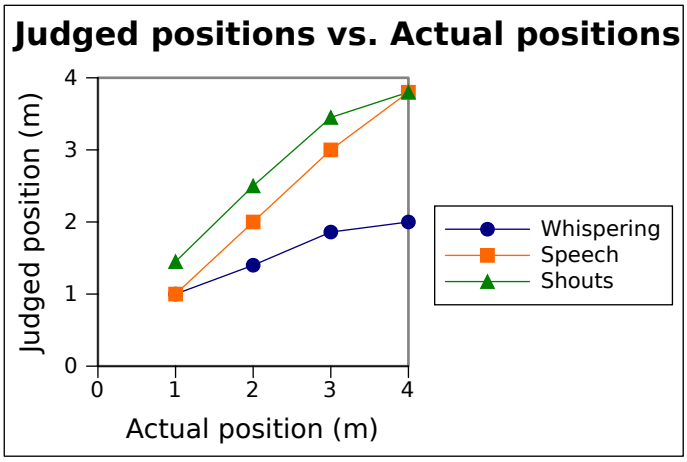


Figure 3.5: Estimated positions v. Actual positions when the following sounds are used: whispering, low-level and conversational-level speech and shouting. This is for sound emitted from a speaker at 0° azimuth in an anechoic chamber (Begault, 1994).

As mentioned, depth estimations are "crude" tasks and the accuracy lies significantly on the use of all available depth cues. Perhaps equally significant is the reliance on familiarity with the emitted sound. These are the factors that contribute to overestimates and underestimates of auditory depths which Zahorik *et al.* (2005) modelled. Using sound stimulus that the user is familiar with in audio augmented point cloud processing could help improve the depth perception exercise, as suggested in the literature.

3.4 Audio Perception with Headphones

Manual point cloud processing will be done on workstations, because of this the user will use headphones. Due to this, headphones will be used in the tests. It is for this reason important that audio perception with headphones be elaborated on, based on existing literature. Headphones are important in sound reproduction, as Begault (1998, pg. 1) noted that "headphone playback is considered optimal for reproducing 3D spatial acoustic imagery because of the relative immunity to acoustical detriments (such as background noise)."

3.4.1 Individualised and Non-individualised Head-related Transfer Functions

As Wightman and Kistler (1989, pg. 859) noted, the key to simulation of free field listening conditions would be to "have waveforms at a listener's eardrums being the same under headphones as in the free field." Furthermore, Wightman and Kistler (1989) argued that this will ensure the listener experiences the same auditory feedback from sources as would be experienced in the free field.

In headphone listening, it is important that the headphones model head-related transfer functions (HRTFs) which would be available in free field listening. Because individuals are different in terms of their anatomy, each individual has his/her own individualised HRTFs. This led Bronkhorst (1995, pg. 2542) to argue that "positions of virtual sources (sources heard through headphones) can generally be estimated with almost the same accuracy as real sources, especially when individualised HRTFs are used."

However, non-individualised HRTFs are used in headphones, this means that one particular set of HRTFs is used for headphones used by people with their own individual HRTFs. This introduces spatial location estimation errors which will vary across different listeners as each listener has his/her own HRTFs (Begault, 1998; Wightman and Kistler, 1989). As headphones will be used in audio augmented point cloud processing, different users might have different experiences due to non-individualised HRTFs. The differences are not expected to be major and might not even be noticeable.

3.4.2 Front-back Confusions

What has been mentioned is that to resolve front-back confusions in the free field, the listener needs to do head-movements on head-movements for resolving front confusions). Wightman and Kistler (1989) observed that with the use of headphones in virtual environments, one of the drawbacks that exist is that head-movements cannot be used to remove front-back confusions. This is because in the free field sound sources are 'removed' from the listener, whereas with headphones they are 'attached' to the listener.

Wenzel *et al.* (1991) and Wightman and Kistler (1989) found that when headphones are used, front-back confusions are twice as high as those in the free field, 11% v 6% for individualised HRTFs and 31% v 19% for non-individualised HRTFs. Front-back confusions will be a limitation, especially when audio objects will exist behind the listener as well.

3.4.3 Externalisation of Sound Sources

Using headphones to estimate the positions of sound sources can lead the listener to believe that the sound source originates from inside the head (Begault, 1994; Wightman and Kistler, 1989). This means that the sound sources are not externalised. The sensation of non-externalisation can be experienced by talking while ears are blocked with the hands – speech will appear to come from inside the head. Non-externalisation is a limitation inherent in headphone listening. Non-externalisation of sound sources impacts their position estimation in virtual environments and makes them appear to be at the edge of the listener's head when they are not (Begault, 1994; Begault, 1998).

A proposed and tested solution to minimising non-externalisation is that of making the virtual environment reverberant. According to Begault (1998, pg. 3), "reverberation has been found to dramatically increase the externalisation of stimuli relative to non-reverberated stimuli, in one case, from 2% to 90%." Reverberant virtual environments lead to experiences that are more real in a sense that, sound waves will experience reflections similar to those in reverberant free field conditions. Making a virtual environment reverberant is important in that it will externalise sound sources by simulating reflections (Begault, 1994). In audio augmented point cloud processing, this will be taken into consideration in order to counter the effect of non-externalised audio sources. It should be noted, however, that reverberant environments

do not necessarily fully guarantee externalisation of sources as it may depend on the source stimuli and HRTFs (Begault, 1994).

3.4.4 Errors in Headphone Position Estimations

The depth and position estimation errors are due partly to the factors that have just been discussed: non-individualisation of HRTFs, front-back confusions and non-externalisation. Another factor is that of experience, where the most experienced users of headphones will make better estimations of depths and directions (Wenzel *et al.*, 1991).

Wenzel *et al.* (1991) conducted a study investigating position estimation with non-individualised headphones. In these experiments, HRTFs of an accurate 'estimator' were modelled and headphones synthesised to these HRTFs. This accurate 'estimator' was omitted from the study as a test subject. The study involved 16 subjects, thereby each of the subjects having non-individualised HRTFs. The subjects were required to make azimuth and elevation estimations in the free field as well. The results of this study revealed that azimuths with headphones were estimated well when compared to azimuth estimations made in the free field. It was found that elevations were generally estimated poorer than azimuths.

Lounsbury and Butler (1979) found that depths were estimated better with headphones for high pass sounds (>4.0 kHz) than for low pass sounds (<1.0 kHz). In this study, other trials were made where azimuths were varied, placing sound sources at azimuths of: $360°, 330°, 300°$ and $270°$. From these trials, Lounsbury and Butler (1979) found proficiency at $330°$. Another finding from this study was that depth estimations were unclear when the intensity cue was excluded as a depths cue.

The studies done by Lounsbury and Butler (1979) and Wenzel *et al.* (1991) provide a good indication of what is to be expected when using headphones. In audio augmented point cloud processing the azimuth of an audio object could be determined better than the elevation. The accuracy of azimuth perception could depend on the audio object's azimuth as suggested. In depth estimations, intensity should not be excluded as a depth cue.

3.5 Discussion

The manner in which the human auditory system perceives sound has been studied. Various audio cues were discussed. Furthermore, the type of information these cues retrieve for the listener were discussed.

The limitations of these cues were stated. This also included other factors that could limit the cues is sound perception. These limitations were then stated in the context of headphones and audio perception will be affected.

Point cloud processing with audio augmentation will make use of stated audio cues perceived with headphones. The stated limitations should be noted in the context of point cloud processing. As the audio cues will be exploited in point cloud processing, their limitations should be avoided where possible.

Chapter 4

Auditory Interface Implementation

An auditory interface was needed for the purpose of extracting spatial information from point clouds using audio. This interface was created by augmenting audio into a point cloud. The process is explained here.

In this context, an auditory interface is an environment that has sound sources emitting audio, with a listener object/virtual listener created to receive this audio. A resource that was used for this, OpenAL (Open Audio Library), is briefly introduced below. The use of OpenAL in creating auditory interfaces in this project will also be highlighted, this is done in section 4.2.

4.1 OpenAL – Open Audio Library

An auditory interface was created for the point cloud using OpenAL. Creative-Labs (2010) described OpenAL as "a cross-platform 3D audio API (application programming interface) used with gaming applications and other audio applications."

OpenAL has mostly been used in gaming applications because of its ability to provide surround sound (Creative-Labs, 2010). Wozniewski and Settel (2007, pg. 1) observed that OpenAL is "purely concerned with the spatialisation of sounds located in the scene, as a result the audio experience is focused around one user who indeed is immersed in 3D sound." It is for this reason that OpenAL was used in this work.

A few key concepts about OpenAL need to be defined. These are, *The Listener Object*, *The Sound Source Object*, *The Buffer Object* and *The Audio Context*. These definitions are provided below. Firstly, the OpenAL coordinate system is explained.

OpenAL Coordinate System

OpenAL emulates real aural environments. In doing this, the listener and the sound source objects, to be defined later, have to be placed in three dimensional space relative to each other. OpenAL uses a right-handed Cartesian coordinate system (RHS) to define the spatial attributes of the listener and sound source objects.

In a default frontal view, the X-axis points to the right, the Y-axis points up and the Z-axis points towards the viewer (Creative-Labs, 2010). Figure 4.1 illustrates the use of the RHS by OpenAL. In this figure, a listener object is shown with one sound source object in the scene.

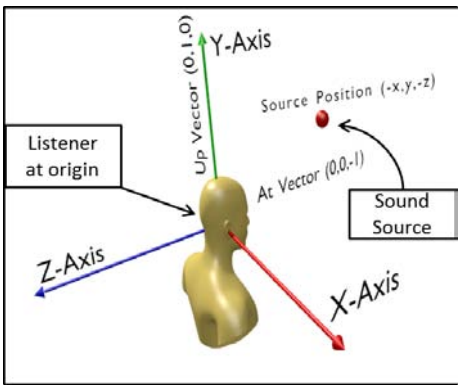


Figure 4.1: A virtual listener and a sound source are shown here. The listener is placed at the origin while the sound source's *3D position* is such that the x and z components are negative and the y component is positive. The listener's orientation is such that its *up vector* is (0,1,0) and its *at vector* is (0,0,-1).

The Listener Object

The listener object was designed to emulate the aural perception of human beings. As a result the OpenAL listener has a *3D position* which can be set.

This *position*, combined with the *position* of the sound source(s) (see below), influences the aural experience the listener receives when audio is emitted.

The listener has other attributes that have to be set. The *master gain* attribute controls the loudness with which the listener experiences emitted audio. If the *master gain* is set to the value of zero, no audio will be heard by the listener. A sound source has a *gain* attribute as well and this also influences the listener's aural experience – this will be mentioned in the text to follow. The listener also has a *velocity vector* that defines its velocity in 3D space relative to the sound sources.

Lastly, the listener has an *orientation vector* to define its orientation in 3D space. The *orientation vector* is split into two vectors, each with three elements: the *up vector* and the *at vector*. The *up vector* defines which way *up* the listener is directed. The *at vector* defines the direction the listener is looking *at*. Figure 4.1 demonstrates some of these listener attributes.

The virtual listener is synced with the user through an audio output device. In this work, headphones were chosen as the most appropriate manner for the user to receive audio. The headphones are connected to a computer which in turn is connected to a screen displaying the visual interface. (See figure 4.5.)

The Sound Source Object

The sound source object is the source of the emitted audio and received by the virtual listener. Only audio whose digital format is supported by OpenAL can be used. The *position* of a sound source can be set in 3D space. This influences the aural experience of a virtual listener when sound is emitted.

The sound source object has other attributes which can be set. The *pitch multiplier* attribute allows the user to change the *pitch/frequency*. The *source gain* helps in adjusting the *gain/volume*. A *source gain* of zero means that a source will not be heard at all when audio is emitted. The *source gain* influences the intensity with which a virtual listener experiences the emitted sound, this is in addition to the virtual listener's *master gain*, as previously mentioned.

Similar to the listener object, the sound source object also has a *velocity vector*. A sound source's *velocity vector* sets its *velocity* relative to that of

the virtual listener in 3D space. The *direction vector* of a sound source sets the *direction* which a sound source is facing. This influences how well the virtual listener will hear the audio. For example, a source directed towards a listener will be heard different to how it would be heard if it were facing away from the virtual listener.

The Buffer Object

The buffer object stores sample audio data that can be emitted by the audio source. The supported data format for a buffer is the PCM (Pulse code modulation), which is ubiquitous as far as digital audio format is concerned.

Buffer objects, unlike the listener and sound source objects, have no spatial attributes. Buffers have the following attributes which can be set: *frequency*, *size*, *bits* and *number of channels*.

The *frequency* is in samples per seconds and measured in hertz (Hz). The *size* attribute represents the size in bytes of the stored audio data. The *bits* attribute represents the number of bits per sample for the audio data (Creative-Labs, 2010). Lastly, the *channels* attribute represents the number of channels for the buffer's audio data.

Only single channel audio data will be used in this work. The reason for this is that multi-channel audio data cannot be used in position estimation tasks. For example, audio data with two channels (stereo) will always be experienced in the same way by the listener's ears even if it is placed closer to one ear, in which case the experience at each ear would be different if it were single channel.

The Audio Context

In a computer program that uses OpenAL and its functions, a call to open a sound device/card needs to be made. The reason for this is so that the sound device acts to process the audio output and play it through available and preferred output device. The available sound device on a machine will then be opened for use.

For audio to be emitted, the sound device will have to be associated with an OpenAL audio context. OpenAL context is created when in a computer

program, listener and sound source objects are created. The creation of an OpenAL context allows for the rendering of audio through sound sources.

Figure 4.2 (after Hiebert (2007)) shows the parts needed in making OpenAL fully functional. Each sound source has to reference a buffer to use its audio data. A listener object has to be created to receive the audio. With listener and sound source objects created, an OpenAL context will have been created. Finally, at execution of the computer program, a call will be made to open a sound device and associate the OpenAL context with it. At this stage OpenAL will be fully functional.

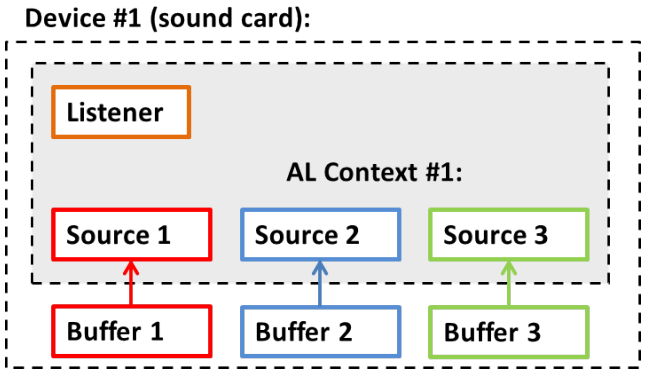


Figure 4.2: Objects (Listener, Source(s) and Buffer(s)). At initialisation, one audio device is opened. One OpenAL context is created, which will have only one virtual listener and one or many sound sources. Each buffer is attached to a sound source, providing it with audio data to emit. After Hiebert, 2007.

The text to follow will explain how OpenAL was employed in this work to create an auditory interface and augment audio into a point cloud.

4.2 Creation of an Auditory Interface

An auditory interface was created for the point cloud – this was within one OpenAL context. The objective here was to turn objects (clusters of points) in point clouds into sound sources and to have a listener object that will listen to the emitted sound.

The point cloud was partitioned using an octree. Figure 4.3 demonstrates the octree subdivision principle by Girardeau-Montaut *et al.* (2005). A bounding

box, containing points is subdivided into eight equal cubes. Each of the eight cubes is subdivided into eight smaller but equal cubes if it still has points or if a pre-set condition is not yet met.

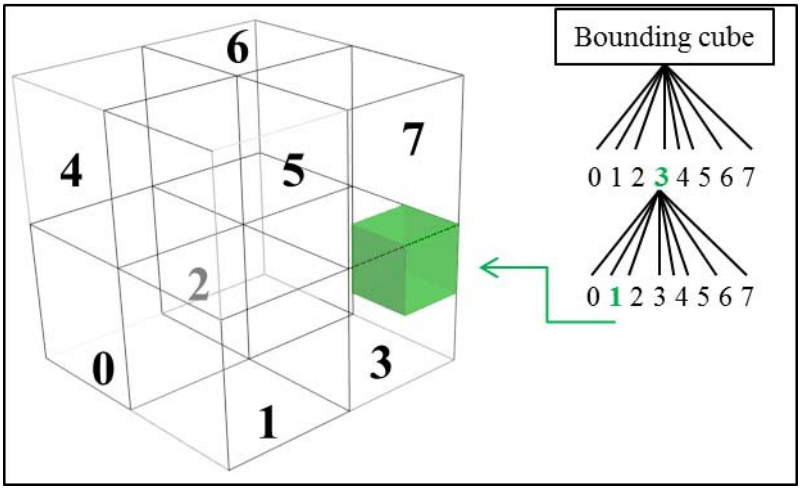


Figure 4.3: Octree subdivision principle. At the first level, the bounding box was subdivided into eight smaller cubes, labelled 0 to 7 in the figure. The green cube shown in this octree occurred at the second subdivision level, where cube labelled 3 was subdivided.

The purpose of this partitioning was to allow for unconnected objects represented in the point cloud to be treated as separate objects. These separate objects are taken as nodes whose purpose will be explained later. Using the octree, the point cloud was partitioned using the *by node width* method.

Figure 4.4 illustrates point cloud partitioning using an octree. Panel (a) of this figure shows the point cloud, whereas panel (b) shows the point cloud partitioned *by node width*, where the node width was set to 2.5 m. Two separate objects are shown in this figure. Object 1 has multiple nodes where the points contained in it are connected (are in close proximity to each other). Object 2 has a single node because the cluster of points contained in it are separated from other clusters of points.

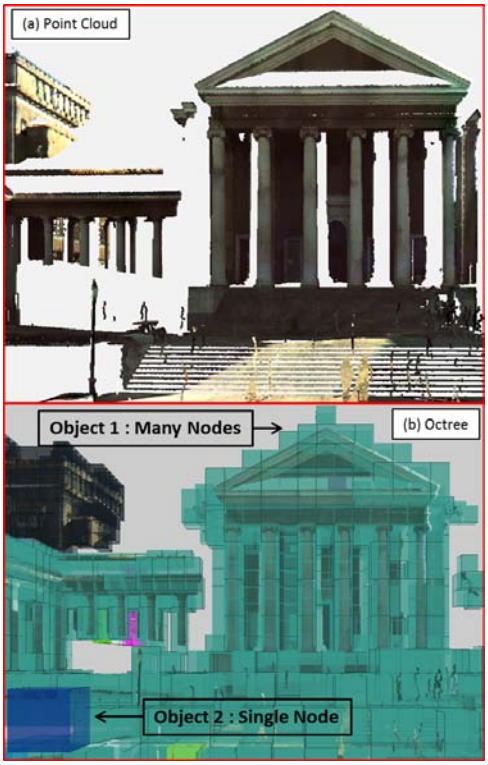


Figure 4.4: (a) A point cloud, (b) A point cloud partitioned *by node width* of 2.5 m using an octree.

With the point cloud partitioned, the objects/nodes in the octree can now be treated as sound sources. Sound sources can be assigned to the centroids of objects. Sound sources were then attached to buffers. The relationship here was such that one buffer was attached to one sound source (see figure 4.2). Each sound source can then be unique from the others. The virtual listener was placed at the origin of the OpenAL coordinate system.

4.3 Experimental Set-up

The experimental set-up adopted here will be explained using figure 4.5. In this figure, the following are shown:

1. The tester wearing headphones (output device).
2. The computer screen (output device) displaying the visual interface.

3. The input devices – mouse and keyboard.

The tester here is wearing Logitech's G35 surround sound headphones to receive emitted audio. (Refer to Logitech (2013) for a technical specification of these headphones.)

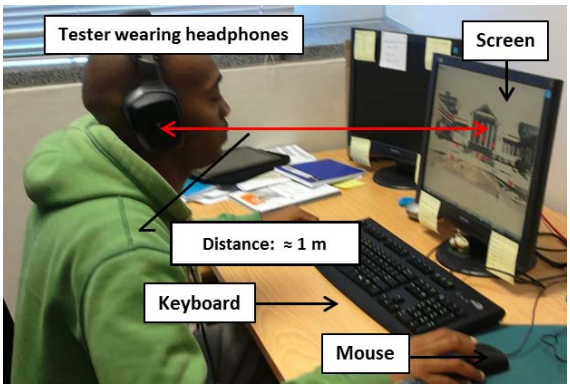


Figure 4.5: The test subject wearing Logitech G35 headphones while interacting with the auditory interface using input devices.

Interaction with the audio augmented point cloud to carry out the tests was done using response locations, explained using figure 4.6. In this figure, an octree partitioned point cloud is shown (the octree is omitted). Five squares are also shown in the figure. These squares are screen projections of five randomly selected octree nodes. These squares offer the tester the means of interacting with the audio augmented point cloud and hereinafter will be referred to as *response locations*. (Figure 4.6 can be assumed to be what the computer screen is displaying to the tester in figure 4.5.)

A sound emitting source can be associated with a response location. The use of response locations offers the tester the ability to determine spatial information of the associated sound source. As shown in figure 4.5, the tester is positioned about 1 m from the computer screen and aligned centrally to it. This arrangement allows the tester to be positioned relative to response locations and to be able to judge sound source spatial information based on this arrangement; this is particularly important for position estimation tests.

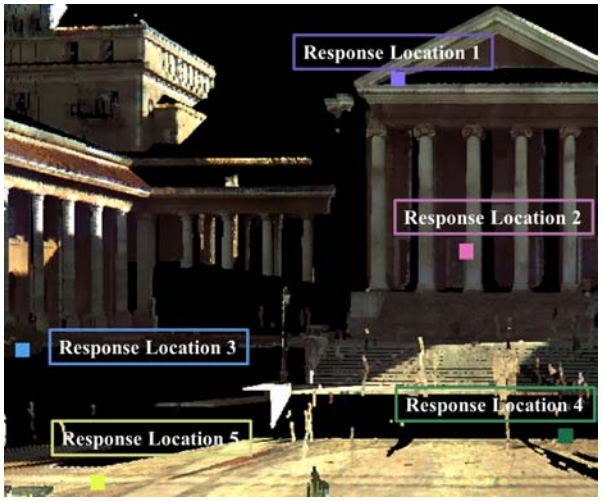
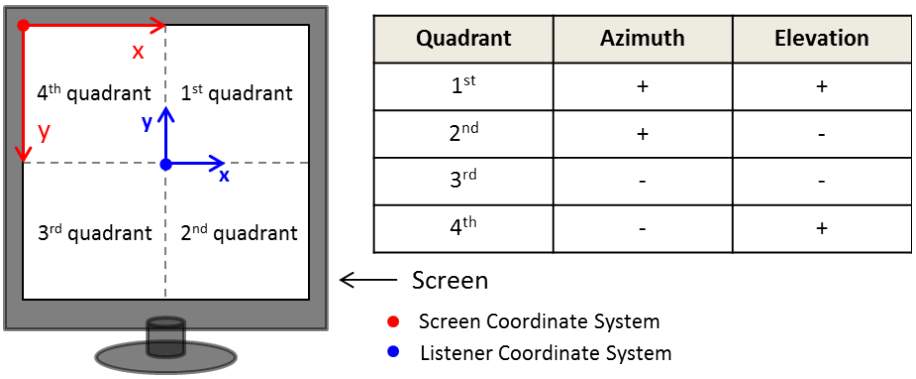


Figure 4.6: A point cloud with randomly chosen nodes made response locations. The octree is omitted in this figure.

4.3.1 Implementation of Position Estimation in Point Clouds

The response locations are presented in the screen's coordinate system. What are needed, however, are the positions of response locations in terms of azimuths and elevations with respect to the tester (and the virtual listener). For this, the positions of the response locations were defined according to the 2D virtual listener system, as shown in figure 4.7 by x and y axes shown in blue.



Quadrant	Azimuth	Elevation
1st	+	+
2nd	+	-
3rd	-	-
4th	-	+

Figure 4.7: The 2D coordinate system of the screen.

The 2D virtual listener coordinate system is a result of a transformation

(rotation plus translation) of the screen's coordinate system. A point's 2D position in the listener coordinate system is therefore defined by equation 4.1,

$$\begin{pmatrix} x_{listener} \\ y_{listener} \end{pmatrix} = \begin{pmatrix} h_w \\ h_h \end{pmatrix} + R \times \begin{pmatrix} x_{screen} \\ y_{screen} \end{pmatrix} \tag{4.1}$$

where, $(x_{listener}, y_{listener})$ is the 2D position in the listener's coordinate system, (h_w, h_h) are the translation parameters where h_w is half the screen width and h_h is half the screen height, (x_{screen}, y_{screen}) is a point's position in the screen coordinate system, and R is the rotation to be applied, defined by equation 4.2.

$$R = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{4.2}$$

The virtual listener's 2D coordinate system was made such that its origin is at the centre of the screen. The screen was divided into four quadrants about its centre. A response location's azimuth and elevation therefore depend on the following two properties:

(a) How far the response location is from the origin.

(b) The quadrant the response location is located in.

The 2D position of a response location on the screen is relative to the tester, therefore corresponds to its 2D position relative to the virtual listener in 3D space. The table in figure 4.7 shows what azimuth and elevation signs a sound source will have in each quadrant. This method of testing is similar to that done by Razavi *et al.* (2005).

4.3.2 Implementation of Depth Estimation in Point Clouds

The depth referred to here is not the euclidean depth, it is the depth along Z-axis (median plane) where the virtual listener is placed. The euclidean depth is given by equation 4.3,

$$d(x, y, z) = \sqrt{(x_l - x_s)^2 + (y_l - y_s)^2 + (z_l - z_s)^2} \tag{4.3}$$

where, x_l, y_l, z_l are the 3D coordinates of the virtual listener and x_s, y_s, z_s are the 3D coordinates of the sound source. The virtual listener is placed at the origin of the OpenAL coordinate system for the tests, therefore x_l, y_l, z_l are all 0.0. Because the x and y coordinates of the target sound source will keep changing, as will be explained later, the easiest depth to help explain aspects of this set of tests is the one along the Z-axis and will simply be referred to as *depth*.

Figure 4.8 shows in top view, audio sources placed at depths A, B and C. This demonstrates that the euclidean depth will be different depending on the depth and the x and y coordinates of the sound source. Sound source intensities are attenuated by the euclidean depth.

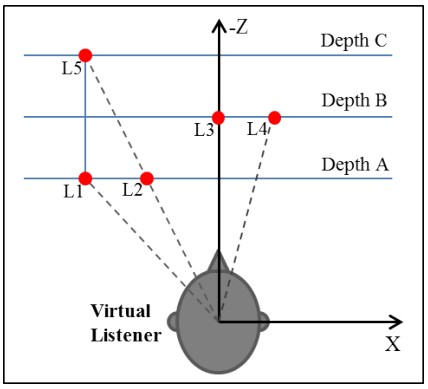


Figure 4.8: Top view of sound source's euclidean depth changing as a function of depth (z coordinate) and x coordinate. The red circles indicate different positions occupied by a sound source.

4.3.3 Implementation of Occupancy of Space in Point Clouds

OpenAL has means of creating reverberant environments for auditory interfaces created through its effects extensions (EFX). Aural experiences from real reverberant environments are simulated in this manner in auditory interfaces. The aural experience from OpenAL depends on the type of reverberant

environment simulated. OpenAL has a number of reverberant environments that can be used and these can be found from Creative-Labs (2010). In this investigation, only the following environments were chosen:

(1) *Cave environment* – reverberation typical of caves.

(2) *Hallway environment* – reverberation experienced in hallways.

(3) *Mountain environment* – reverberation of a mountainous environment, a large open area.

(4) *Room environment* – reverberation of a normal room.

A *waterdrop* wave sound was used as a sound of choice for this investigation. In these tests, the environment the virtual listener was immersed in was randomly picked from the four listed environments. The test subject therefore did not know which environment the virtual listener was immersed in and had to make that decision based on the aural experience when sound was emitted. This aim here was to test if reverberations were experienced by the test subject.

The mouse and the keyboard act as input devices for the tester to interact with the software. This is done in a similar way as in Edwards's (1989) work. The mouse was used particularly in position estimation tests, whereas the keyboard was used in depth and environment ambience tests. These interactions will be explained further when the respective tests are discussed. Through all the tests, one target sound source whose spatial information need to be determined was be used. The type of audio emitted by this sound source is mentioned when each problem is discussed.

Chapter 5

Limitations of Audio Augmented Processing

This chapter is split according to each research problem to be investigated in sections 5.1, 5.2 and 5.3. Each of these sections are split into subsections which explain how the tests were done and how the data arising from the tests will be analysed in Chapter 6.

5.1 Sound Source Position Estimation

The first research problem as outlined in Chapter 1 is about position estimation points in audio augmented point clouds for their processing. This is in terms of azimuth and elevation of audio objects. The steps taken to do this investigation are explained here.

5.1.1 Tests for the Sound Source Position Estimation

Position estimation of the target sound source was tested as a function of the following properties:

1. Number of response locations.
2. Number of noise sound sources.
3. Azimuthal and Elevational Spatial Distributions of Response Locations.

Position Estimation as a Function of Response Locations

The tester was required to estimate the position of the target sound source when presented with a varying number possible response locations a sound source can be associated with. The purpose of this was to investigate the user's ability to locate or position an audio augmented point in the point cloud when numerous candidates exist. Six tests (25 trials each) were done in this manner, where the tester was presented with 2, 3, 4, 5, 6 and 7 response locations in each test, where the target sound source was located at only one of them.

Position Estimation as a Function of Number of Noise Sources

The purpose here was to investigate position estimation limitations in a situation where various areas of interest in the point cloud are each attached to an audio object and all are emitting audio at the same time. If the user wants to focus on one audio object, the others could become noise and possibly hinder the position estimation. Here too, the tester was presented with 2, 3, 4, 5, 6 and 7 response locations in each test. In the first instance, in each of these tests a noise sound source was added. In the second instance, two noise sound sources were added. The tester therefore needed to estimate the position of the target sound source in the presence of one or two noise sources.

Figure 5.1 illustrates a situation in the third instance where two noise sources were presented together with the target sound source. The noise sources occupied positions indicated with this symbol: $\times$. The target sound source could occupy any of the seven response locations as marked by this symbol: $+$. The tester was required to click on the correct response location to locate the target sound source while noise sources were emitting sound as well.

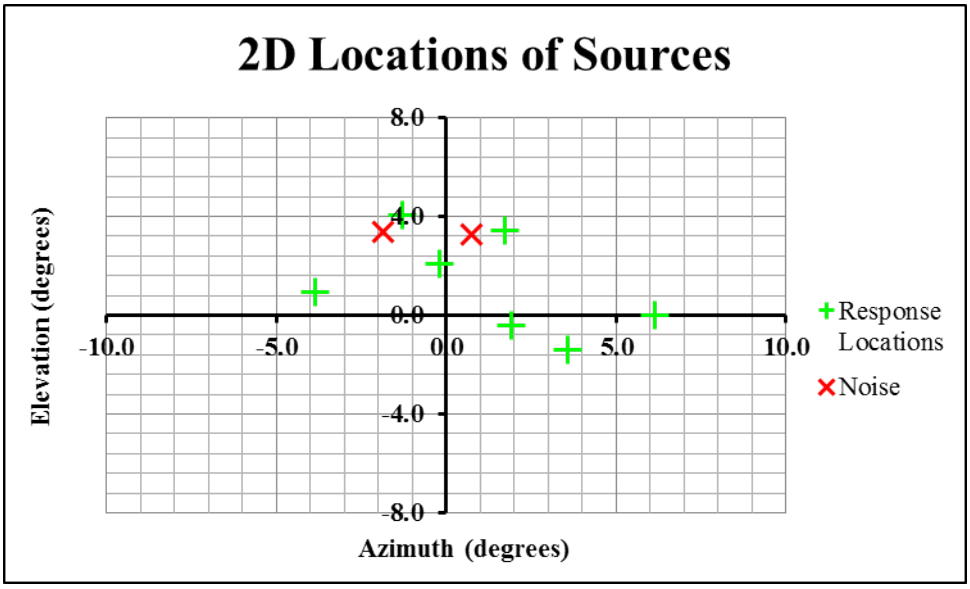


Figure 5.1: Two noise sources with seven possible response locations for the target sound source.

Position Estimation as a Function of Azimuthal and Elevational Spatial Distributions of Response Locations

In this set of tests the tester is presented with the same number of response locations as in the other two sets. The investigation here is on how the distribution of response locations on the screen in terms of azimuth and elevation influence position estimation. This investigates position estimation an audio augmented point when it is among a cluster of other points that could be augmented with audio.

In all the tests, the tester had to click on the presented response locations to estimate the position of the target sound source. Carlile *et al.* (1997, pg. 180) referred to this method of testing by stating that the "most straightforward means of indicating the perceived locations was for the subject to identify the speaker to be generating the sound." There was a total of 25 trials in each test. The total number of tests done was 18, where each position estimation property had six tests.

The intensity of the target sound source was programmatically made not to be attenuated with depth. Intensity changes were therefore only related to the location of the sound source with respect to the virtual listener in both azimuth and elevation and not related to depth changes. A sine tone gen-

erated using the Audacity software (Audacity-Team, 2012) was loaded in a buffer that was used by the target sound object.

5.1.2 Analyses of the Position Estimation Test Results

Azimuth difference is the difference between the actual azimuth of the target sound source and the user's estimation of the azimuth. Similarly, elevation difference is the difference between the actual elevation of the target sound source and the user's estimation of the elevation. The azimuth and the elevation differences are used to determine if the position of the sound source was properly estimated or not through interaction with the response location.

Figures 5.2 and 5.3 best illustrate how the tests will be carried out and the expected format of the results. As sound is being emitted from a source, the test subject will click on the response location that is potentially 'housing' the target sound source. Two possible scenarios will arise from each mouse click and they are as follows:

- Scenario 1: If the correct response location is clicked on, then both the azimuth and elevation differences will be zero, meaning that the target sound source has been found. Figure 5.2 illustrates this idea.

- Scenario 2: If the incorrect response location is clicked on, then both the azimuth and elevation differences will not be zero, meaning that the target sound source was not located. Figure 5.3 illustrates this idea.

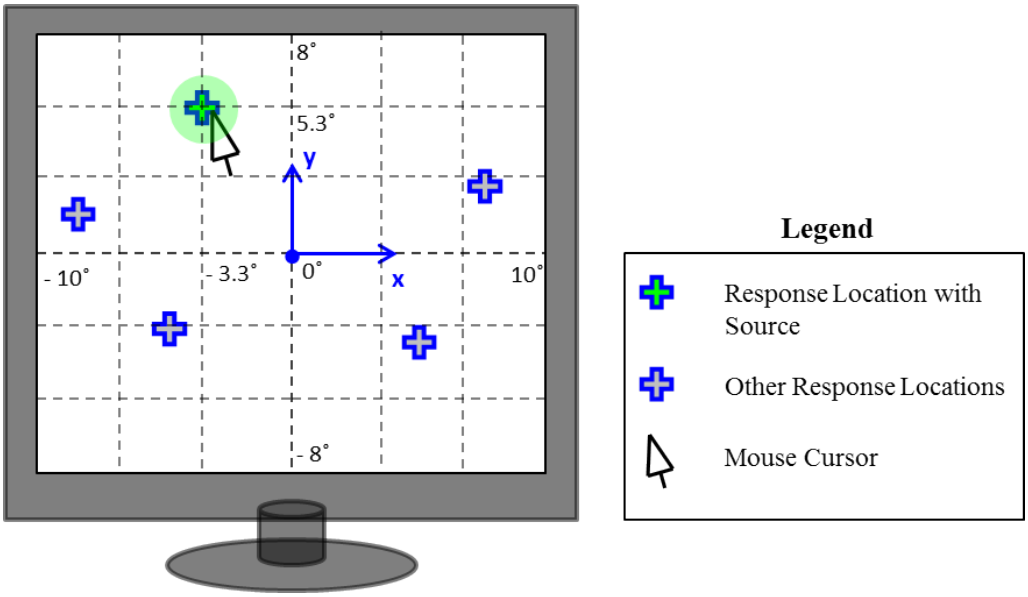


Figure 5.2: Correct position estimation made by the tester. Both azimuth and elevation differences are zero.

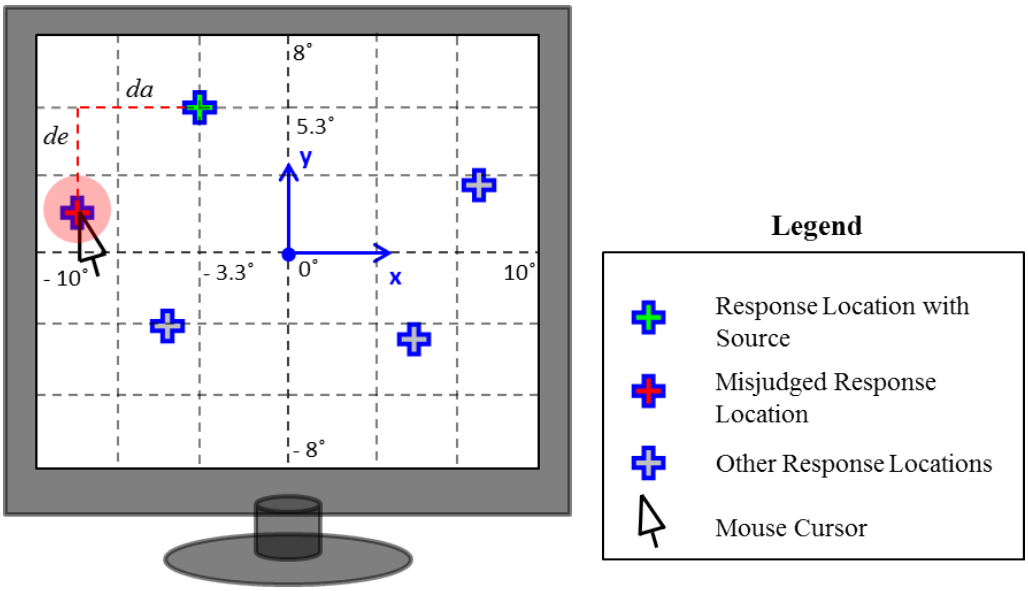


Figure 5.3: Incorrect position estimation made by the tester. Both azimuth (da) and elevation (de) differences are non-zero.

The azimuth and elevation differences arising from the mouse clicks will then give an indication of how well position estimation was done in terms of the

target sound object's azimuth and elevation. Figure 5.2 shows a situation where correct position estimation is made by the tester. However, figure 5.3 demonstrates a situation where incorrect position estimation is made, resulting in azimuth and elevation differences. In this example, the target sound source is at azimuth and elevation position of (−3.3°,5.3°) but the tester estimated it to be at about (−8.3°,1.3°). This results in error of estimation of (−5°,−4°), shown using *da* and *de* (azimuth and elevation differences) symbols in the figure. Multiple tests were carried out in this manner and the results from each test will be tabled and presented in Chapter 6.

5.2 Sound Source Depth Estimation

The second problem entails the estimation of depth/distance of/to sound sources from the virtual listener. This will investigate how depth to a an audio augmented area of interest can be determined. The steps carried out in doing this investigation are explained here.

5.2.1 Tests for the Sound Source Depth Estimation

The aim here is to determine the depth of the target sound object at various response locations on the screen. (As in the position estimation test, the target sound source used was the sine tone.) Unlike in the previous problem, one response location at a time would appear on the screen, rather than multiple. The displaying of a response location on the screen was done systematically – this is illustrated using figure 5.4.

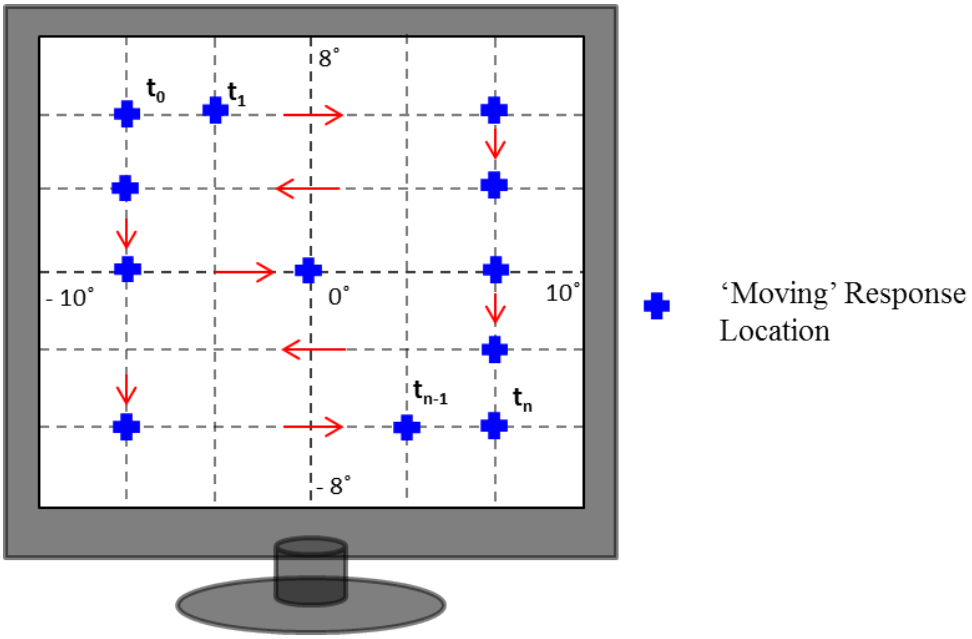


Figure 5.4: Displaying of one response location on the screen changing with time.

Shown in figure 5.4 is a response location 'moving' with time. At time t_0, the response location appears at the first location on the screen. After a set number of seconds, the response location appears at the second location, this is at time t_1. This process continues until the response location appears at the last location at time t_n, where n is the total number of locations where the response location needs to appear, minus one. The arrows in the figure indicate the order of appearance of the response location on the screen. As previously stated, the target sound source had the same coordinates as those of the node shown as the response location appearing at that time. This was done so that depth tests can be done at as many locations as possible.

At each response location, the target audio source would have a randomly chosen depth away from the virtual listener. Depth was chosen from two sets containing three depth values each: {-0.75, -1.5, -2.25}m and {-1.0, -1.75, -2.5}m. (Hereinafter, these sets will simply be referred to as Set A and Set B, respectively.) The reason for making this random was so that the tester would have to estimate each time what the depth was. Determination of depths of sound sources is a crude exercise, as previously noted. This is why the target sound source was not made to vary between multiple depths with small variations between them. The idea here was to estimate which depth the target sound source had.

In Chapter 3, it was stated that the intensity of a sound source changes in relation to the depth from the listener. It was then noted that this change of intensity can act as a cue to provide sound source depth information. In the process of this research, it emerged that for this work, intensity alone is not enough to act as a depth cue. As a result, pitch variation was used to act as a depth cue too. This means that at different depths, the sine tone (target sound source) had different frequencies, leading to these pitch variations. Figure 5.5 helps explain this more. (Figure 5.5 shows frequency-depth relationship for Set A depths only, the same relationship exists for Set B depths and was omitted here.)

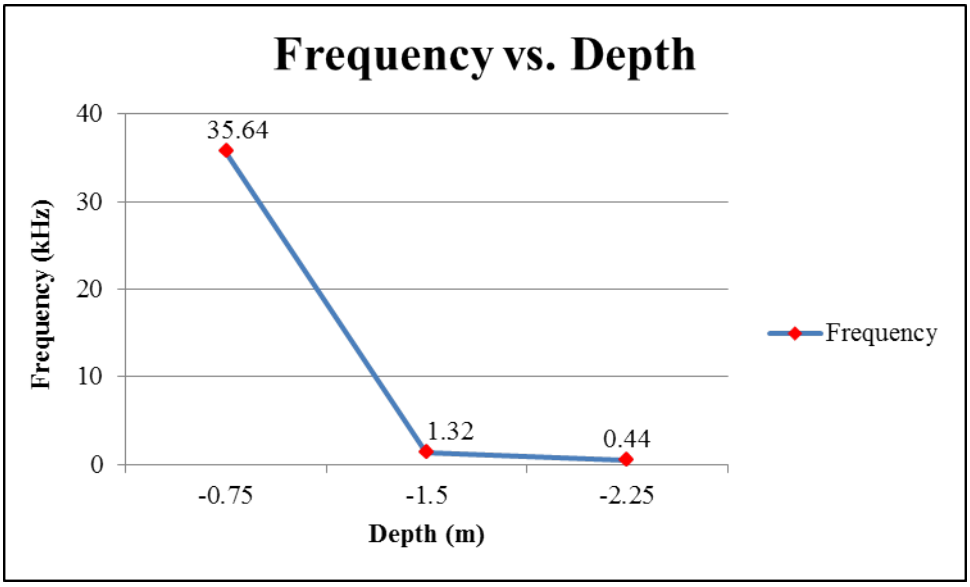


Figure 5.5: The Frequency v. Depth curve.

Figure 5.5 shows different frequencies used for different depth values. The sine tone was generated with a default frequency of 0.44 kHz (440 Hz) using the Audacity software. As an indicator of closeness of the target sound source to the virtual listener, the frequency was increased as the depth decreased (see figure 5.5). The frequency-depth pairing was therefore as follows for both sets of depths: {35.64:-0.75, 1.32:-1.5, 0.44:-2.25} and {35.64:-1.0, 1.32:-1.75, 0.44:-2.5} (kHz:m). The frequencies were calculated as indicated by equation 5.1. These frequency values were chosen because between them there were good noticeable aural differences, i.e., the pitch changes were not discreet.

$$
\begin{aligned}
freq_{(-0.75)} &= 0.44 \times 3.0^4 \\
freq_{(-1.5)} &= 0.44 \times 3.0^1 \\
freq_{(-2.25)} &= 0.44 \times 3.0^0
\end{aligned} \tag{5.1}
$$

Given that the target sound source had different intensity and pitch (frequency) at each depth, the tester had to make depth estimations based on these depth signatures. These estimations did not depend on the tester interacting with software with the mouse as in the position estimation problem. The tester simply had to press a number on the keyboard to make the depth estimations. The choices were as follows for Set A: key 0 for depth -0.75 m, key 1 for depth -1.5 m and key 2 for depth -2.25 m. Similarly, for Set B they were: key 0 for depth -1.0 m, key 1 for depth -1.75 m and key 2 for depth -2.5 m.

Depth estimations were made at a total of 72 response locations for each test. Figures 5.6 and 5.7 show response locations and the frequency-depth pairs that the target sound object had at each response location for first tests of Set A and Set B depths. (Figures for second and third tests for the respective sets show different frequency-depth pairs at different response locations. These are omitted here as they illustrate the same idea.) The sizes of the 'bubbles', i.e., response locations, indicate the depth value – the smallest 'bubble' for the smallest depth, etc.

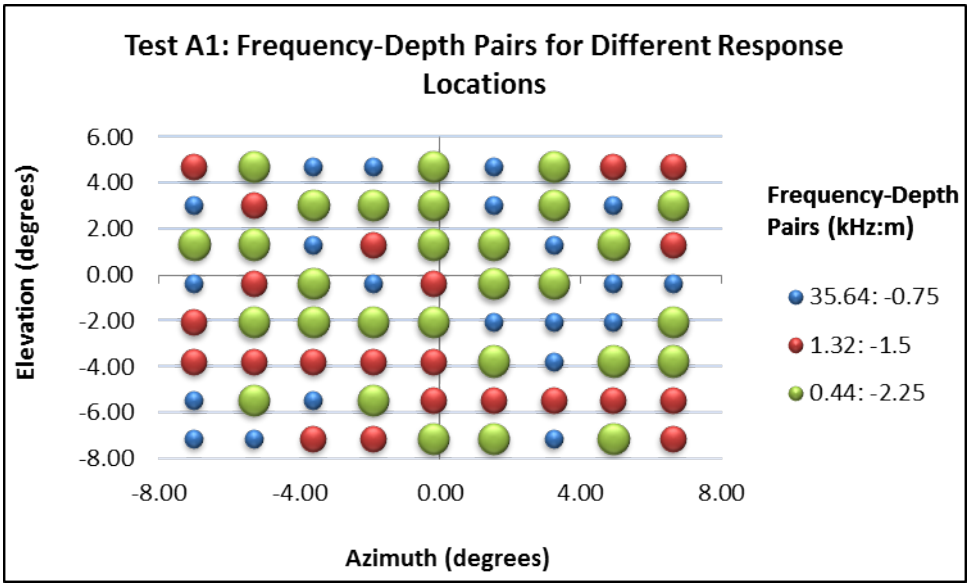


Figure 5.6: The Frequency-Depth pairs for different response locations for first test of set A depths.

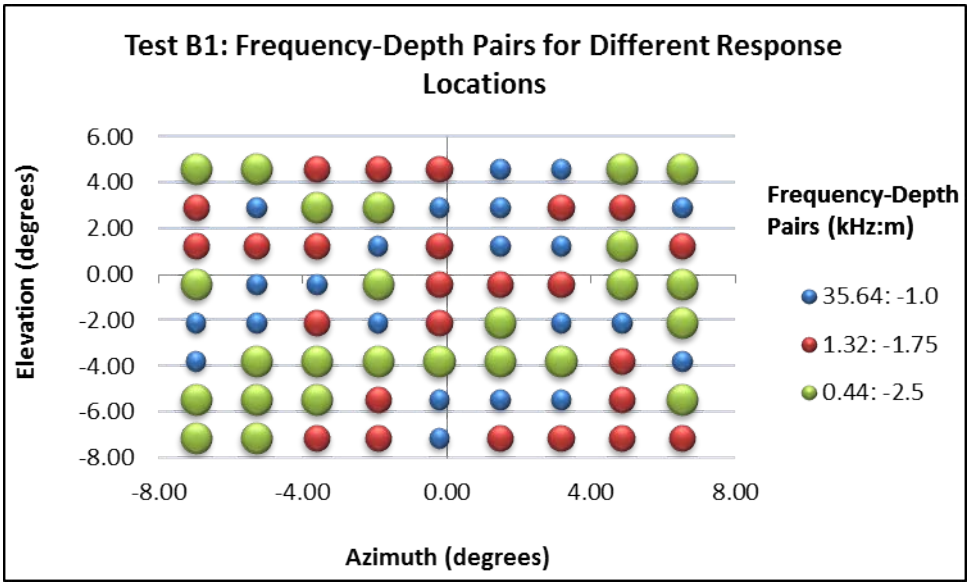


Figure 5.7: The Frequency-Depth pairs for different response locations for first test of set B depths.

5.2.2 Analyses for the Sound Source Depth Estimation Problem Test Results

The analysis of depth estimations will be made based on whether the depth was well-estimated or mis-estimated at a response location. If mis-estimated, it will be indicated whether this was an underestimation or an overestimation. Underestimation is when the depth was estimated to be lower than it actually is, and overestimation is when it was estimated to be higher than it actually is.

Furthermore, the magnitude of the mis-estimation will be shown for both underestimated and overestimated depths. Because of the depth values in these sets {-0.75 m, -1.5 m, -2.25 m} and {-1.0 m, -1.75 m, -2.5 m}, underestimations can only be -0.75 or -1.5 m, whereas overestimations can only be 0.75 m or 1.5 m. Figure 5.8 shows well-estimated and mis-estimated instances, with all types of mis-estimation shown. The sizes of the 'bubbles' indicate the absolute magnitudes of the mis-estimation; for well-estimated depths a magnitude of 1.0 was used to provide the sizes. What is shown in figure 5.8 serves as an example of the types of results to be expected for all tests.

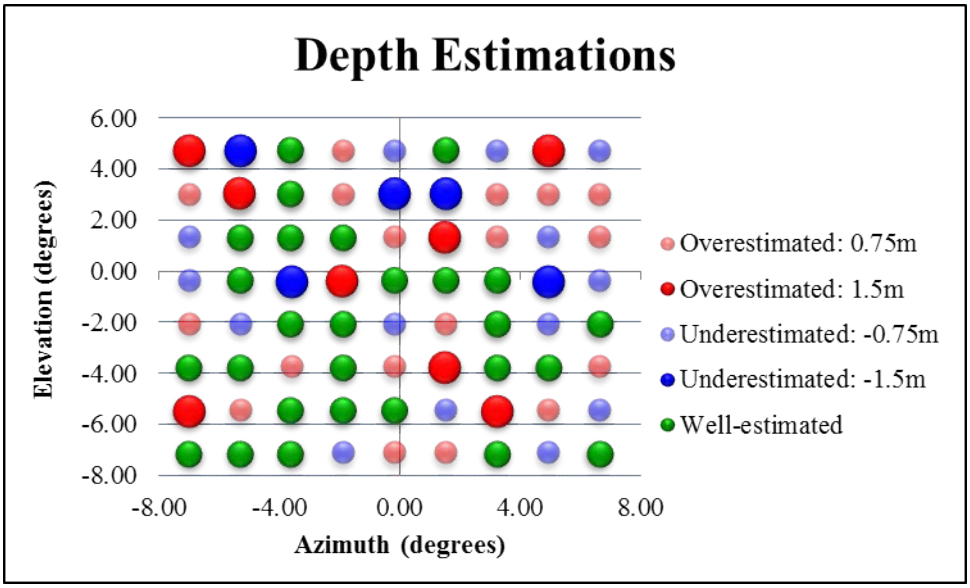


Figure 5.8: Expected format of depth estimation results. Shown are well-estimated and mis-estimated depths.

5.3 Occupancy of Space

Here the investigation is about whether the tester can determine the occupancy of space judging from its ambience when a sound is emitted. This could serve to provide information about existing objects which the user can not visually witness. This will investigate the existence of unseen events and their surroundings when the user processes a point cloud.

Three sets of tests with 50 trials each were done. In each trial, the test subject was required to make a judgement of which environment the virtual listener was immersed in by pressing a key on the keyboard. The keyboard was used as an input device.

Accurately judging the environment the listener is immersed in, indicates that the test subject does notice the reverberation. Moreover, it shows the tester's ability to identify the nature of those reflections in terms of whether they are from a closed or an open space. The results are provided in section 6.3, together with the analyses.

Chapter 6

Results and Analyses

In this chapter, the results from tests carried out as explained in Chapter 5 are given. The analyses of the results in terms of their implications in audio augmented point cloud processing are provided.

The results provided here are based on audio augmented to the Jameson Plaza point cloud. Jameson Plaza is located on the Upper Campus of the University of Cape Town. This point cloud was provided courtesy of the Zamani Project and was used here with permission.

6.1 Position Estimation of Target Sound Source

Position estimation was done in terms of the azimuths and elevations of the target sound source with respect to the virtual listener's position in 3D space. The objective of these tests was to estimate azimuths and elevations of the target sound source in the audio augmented point cloud.

6.1.1 Position Estimation as a Function of Response Locations

The tester was required to locate the target sound source when presented with a varying number possible response locations. Six tests (25 trials each) were done, where the tester was presented with 2, 3, 4, 5, 6 and 7 response locations in each test, where the target sound source was located at only one response location. The results given here are split into average azimuth errors and average elevation errors as a function of response locations.

Average Azimuth Errors as a Function of Response Locations

Figure 6.1 demonstrates average position estimation errors in azimuth changing as a function of the number of response locations. The minimum and maximum errors are indicated using error bars, which can be seen in the figure. Presented with the target sound source and two possible response locations, the tester could accurately locate the sound source in all 25 trials. This is demonstrated in the figure, where both the minimum and maximum errors were 0°, leading to average error of 0°.

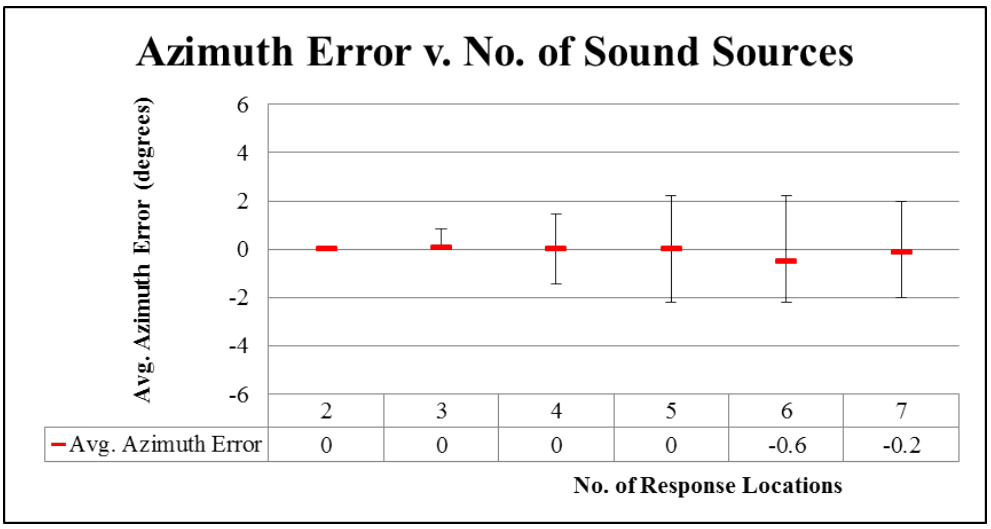


Figure 6.1: Average azimuth position estimation errors as a function of number of response locations. Largest average azimuth error ($-0.6°$) observed with six response locations.

The average azimuth errors changed with the number of response locations. However, this was not proportional as the average azimuth error when there were six possible response locations was the largest. This error was $-0.6°$, with minimum and maximum errors of $-2°$ and $2°$, respectively. It is larger than the average azimuth error for seven response locations, which was $0°$. Although for five response locations the average azimuth error was found to be $0°$, the minimum and maximum errors were found to be larger than for seven response locations. The reason for this and the lack of proportional change of average azimuth error with number of response locations will become evident in subsection 6.1.3.

These results indicate that the user can potentially locate an audio aug-

mented object well in terms of its azimuth in a situation where there are two objects that could be associated with the audio object. This accuracy changes as the number of points in a point cloud that could be associated with the audio object increases.

Average Elevation Errors as a Function of Response Locations

The average elevation errors experienced when the number of response locations increased are reported in figure 6.2. Minimum and maximum errors are indicated using error bars here as well. As with average azimuth errors, when presented with the target sound source with two response locations, the average elevation error made by the tester was 0°, with minimum and maximum errors both being 0°.

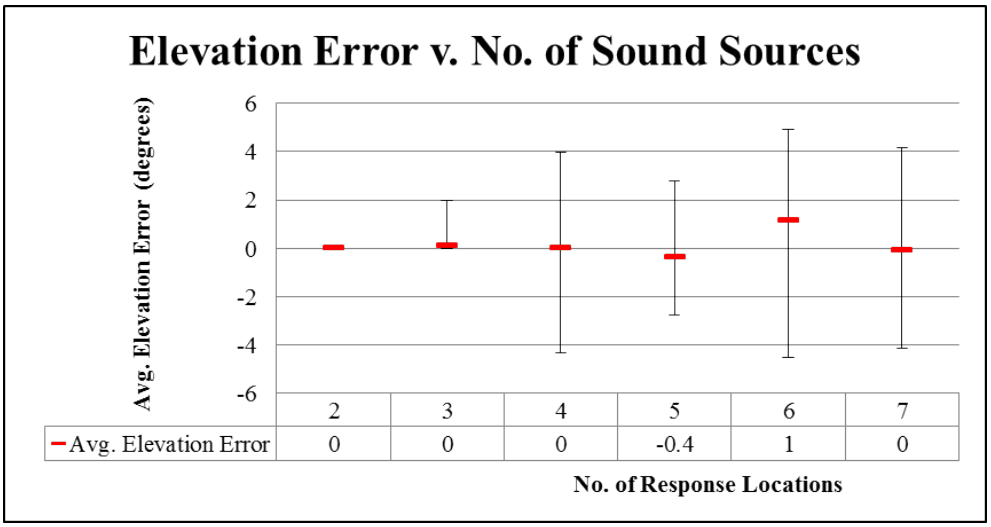


Figure 6.2: Average elevation position estimation errors as a function of number of response locations. Largest average elevation error (1°) observed with six response locations.

The average elevation error was found to change with the number of response locations in the same manner as the average azimuth error. In this case too, with six response locations the largest average elevation error was found. This error was 1°, with minimum and maximum errors of $-4.5°$ and 4.9°, respectively. The second largest average elevation error was $-0.4°$, which occurred with five response locations. Similarly, the results found for average

elevation errors will become clear in the discussion in subsection 6.1.3.

The implications of these results are similar to those observed in the azimuth problem. Where there are two points/areas that could be associated with an audio emitting object in the point cloud, the determination of the elevation is accurate. Here too the accuracy changes as the number of objects that could be associated with the audio object changes.

6.1.2 Position Estimation as a Function of the Number of Noise Sources

Position estimation of the target sound source was tested when there were other sound emitting sources – noise sources. The results are provided below.

Average Azimuth Errors as a Function of Response Location – with Noise Sound Sources

The average azimuth errors when there were no noise sources were reported in figure 6.1. In figure 6.3, the average azimuth errors for all response locations with one noise source introduced are shown. With two possible response locations, the tester could locate the target sound source accurately while a noise source was also emitted. In this case the average azimuth error was 0°, with both the minimum and maximum errors being 0° as well.

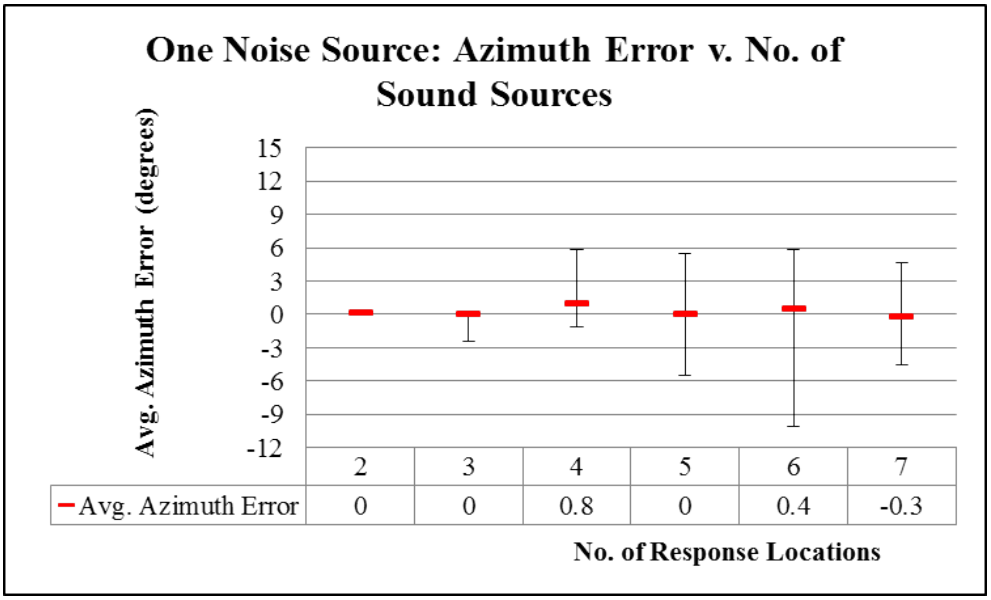


Figure 6.3: Average azimuth position estimation errors as a function of number of response locations – with one noise source. With one noise source present, the largest average azimuth error (0.8°) was observed with four response locations.

The largest average azimuth error occurred when there were four response locations. This error was 0.8°, with minimum and maximum errors of $-1°$ and 5.9°, respectively. This error is larger than the average error of $-0.6°$ which was the largest in the case where there were no noise sources. With six response locations, large minimum and maximum errors of $-10°$ and 5.8° were found, with average error of 0.4°, which was the second largest. As before, the average azimuth error was found to change with the number of response locations, albeit not changing proportionally.

The average azimuth errors found when two noise sources were presented with the target sound source are shown in figure 6.4. Unlike in previous cases, the smallest average error was found when there were five possible response locations, not when there were two. This value was 0°, where the minimum and maximum errors were $-3.9°$ and 5.4°, respectively. The largest average error occurred when there were six possible response locations. This error was 2°, with minimum and maximum errors of $-5.7°$ and 12.3°, respectively. This error is also larger than those in the previous two cases, $-0.6°$ and 0.8°.

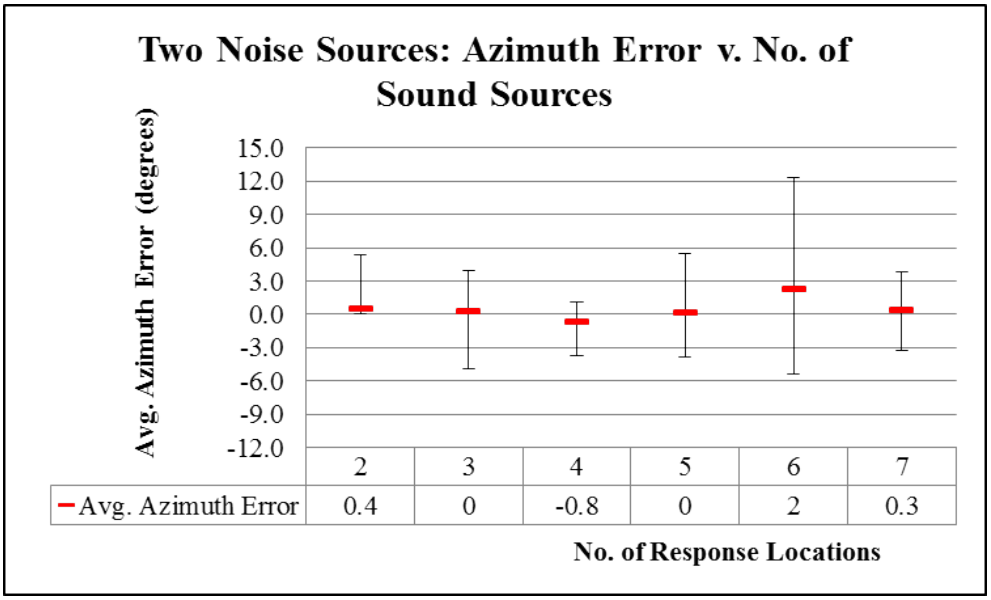


Figure 6.4: Average azimuth position estimation errors as a function of number of response locations – with two noise sources. With two noise sources present, the largest average azimuth error (2°) was observed with six response locations.

The existence of noise appears to influence position estimation in terms of azimuth. The results above suggest that in a case where the user needs to locate an audio augmented object in a point cloud in terms of azimuth, the accuracy will be affected by other audio emitting objects. This is more evident as when there are two noise sources in the audio augmented point cloud. The extent to which noise sources influence position estimation also depends on the number of objects in the point cloud that could be associated with the audio object.

Average Elevation Errors as a Function of Response Locations – with Noise Sound Sources

The average azimuth errors when there were no noise sources were reported in figure 6.1. The average elevation errors when the target sound source was presented with one noise sound source are given in figure 6.5.

When the target sound source was presented with two possible response locations and one noise source, no elevation errors were observed. The largest average elevation error of $-0.6°$, occurred when there were seven possible

response locations. The minimum and maximum errors in this instance were $-6.6°$ and $4°$, respectively. These minimum and maximum errors were also larger than the ones observed for other sets of response locations.

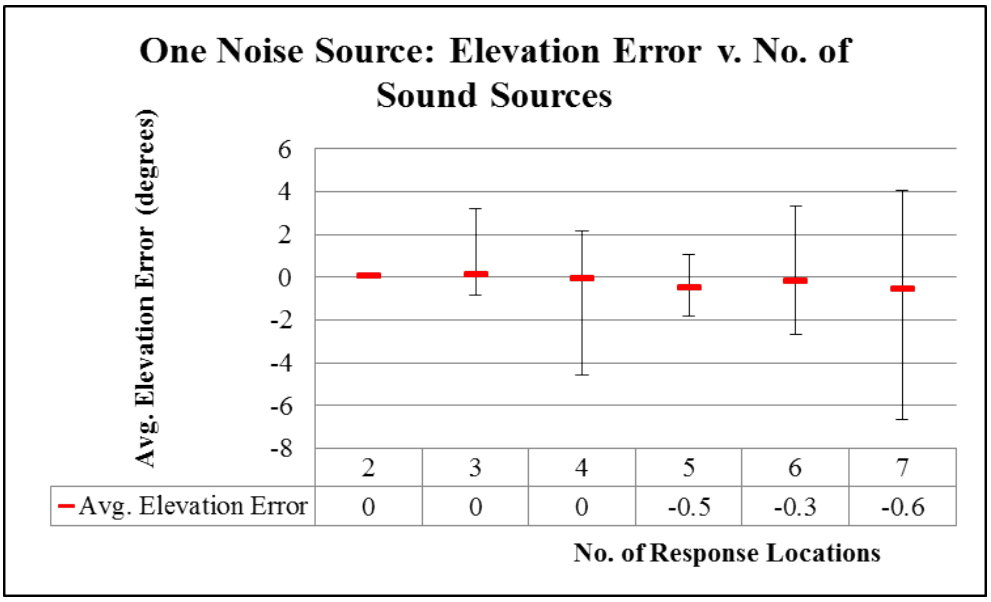


Figure 6.5: Average elevation position estimation errors as a function of number of response locations – with one noise source. With one noise source present, the largest average elevation error ($-0.6°$) was observed with seven response locations.

Shown in figure 6.6 are the average elevation errors observed when two noise sources were presented with the target sound source. The smallest average elevation error was observed for two response locations. This error was $0°$, where the minimum and maximum errors were $-0.5°$ and $0°$, respectively. In previous cases, the average elevation error for two response locations was $0°$, with both the minimum and maximum elevation errors being $0°$ as well.

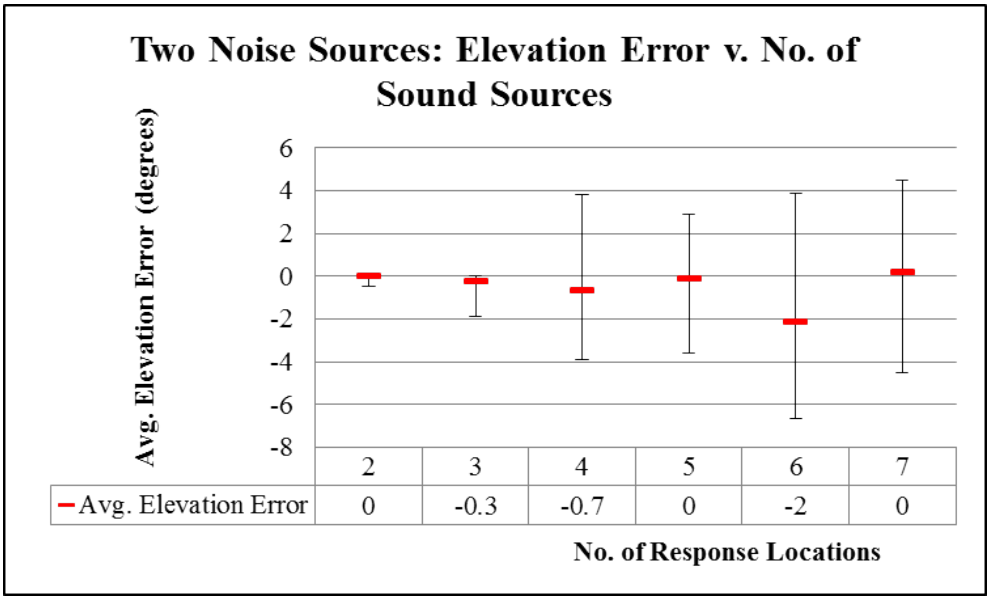


Figure 6.6: Average elevation position estimation errors as a function of number of response locations – with two noise sources. With two noise sources present, the largest average elevation error ($-2°$) was observed with six response locations.

The largest average elevation error of $-2°$, was observed when there were six possible response locations. The minimum error was $-6.7°$, while the maximum error was $3.9°$. This average elevation error is larger than in the other two cases (when there were no noise sources ($1°$) and when there was one ($-0.6°$)).

Based on the quoted results, with elevation too, having noise appears to influence the position estimation. As in the azimuth problem, this is more evident when there are two noise sources in the audio augmented point cloud. The number of objects that could be associated with the audio object also played a role here.

6.1.3 Position Estimation as a Function of Spatial Distribution of Response Locations

This investigation looked at whether the level of clustering of the response locations had an impact on target sound source position estimation. To illustrate this idea of spatial distribution, figure 6.7 shows four relatively spread

out response locations and one noise source. In contrast, figure 6.8 shows greater clustering of four response locations and two noise sources.

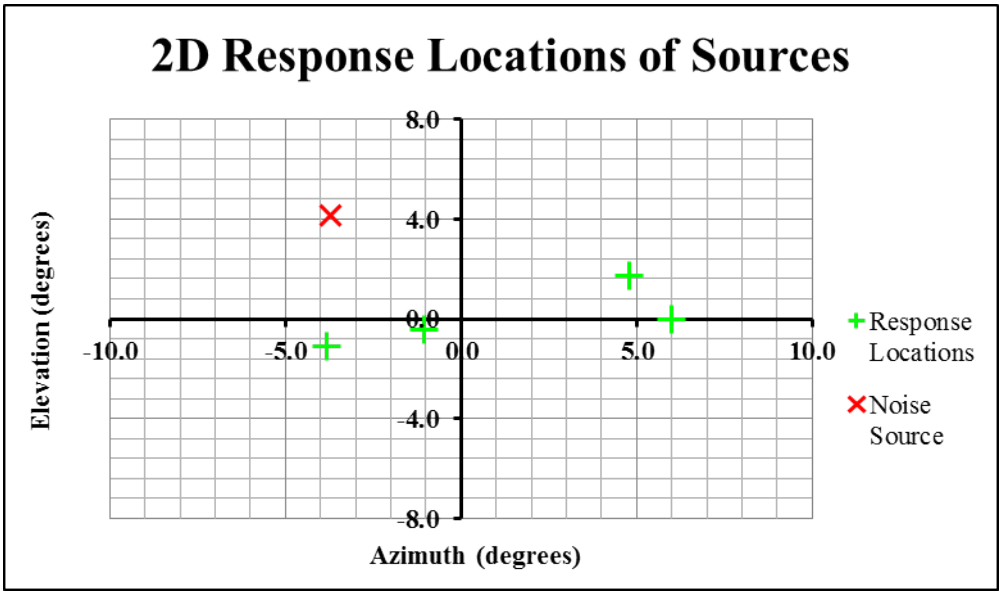


Figure 6.7: Spatial distribution of four response locations and one noise source.

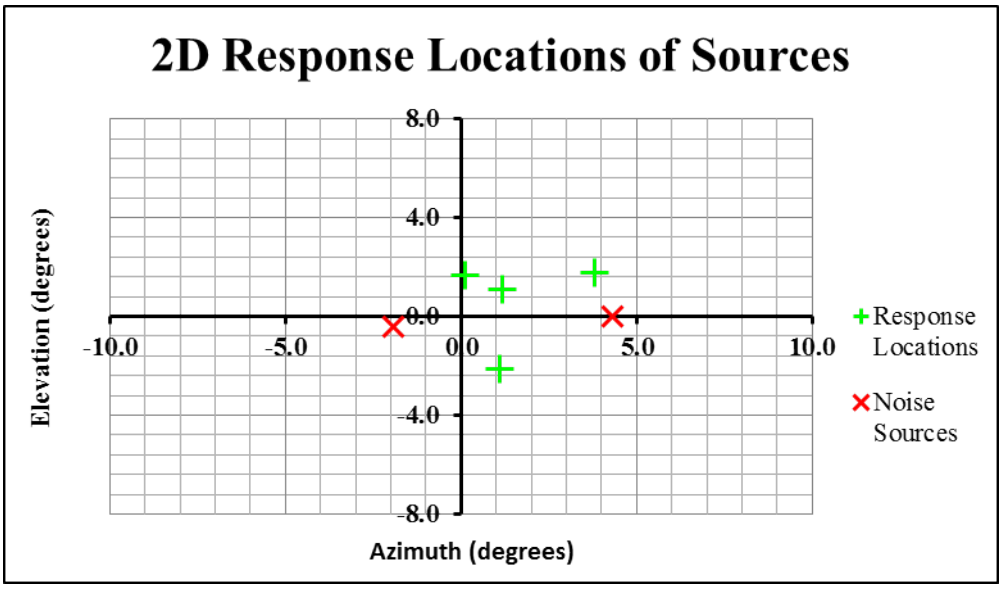


Figure 6.8: Spatial distribution of four response locations and two noise sources.

Spatial distributions of response locations in terms of both azimuth and elevation were calculated. These spatial distributions were determined as standard deviations, by determining how response locations (and noise source locations where applicable) were spread around the mean location. This was done for all response locations (2, 3, 4, 5, 6, 7) and also when noise sources were present. In figure 6.7 the azimuth and elevation standard deviations were 4° and 1.9°, respectively. In figure 6.8 these values were 2.3° and 1.4°, indicating greater clustering in this situation. The results will be provided for average azimuth and elevation position estimation errors separately using figures 6.9 to 6.14.

Average Azimuth Errors as a Function of Azimuthal Spatial Distribution of Response Locations

Figure 6.9 shows average azimuth errors for all response locations, with only the target sound source presented. The largest average azimuth error of $-0.6°$, was observed for six response locations. This error corresponded with the large azimuthal spatial distribution of 3.7° of these six response locations. Response locations with lesser azimuthal spatial distributions yielded lesser average azimuth errors.

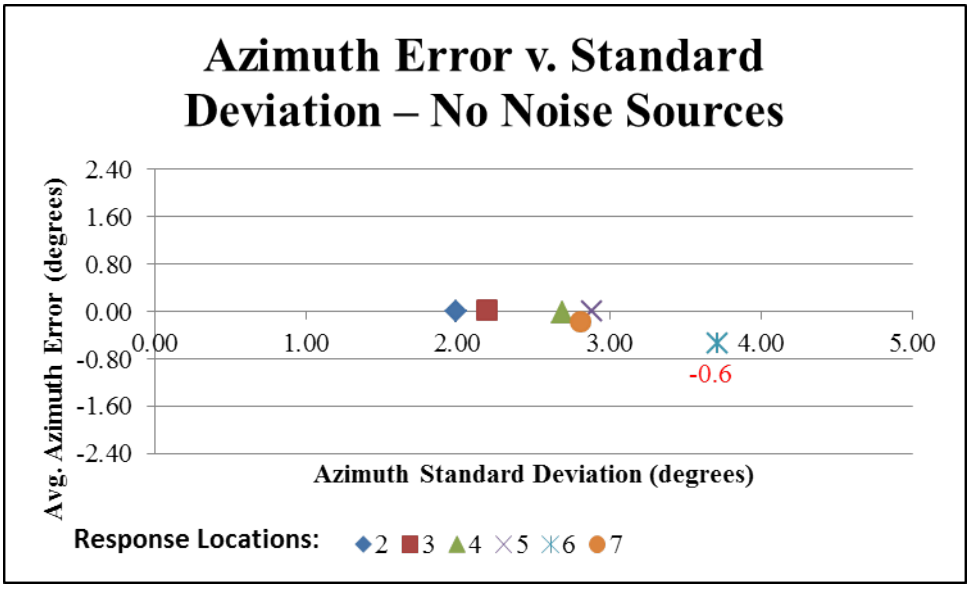


Figure 6.9: Position estimation as a function of azimuthal spatial distributions – No noise sources in this test. The largest average azimuth error ($-0.6°$) was observed with six response locations.

Figure 6.10 shows the influence of azimuthal spatial distribution for all response locations when the target sound source was presented with one noise sound source. The largest average error was 0.8°, which occurred for four response locations. The second largest azimuth error, 0.4°, occurred for six response locations. These two cases also gave rise to the largest azimuthal spatial distributions of 4.4° and 4°, for six and four response locations respectively.

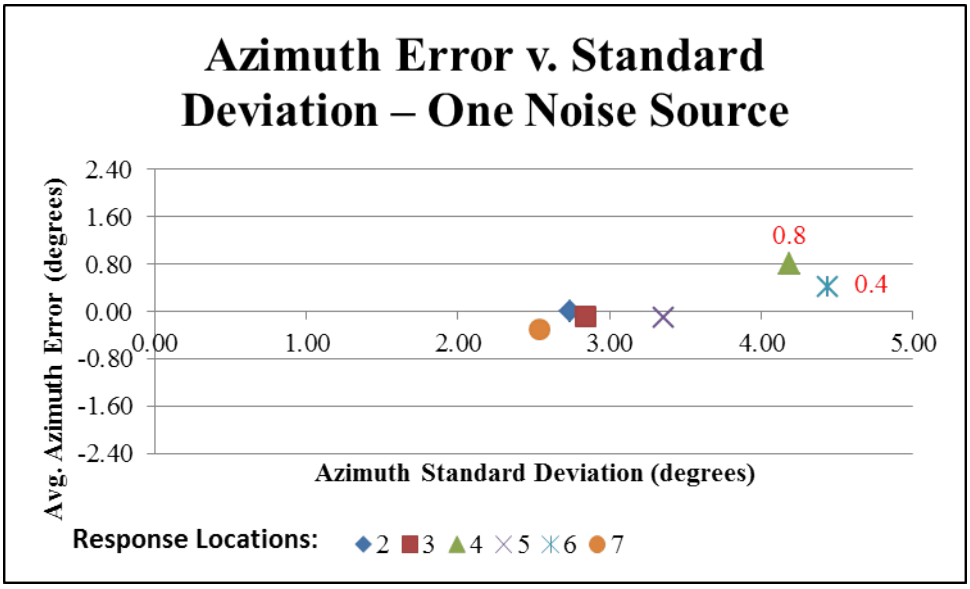


Figure 6.10: Position estimation as a function of azimuthal spatial distributions – One noise source in this test. The largest average azimuth error (0.8°) was observed with four response locations.

Figure 6.11 shows the azimuth errors when the target sound source was presented with two noise sources. In this case the largest average azimuth error of 2° occurred for six response locations. This corresponded with the largest azimuthal spatial distribution of 4.6°. The second largest average error of −0.8°, occurred for four response locations. This however, corresponded with the smallest azimuthal spatial distribution of −2°.

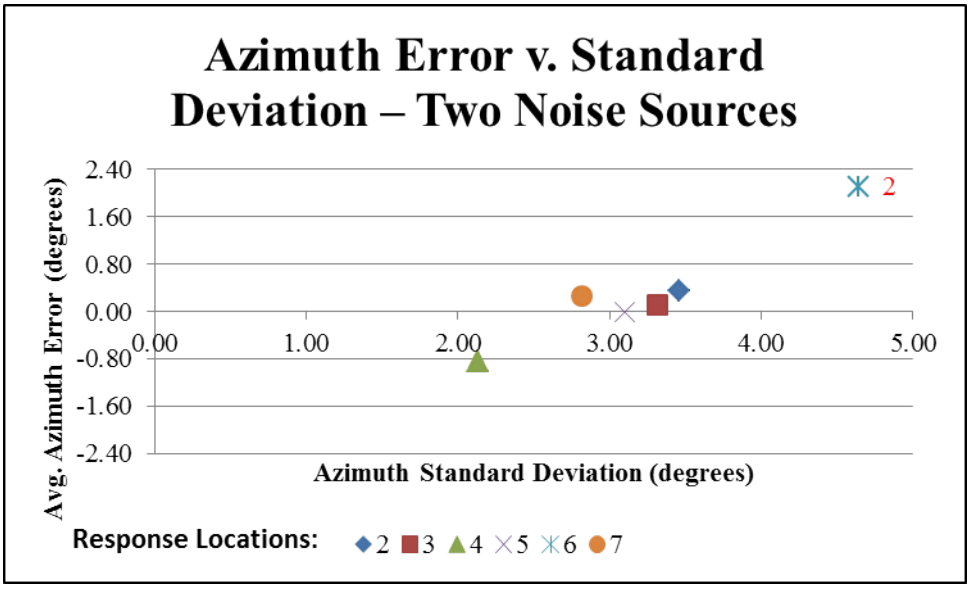


Figure 6.11: Position estimation as a function of azimuthal spatial distributions – Two noise sources in this test. The largest average azimuth error (2°) was observed with six response locations.

The level of clustering of objects that could be augmented with audio in the point cloud in terms of azimuth influences the position estimation accuracy. In the results quoted above this is more evident when the number of such objects increases. With six objects, having the widest spread, the position estimation accuracy was the worst.

Average Elevation Errors as a Function of Elevational Spatial Distribution of Response Locations

Figure 6.12 shows the largest average elevation error of 1° occurring for six response locations, when a target sound source was presented in the absence of noise sources. The elevational spatial distribution for six response locations in this instance was not the largest, however. The largest value of 2°, occurred for five response locations, in which case the average error was −0.4°, which was the second largest error.

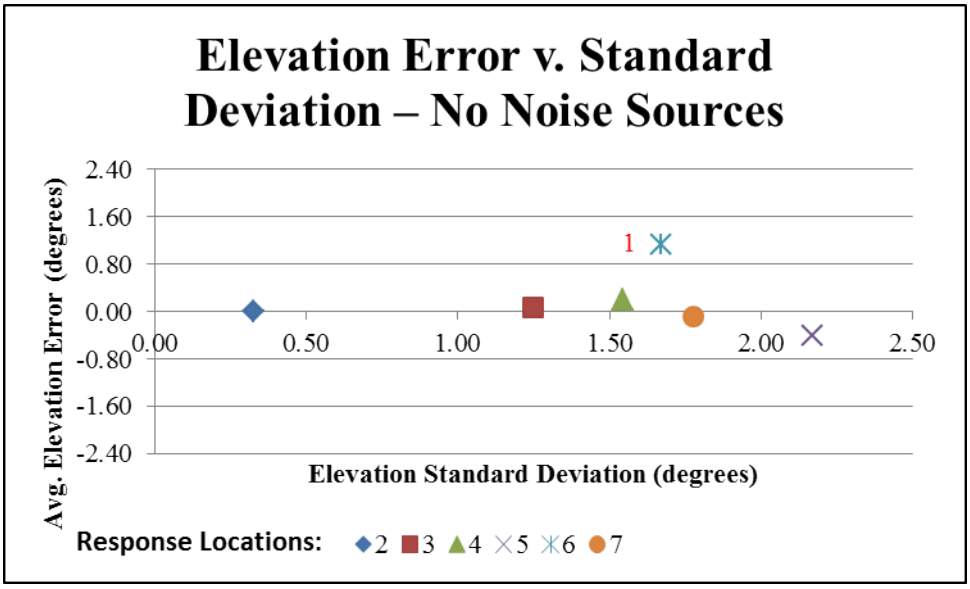


Figure 6.12: Position estimation as a function of elevational spatial distributions – No noise sources in this test. The largest average elevation error (1°) was observed with six response locations.

In the case where one noise source was presented with the target sound source, the largest average elevation error observed was −0.6° and occurred for seven response locations. This corresponded with the largest elevational spatial distribution of 2° (see figure 6.13). In contrast, the second largest average elevation error of −0.5° corresponded with the second smallest elevational spatial distribution of 0.7° – this was for five response locations.

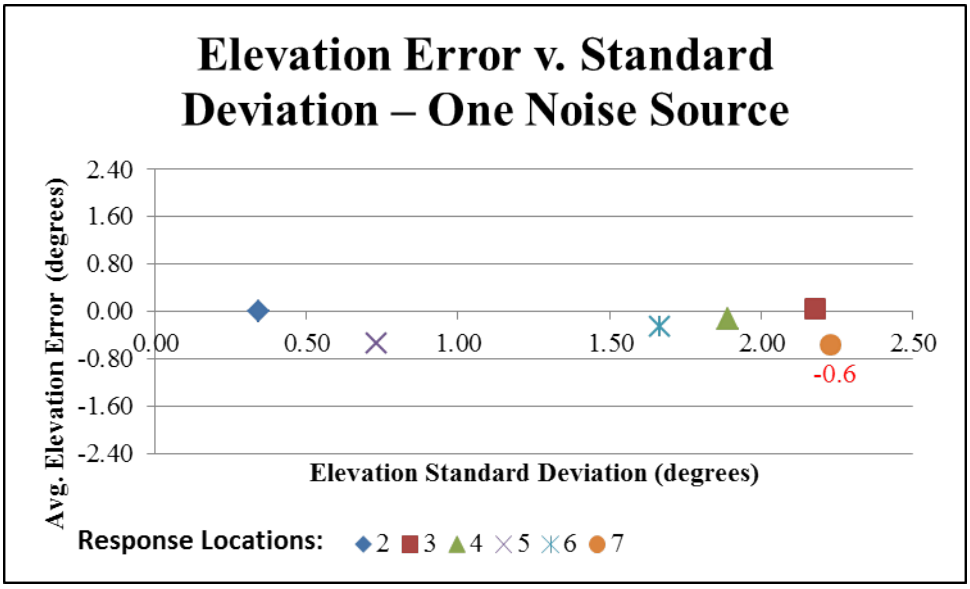


Figure 6.13: Position estimation as a function of elevational spatial distributions – One noise source in this test. The largest average elevation error ($-0.6°$) was observed with seven response locations.

Figure 6.14 shows results in the case where two noise sources were presented with the target sound source. Again the largest average elevation error occurred for six response locations and it was $-2°$. This corresponded with the largest elevational spatial distribution which was $2°$. The second largest average elevation error of $-0.7°$, occurred for four response locations, where the elevational spatial distribution was $1.4°$, which was the third smallest.

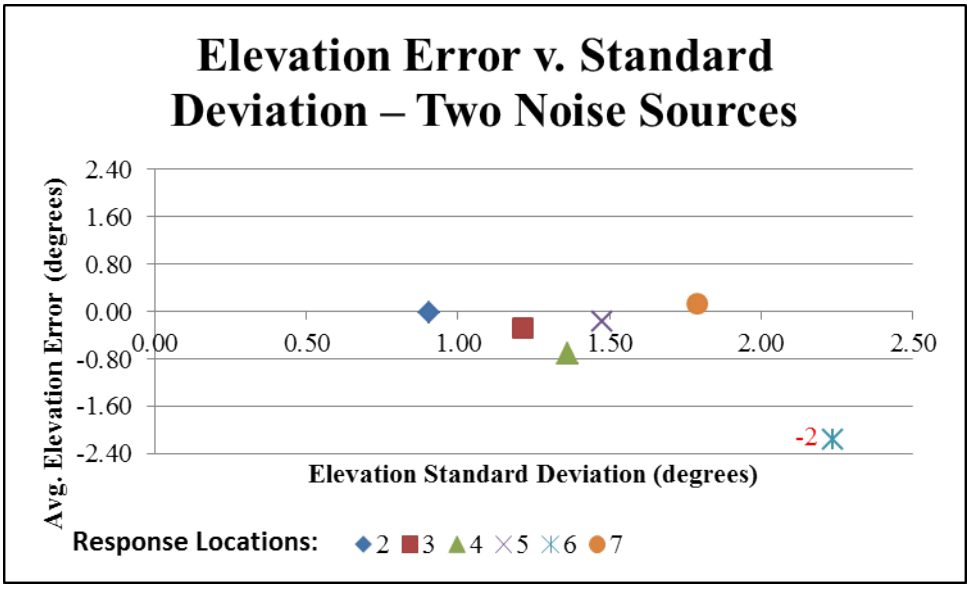


Figure 6.14: Position estimation as a function of elevational spatial distributions – two noise sources in this test. The largest average elevation error ($-2°$) was observed with six response locations.

The clustering in terms of elevation had an impact on position estimation accuracy here as well. This also corresponds with the number of objects potentially augmented with audio. The largest elevation errors were between six and seven potentially audio augmented objects. The largest error ($-2°$) was observed when the spatial distribution was the largest ($2°$).

6.1.4 Analyses of Position Estimation of the Target Sound Source Results

The results above indicate that the position estimation errors in both azimuth and elevation are related to the number of response locations presented to the tester. This relationship, however, is not a directly proportional one as was seen in the results. While processing a point cloud using audio augmented objects, the user needs to be cognisant of this. This means that if possible, the user work on a single problem at a time, in which case only one object would be augmented with audio. This will potentially improve the accuracy of estimating position such and object and improve the processing experience.

The introduction of noise sources had an impact on position estimation accuracy for both azimuth and elevation. The maximum absolute average

azimuth and elevation errors as a function of the number of noise sources are provided in figure 6.15. These results imply that while the user is focusing on one audio augmented object, if others exist then their audio should be detached or they should not emit audio. The existence of other audio augmented objects could act as noise and affect the position estimation and therefore the processing experience. This should be avoided if possible. The findings are in agreement with what Begault (1994) noted when observing that background noise is problematic.

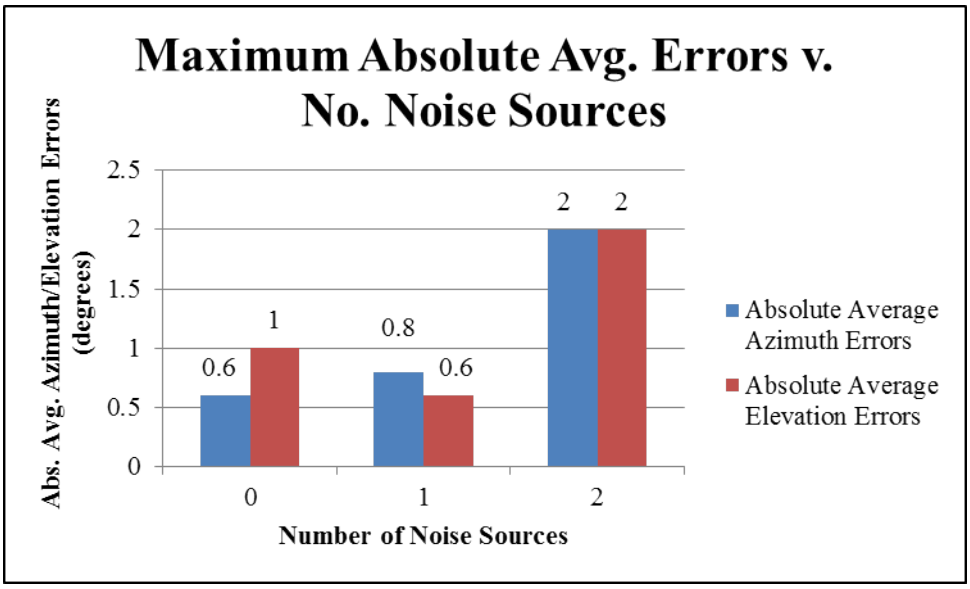


Figure 6.15: Maximum absolute errors as a function of the number of noise sources. The largest azimuth and elevation errors (both 2°) occurred when two noise sources were emitted.

The spatial distribution of response locations appears to have the most impact on position estimation accuracy. This can be viewed in two ways: 1) with tightly clustered response locations there was more confusion as to where the target sound source was located, and 2) with spread out response locations an error in one instance/trial could significantly affect the average error. In almost all cases, the maximum absolute average errors in both azimuth and elevation occurred when the spatial distributions were the largest. In situations where it is not avoidable to have numerous objects that could be augmented with audio, the user must expect that the spatial distribution of these objects will affect the position estimation accuracy. This will in turn affect the position estimation of objects of interest in the point cloud if there

are numerous objects that could be augmented with audio.

Although this is hard to quantify, it appears that of the binaural cues, the inter-aural intensity difference (IID) was more prevalent as a position estimation cue than the inter-aural time difference (ITD). The reason for this is that the target sound source's intensity changed, depending on where it was relative to the virtual listener. Despite some of these changes being discreet, the tester could still detect them to a certain extent. It is unclear whether time of arrival of the emitted sound to the virtual listener changed based on where the target sound source was. Despite this, the ITD could have still played a role in the position estimation process – this is still hard to quantify as ITDs are almost always less than 1 ms.

Because of the screen where response locations were displayed, the azimuth range of response locations was $-10°$ to $+10°$ and the elevation range was $-8°$ to $+8°$. This meant that response locations were distributed closer to the virtual listener's median plane. As was observed in the literature, both IID and ITD increase as the sound source approaches azimuths of $\pm 90°$, i.e., nearer to the inter-aural axis. Along the median plane, small IID and ITD are experienced – this played a role in position estimation of the target sound source here, as small IIDs and ITDs were discreet in some instances.

As the tester was immersed in the auditory environment using headphones, the role of the Head-related Transfer Function in position estimation was limited. As previously noted, generalised HRTFs are used for headphones. Here too it is difficult to quantify how much impact this had on position estimation accuracies. Non-externalisation of the target sound source and the inability to do head movements to help in the position estimation also likely impacted the outcomes from these tests.

The maximum average azimuth and elevation errors from these tests are still smaller than those found by Brungart *et al.* (1999) where they were $12.6°$ and $11.3°$ for azimuth and elevation, respectively. The reasons could be alluded to the fact that different test conditions were used here. Brungart *et al.*'s (1999) tests involved a number of test subjects, whereas only one was used here.

6.2 Depth of Target Sound Source

Depth of the target sound source with respect to the virtual listener was estimated at different response locations. At each response location, the target sound source was free to move to any of the three choices of depth. At each depth the target sound source played with a pitch/frequency associated with that depth. Given that sound intensity changes with depth, both intensity and frequency were used as depth signatures. This investigated if depths of audio augmented objects in point clouds could be estimated using intensity and pitch as depth cues and aid the processing. The results from depth tests will be shown here. The results will then be analysed and contextualised.

6.2.1 Depth Estimation Results

The results shown here are for both sets of tests, Set A and Set B. Set A depth choices were -0.75, -1.5 and -2.25 m while Set B depth choices were -1.0, -1.75 and -2.5 m. Only Test A1 (A1 meaning test 1 of Set A) for Set A results are shown, while for Set B it is only Test B1 results that are shown. The reason for omitting tests A2, A3, B2, and B3 is that the same results were obtained in them as those obtained in A1 and B1 respectively.

Figure 6.16 shows Test A1 results. The tester made correct depth estimations of the target sound source for all response locations. The random change of depth of the target sound source with respect to the virtual listener did not affect the estimations. As the target sound source 'moved' to the currently displayed response location, the euclidean distance between the source and the virtual listener also changed, in turn changing the sound source's intensity. However, the use of different frequencies as indicators of depth aided with the depth estimations.

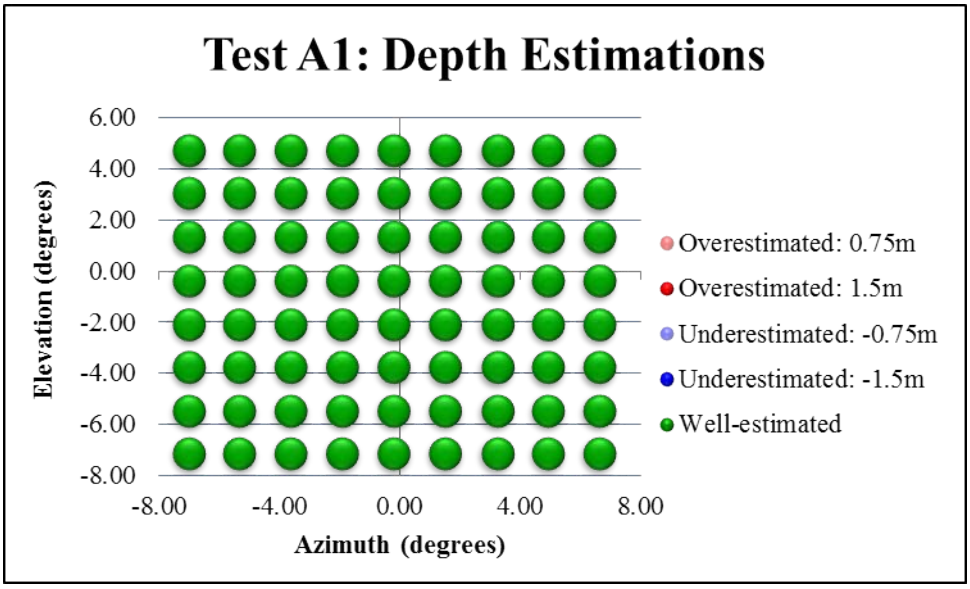


Figure 6.16: Depth estimation results from Test A1. Correct estimations made in all instances.

Correct depth estimation for all response locations were made in Test B1 as well. The results of these estimations are shown in figure 6.17. Here too the random change of depth did not affect the tester's ability to make correct estimations. Even though the intensity was lower for Set B tests as a result of greater depth values, the different frequencies still served as good depth signatures.

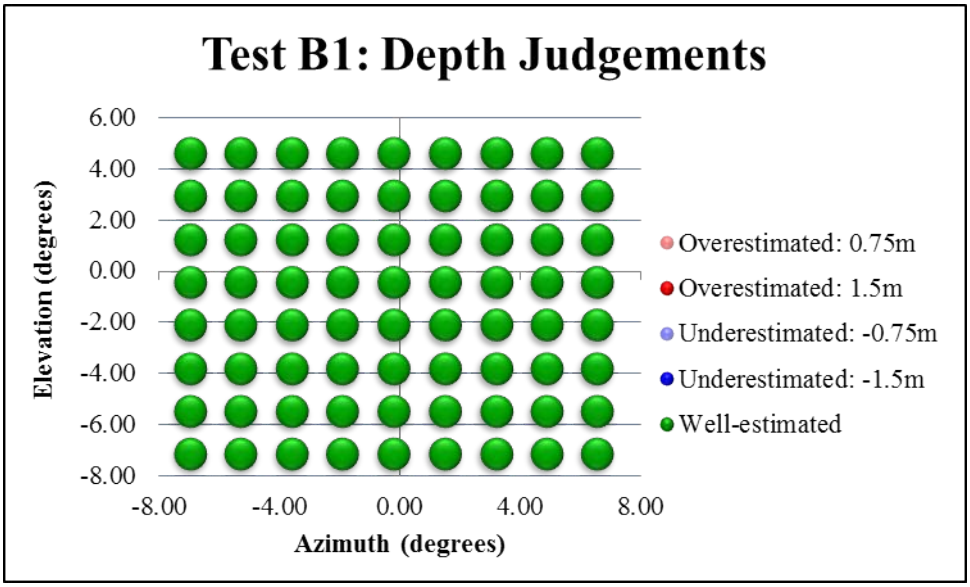


Figure 6.17: Depth estimation results from Test B1. Correct estimations made in all instances.

6.2.2 Depth Estimation Analyses

Depth estimations were made with the aid of intensity and frequency as depth cues. The increase in intensity with a decrease in depth and vice versa, occurs naturally for sound sources. The change of intensity was small, leading to crude depth estimations and as a result, frequency changes were introduced to help in making these depth estimations. The given results suggest that this was a beneficial undertaking.

While doing the depth estimation tests, it emerged that the tester needed to concentrate on the sine tone being emitted by the target audio object without distractions. With changes in the euclidean depth between the virtual listener and the target sound source, intensity changes became harder to detect. To some extent, this was true with regards to frequency changes too, hence the need to concentrate.

In particular, depth estimations for Set B depths were the hardest to make. The reason for this is that Set B depths were slightly greater than those of Set A, {-0.75,-1.5,-2.25} v. {-1.0,-1.75,-2.5}. As a result, the intensities at Set B depths were lower, leading to discreet changes when the target sound source changed depths. The frequency variations became discreet too, particularly when the target sound source was at the peripheral response locations, where

the euclidean depths were greatest.

The tester required some time to make the depth estimations. Response locations appeared on the screen at a rate of 4 s, meaning that at each response location the target sound source played the sine tone for 4 s, giving the tester 4 s to make the estimation. In some cases, particularly for Set A tests, the tester could make estimations in about 2 s. It took the tester longer to make the estimations in some instances, especially for Set B tests. This is alluded to the fact that the variations were more discreet for these tests.

Although a 100% depth estimation accuracy was obtained for tests of both sets, i.e., no mis-estimations occurred, this does not imply that depth estimations can be made error free. The reviewed literature suggests, making depth estimations is a crude exercise. As a result, the focus here was on the depths that would offer less discreet variations for the sake of detecting if the depth has changed or not.

The stated results suggest that depths of audio augmented objects in point clouds can be estimated accurately using intensity and pitch as depth cues. The changes in these depths can be detected using these cues. Discreet depth changes could be harder to detect, however. Therefore, while performing point cloud processing which requires depth estimation, the user must be cognisant of this.

Using more depth variations for each set of depths could have led to results being more like as illustrated in figure 5.8. The results could have likely behaved as predicted by Zahorik *et al.* (2005) where underestimation increases with depth. As pointed out, depth estimations were made by pressing a number key on the keyboard. Having more depth variations was problematic because there were more keys to choose from, leading to errors resulting from pressing the wrong key even if the tester knew what the correct depth was.

Calibration was needed before the tests could be carried out. This was so that the tester could become familiar with the sound source since familiarity aids in depth estimations. This entailed the tester having to programmatically change the depth of the target sound source to get an aural impression at different depths. The calibration was done so that the tester knew what intensities and frequencies to expect at what depths. This calibration process proved vital in the tests as the tester had a point of reference.

6.3 Occupancy of Space

Here the investigation is on the existence of unseen events in point cloud processing using reflections of audio augmented in a point cloud. In this investigation, the test subject was required to identify the type of reverberant environment a sound source was in, with four choices to choose from, *cave*, *hallway*, *mountain*, and *room* environments.

The tests were carried out as explained in section 5.3 and the results are provided in figure 6.18. The ambience judgement accuracies are provided in terms of the percentages, showing how successful the tester was in judging which ambient environment the sound source was emitted from.

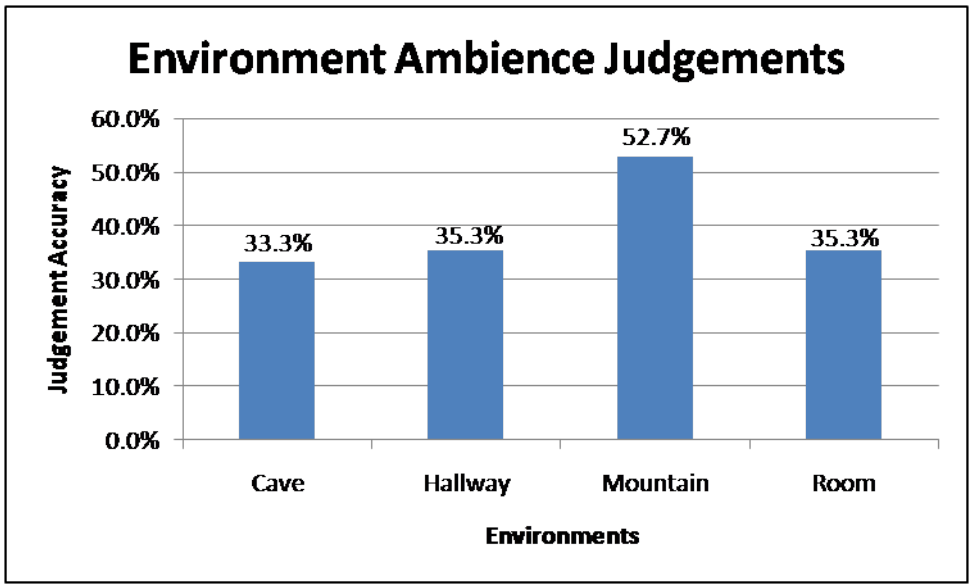


Figure 6.18: Environment ambience judgements. Highest ambience perception accuracy (52.7%) for the 'mountain' environment.

As shown in the figure, the tester could identify if the sound source was being emitted from a mountainous environment 52.7% of the time. For cave, hallway and room environments, the accuracies were 33.3%, 35.3% and 35.3%, respectively. The reverberation that the sound source experienced in mountainous environments was different to how it was experienced in the other environments. As a result, the reflections experienced by the virtual listener in mountainous environments were different to those of the other environments.

Mountainous environments are open, whereas cave, hallway and room environments are closed. This is the reason why the reflections of mountainous environments are so different and easier to distinguish from the others. The reflections experienced in closed environments are very similar, hence the difficulty in separating one from the others. For example, if the sound source was being emitted from a room environment, the tester would more likely judge it to be in a hallway or cave environment rather than a mountainous one. Furthermore it was difficult to distinguish if the sound source was emitted from a room or a hallway environment as the two make sound reflect in a very similar way.

The ability of the tester to identify the ambience of an environment can provide information about the occupancy of space of a point cloud. This implies that judging by the nature of the reflections, the user might be able to infer the unseen events in audio augmented point cloud processing.

To improve the accuracy of identifying the ambience of an environment, the user needs to listen to sound being reflected in different environments and get an understanding of how it is reflected. In doing the tests, the test subject had to go through a training process in order to understand these different reflections. The ability to determine how sound is reflected in a particular situation, can potentially inform the user the nature of the unseen event and its surroundings while processing a point cloud.

Chapter 7

Conclusions and Recommendations

Provided here are concluding remarks which will be based on each research question. Lastly, recommendations are made.

7.1 Conclusions

The auditory system can provide spatial information about objects where the visual cues are limited. Various auditory cues are used by the auditory system to do this, as already discussed. Using these cues, estimates of the position, depth and occupancy of space of an object of interest in point cloud processing by using audio objects associated to this object were sought for. This was so that point cloud processing can be enhanced by augmenting point clouds with audio.

7.1.1 Position Estimation

Positions of audio augmented objects in point clouds can be estimated with a fair level of accuracy. The available cues can be used to perform this position estimation, even though the estimation is limited by the use of headphones. In particular, the binaural cues play a significant role in the position estimation, moreover the inter-aural intensity difference cue. This can serve to assist in point cloud processing.

In the situation where there were no noise sources, absolute azimuth and

elevation accuracies were 0.6° and 1°, respectively. Considering the distance of 1 m between the screen and the tester, these translate to x and y screen values of about 0.01 and 0.02 m, respectively. The case of two noise sources (both azimuth and elevation of 2°) translates to x and y values of about 0.04 and 0.04 m, respectively.

The noise free augmentation of audio in point clouds promises to be beneficial. Judging by the stated results, it appears that in point cloud processing task, positions objects of interest in scans can be estimated fairly well. The accuracy would drop if other objects in the point cloud are augmented with audio and emitting it, however.

7.1.2 Depth Estimation

Depth estimation of sound sources using intensity and frequency variations as cues can prove useful in roughly detecting the depth of points/objects of interest in point clouds. This can be useful in point cloud processing tasks. For example a user can get a rough sense of how far out outliers are in a noisy point cloud and do the necessary cleaning.

Depths to/of unseen areas of interest in a point cloud can be detected, even though this will be crude. The aspect of familiarisation with a particular tone and its intensity at different levels could prove to be vital in auditory interfaces. The value of being familiar with a sound stimuli was demonstrated by Begault (1994). The user of the auditory interface will therefore need to be familiar to sound stimuli used to help in point clouds processing.

The inability to detect fine depth variations using audio cues could be limiting in audio augmented point cloud processing. In this respect, point cloud processing that depends on audio depth could require some time for the user to master. Because correct depth estimations at all response locations for all depth variations were made, processing instances where fine depth variations are not particularly required (e.g., coarse registration), can benefit from the use of intensity and frequency as depth cues.

7.1.3 Occupancy of Space

In the open environment that was tested, mountainous environment, the accuracy of estimating if the virtual user was in this environment was 52.7%.

In closed environments, cave, hallway and room environments, the accuracies were 33.3%, 35.3% and 35.3%, respectively. In the initial stages of the tests these were lower. With thorough knowledge of sound reflections in different environments, with practice, the user can be able to determine the nature of reflections more accurately. One can also get a sense of how clustered an environment is as that affects sound reflections depending on where the sound source is placed.

Using the nature of these reflections, the user can possibly infer the occupancy of space where visual cues are limited. This could be done when the user knows how the sound is reflected for an area of interest with a particular type of occupancy. As seen, some events might lead to similar reflections and therefore confusing the user. Familiarity with sound reflections could lead to better results and therefore enhance point cloud processing using augmented audio.

The objective of extracting spatial information of features of interest in point clouds has been realised. The cues used in extracting specific information have been identified. The accuracies of these cues and their limitations have also been stated. In point cloud processing tasks the relevant audio cues for extracting spatial information that is of interest to the user can be used to assist in processing the point cloud. In this audio augmented processing, the user needs to take note of the stated achievable accuracies and the stated limitations. Section 7.2 will recommend aspects that can possibly help improve point cloud processing using sound or help advance future work in this regard.

7.2 Recommendations

Unexpected limitations were encountered. These limitations likely affected the results as obtained from carried out tests. Effort was made to lessen their effects where possible.

Computer Screen Size

The manner in which position estimations were tested was limiting in a sense that the used 17 inch computer screen could only accommodate azimuth range of -10.0° to $+10.0^\circ$ and elevation range of -8.0° to $+8.0^\circ$. This meant that areas where binaural cues are most efficient could not be exam-

ined. Such areas are in the regions of roughly $\pm 30^{\circ}$ to $\pm 90^{\circ}$. Meaning that audio augmented point cloud processing could not be investigated at these regions.

The effect of the screen size on position estimation was not factored in. The influence of the screen size on position estimation needs to be investigated. This should look at the limitations imposed on audio augmented point cloud processing where position estimations of points are required.

Head Tracking

As the listening was not done in the free field (using headphones), the listener could not use head movements to better make estimations, particularly position estimations. It would be worthwhile to investigate if head movements of the user can be made to influence those of the virtual listener.

The investigation into head movements would possibly reveal if better position estimations can be made. This would have benefits in audio augmented point cloud processing because of better position estimations.

Audio Augmentation Per Point Cloud Processing Problem

Audio augmented point cloud processing needs to be tested on a case by case basis. In this research, examples of how audio would be augmented in various point cloud processing problems were given. Investigations of practically carrying out these augmentations for each problem would be worthwhile.

Every point cloud processing problem is unique. This would therefore lead to unique augmentation and unique limitations associated with the problem.

Estimating Other Spatial Information

In this study, the focus has been on position estimation, depth estimation and occupancy of space estimation. Audio augmentation is not limited to these. Investigations into retrieving other information would be worthwhile.

Retrieving information such as sizes and shapes of objects using audio cues could assist in point cloud processing. Here too, the limitations of this in point cloud processing would be studied.

Bibliography

Akeroyd, M. A. (2006), 'The psychoacoustics of binaural hearing', *International journal of audiology* **45**(S1), 25–33.

Alexa, M., Behr, J., Cohen-Or, D., Fleishman, S., Levin, D. and Silva, C. T. (2003), 'Computing and rendering point set surfaces', *Visualization and Computer Graphics, IEEE Transactions on* **9**(1), 3–15.

Audacity-Team (2012), 'Audacity'.
URL: *http://www.audacity.sourceforge.net/*

Begault, D. (1994), *3D Sound for Virtual Reality and Multimedia*, Academic Press Inc.

Begault, D. R. (1998), Auditory factors and non-auditory factors that potentially influence virtual acoustic imagery, *in* 'AES 16^{th} International conference on Spatial Sound Reproduction'.

Bosse, M., Zlot, R. and Flick, P. (2012), '*Zebedee*: Design of a spring-mounted 3-d range sensor with application to mobile mapping', *Robotics, IEEE Transactions on* **28**(5), 1104–1119.

Brenner, C., Dold, C. and Ripperda, N. (2007), 'Coarse orientation of terrestrial laser scans in urban environments', *ISPRS Journal of Photogrammetry and Remote Sensing* **63**(1), 4–18.

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

Bronkhorst, A. W. and Houtgast, T. (1999), 'Auditory distance perception in rooms', *Nature* **397**(6719), 517–520.

Brungart, D. S., Durlach, N. I. and Rabinowitz, W. M. (1999), 'Auditory localization of nearby sources. ii. localization of a broadband source', *The Journal of the Acoustical Society of America* **106**(4), 1956–1968.

Brungart, D. S. and Rabinowitz, W. M. (1999), 'Auditory localization of nearby sources. head-related transfer functions', *The Journal of the Acoustical Society of America* **106**(3), 1465–1479.

Bucksch, A. and Lindenbergh, R. (2008), 'Campino a skeletonization method for point cloud processing', *ISPRS journal of photogrammetry and remote sensing* **63**(1), 115–127.

Carlile, S., Leong, P. and Hyams, S. (1997), 'The nature and distribution of errors in sound localization by human listeners', *Hearing research* **114**(1), 179–196.

Creative-Labs (2010), 'Openal'.
URL: *connect.creativelabs.com/openal/default.aspx*

Edwards, A. D. (1989), 'Soundtrack: An auditory interface for blind users', *Human-Computer Interaction* **4**(1), 45–66.

Fabio, R. (2003), 'From point cloud to surface: the modeling and visualization problem', *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* **34**(5), W10.

Garland, M. and Heckbert, P. S. (1997), Surface simplification using quadric error metrics, *in* 'Proceedings of the 24th annual conference on Computer graphics and interactive techniques', ACM Press/Addison-Wesley Publishing Co., pp. 209–216.

Girardeau-Montaut, D., Roux, M., Marc, R. and Thibault, G. (2005), 'Change detection on points cloud data acquired with a ground laser scanner', *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* **36**(part 3), W19.

Handel, S. (1989), *Listening*, Massachusetts Institute of Technology.

Held, C. (2012), 'Creating 3d models of cultural heritage sites with terrestrial laser scanning and 3d imaging'.

Hiebert, G. (2007), *OpenAL Programmer's Guide*.
URL: *http://connect.creativelabs.com/openal/Documentation/Forms/AllItems.aspx*

Kapralos, B., Jenkin, M. R. and Milios, E. (2003), 'Auditory perception and spatial (3d) auditory systems', *Department of Computer Science, York University, Tech. Rep. CS-2003-07* .

Kendall, G. S. (1995), 'A 3-d sound primer: directional hearing and stereo reproduction', *Computer music journal* pp. 23–46.

Kolluri, R., Shewchuk, J. R. and O'Brien, J. F. (2004), Spectral surface reconstruction from noisy point clouds, *in* 'Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing', ACM, pp. 11–21.

Lee, J., Post, R. and Ishii, H. (2011), Zeron: Mid-air tangible interaction enabled by computer controlled magnetic levitation, *in* 'Proceedings of the 24th annual ACM symposium on User interface software and technology', ACM, pp. 327–336.

Lewis, A. (2013), 'Wordweb online lookup'.
URL: *http://www.wordwebonline.com/search.pl?w=acoustics*

Linsen, L. (2001), Point cloud representation, Technical report, Universität Karlsruhe.

Logitech (2013), 'G35 7.1 surround sound gaming headset'.
URL: *http://gaming.logitech.com/en-za/home*

Lounsbury, B. and Butler, R. (1979), 'Estimation of distances of recorded sounds presented through headphones', *Scandinavian audiology* **8**(3), 145–149.

Mahmoudi, M. and Sapiro, G. (2009), 'Three-dimensional point cloud recognition via distributions of geometric distances', *Graphical Models* **71**(1), 22–31.

Makous, J. C. and Middlebrooks, J. C. (1990), 'Two-dimensional sound localization by human listeners', *The journal of the Acoustical Society of America* **87**(5), 2188–2200.

Mederos, B., Velho, L. and de Figueiredo, L. H. (2003), Robust smoothing of noisy point clouds, *in* 'Proc. SIAM Conference on Geometric Design and Computing', Vol. 2004, p. 13.

Mereu, S. W. and Kazman, R. (1996), 'Audio enhanced 3d interfaces for visually impaired users', *ACM SIGCAPH Computers and the Physically Handicapped* **1**(57), 72–78.

Mershon, D. H. and King, L. E. (1975), 'Intensity and reverberation as factors in the auditory perception of egocentric distance', *Perception & Psychophysics* **18**(6), 409–415.

Pauly, M., Gross, M. and Kobbelt, L. P. (2002), Efficient simplification of point-sampled surfaces, *in* 'Proceedings of the conference on Visualization'02', IEEE Computer Society, pp. 163–170.

Rabbani, T., Dijkman, S., van den Heuvel, F. and Vosselman, G. (2007), 'An integrated approach for modelling and global registration of point clouds', *ISPRS journal of Photogrammetry and Remote Sensing* **61**(6), 355–370.

Rayleigh, L. (1875), 'On our perception of the direction of a source of sound', *Proceedings of the Musical Association* pp. 75–84.

Razavi, B., O'Neill, W. and Paige, G. D. (2005), Both interaural and spectral cues impact sound localization in azimuth, *in* 'Neural Engineering, 2005. Conference Proceedings. 2nd International IEEE EMBS Conference on', IEEE, pp. 587–590.

Schall, O., Belyaev, A. and Seidel, H.-P. (2005), Robust filtering of noisy scattered point data, *in* 'Proceedings of the Second Eurographics/IEEE VGTC conference on Point-Based Graphics', Eurographics Association, pp. 71–77.

Shinn-Cunningham, B. G. (2000), Distance cues for virtual auditory space, *in* 'Proceedings of the First IEEE Pacific-Rim Conference on Multimedia', pp. 227–230.

Song, H. and Feng, H.-Y. (2008), 'A global clustering approach to point cloud simplification with a specified data reduction ratio', *Computer-Aided Design* **40**(3), 281–292.

Thurlow, W. R., Mangels, J. W. and Runge, P. S. (2005), 'Head movements during sound localization', *The Journal of the Acoustical society of America* **42**(2), 489–493.

Väljamäe, A. (2005), Self-motion and presence in the perceptual optimization of a multisensory virtual reality environment, Technical report, Chalmers University of Technology.

Vorländer, M. (2008), *Auralization*, Springer.

Vosselman, G., Gorte, B. G., Sithole, G. and Rabbani, T. (2004), 'Recognising structure in laser scanner point clouds', *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* **46**(8), 33–38.

Vosselman, G. and Maas, H. (2010), *Airborne and Terrestrial Laser Scanning*, Whittles Publishing.

Wenzel, E. M., Wightman, F. L. and Kistler, D. J. (1991), Localization with non-individualized virtual acoustic display cues, *in* 'Proceedings of the SIGCHI Conference on Human Factors in Computing Systems', ACM, pp. 351–359.

Weyrich, T., Pauly, M., Keiser, R., Heinzle, S., Scandella, S. and Gross, M. (2004), Post-processing of scanned 3d surface data, *in* 'Proceedings of the First Eurographics conference on Point-Based Graphics', Eurographics Association, pp. 85–94.

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

Woo, H., Kang, E., Wang, S. and Lee, K. H. (2002), 'A new segmentation method for point cloud data', *International Journal of Machine Tools and Manufacture* **42**, 167–178.

Wozniewski, M. and Settel, Z. (2007), User specific audio rendering and steerable sound for distributed virtual environments, *in* 'Proceedings of the 13th International Conference on Auditory Display'.

Xie, H., McDonnell, K. T. and Qin, H. (2004), Surface reconstruction of noisy and defective data sets, *in* 'Visualization, 2004. IEEE', IEEE, pp. 259–266.

Xie, Z., Xu, S. and Li, X. (2010), 'A high-accuracy method for fine registration of overlapping point clouds', *Image and Vision Computing* **28**(4), 563–570.

Zahorik, P., Brungart, D. S. and Bronkhorst, A. W. (2005), 'Auditory distance perception in humans: A summary of past and present research', *Acta Acustica united with Acustica* **91**(3), 409–420.