An Investigation of Eyes-Free Spatial Auditory Interfaces for Mobile Devices: Supporting Multitasking and Location-Based Information

Yolanda Vazquez-Alvarez

Submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

School of Computing Science

College of Science and Engineering, University of Glasgow

July 2013

Abstract

Auditory interfaces offer a solution to the problem of effective eyes-free mobile interactions. However, a problem with audio, as opposed to visual displays, is dealing with multiple simultaneous information streams. Spatial audio can be used to differentiate between different streams by locating them into separate spatial auditory streams. In this thesis, we consider which spatial audio designs might be the most effective for supporting multiple auditory streams and the impact such spatialisation might have on the users' cognitive load.

An investigation is carried out to explore the extent to which 3D audio can be effectively incorporated into mobile auditory interfaces to offer users eyes-free interaction for both multitasking and accessing location-based information. Following a successful calibration of the 3D audio controls on the mobile device of choice for this work (the Nokia N95 8GB), a systematic evaluation of 3D audio techniques is reported in the experimental chapters of this thesis which considered the effects of multitasking, multi-level displays, as well as differences between egocentric and exocentric designs.

One experiment investigates the implementation and evaluation of a number of different spatial (egocentric) and non-spatial audio techniques for supporting eyes-free mobile multitasking that included spatial minimisation. The efficiency and usability of these techniques was evaluated under varying cognitive load. This evaluation showed an important interaction between cognitive load and the method used to present multiple auditory streams. The spatial minimisation technique offered an effective means of presenting and interacting with multiple auditory streams simultaneously in a selective-attention task (low cognitive load) but it was not as effective in a divided-attention task (high cognitive load), in which the interaction benefited significantly from the interruption of one of the stream.

Two further experiments examine a location-based approach to supporting multiple information streams in a realistic eyes-free mobile environment. An initial case study was conducted in an *outdoor* mobile audio-augmented exploratory environment that allowed for the analysis and description of user behaviour in a purely exploratory environment. 3D audio was found to be an effective technique to disambiguate multiple sound sources in a mobile exploratory environment and to provide a more engaging and immersive experience as well as encouraging an exploratory behaviour. A second study extended the work of the previous case study by evaluating a number of complex multi-level spatial auditory displays that enabled interaction with multiple location-based information in an *indoor* mobile audio-augmented exploratory environment. It was found that a consistent exocentric design across levels failed to reduce workload or increase user satisfaction, so this design was widely rejected by users. However, the rest of spatial auditory displays tested in this study encouraged an exploratory behaviour similar to that described in the previous case study, here further characterised by increased user satisfaction and low perceived workload.

Table of Contents

List of Tables

List of Figures

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(Yolanda Vazquez-Alvarez)

Contributing Publications

Vazquez-Alvarez, Y. and S. Brewster (2009) Investigating background & foreground interactions using spatial audio cues. In CHI '09 Extended Abstracts on Human Factors in Computing Systems, pp. 3823-3828. Boston, MA, USA: ACM Press.

Vazquez-Alvarez, Y. and S. Brewster (2009) Audio minimization: Applying 3D audio techniques to multi-stream audio interfaces. In M. E. Altinsoy, U. Jekosch, and S. Brewster (eds.), HAID 2009, vol. LNCS 5763. Dresden, Germany: Springer.

Vazquez-Alvarez, Y. and S. A. Brewster (2010) Designing spatial audio interfaces to support multiple audio streams. In Proceedings of MobileHCI 2010, pp. 253-256. Lisbon, Portugal: ACM Press.

Vazquez-Alvarez, Y. (2010) Designing spatial audio interfaces for mobile devices: supporting multitasking and context information. In Proceedings of MobileHCI 2010 (Doctoral Consortium), pp. 481-482. Lisbon, Portugal: ACM Press.

Vazquez-Alvarez, Y. and S. A. Brewster (2011) Eyes-free multitasking: The effect of cognitive load on mobile spatial audio interfaces. In Proceedings of CHI 2011, pp. 2173-2176. Vancouver, Canada: ACM Press.

Vazquez-Alvarez, Y., I. Oakley, and S. A. Brewster (2012) Auditory display design for exploration in mobile audio-augmented reality. *Personal and Ubiquitous Computing*, **16**, 8, 987-999.

Acknowledgements

First of all, I would like to thank my supervisor, Prof. Stephen Brewster, for his support, invaluable advice and friendship. I would also like to take the opportunity to thank my second supervisor Prof. Roderick Murray-Smith.

I am grateful to my Ph.D. examiners, Prof. Matt Jones and Dr. Alessandro Vinciarelli for their valuable feedback that improved the quality of my thesis.

My colleagues and members of the Glasgow Multimodal Interaction Group, both past and present, have also been a source of support and inspiration. In particular, I want to thank Dr. David McGookin for the friendship he has offered over the years and for being there during some tough times. Additional thanks go to Chris McAdam, Dr. Georgios Marentakis, Dr. Marilyn McGee-Lennon, Dr. Graham Wilson, Wanda Diaz, Martin Halvey and David Warnock for their enthusiasm and support.

My warmest thanks are due to my co-authors and colleagues, Dr. Ian Oakley, Dari Trendafilov and Saija Lemmela for their collaboration and support at all stages. Additionally, I would like to thank all the people at the Madeira Interactive Technologies Institute (M-ITI), Funchal (Portugal) and at the Nokia Research Center (NRC-Helsinki), that made my stay both a productive and enjoyable one.

I would also like to acknowledge the administrative staff at the School of Computing Science and in particular Mrs May Gallagher, Mrs Helen McNee and Mr Stewart Macneill for their cheerfulness and support.

Last but not least, I would like to thank the people who have been by my side unconditionally throughout the duration of my Ph.D. My dearest thanks to my husband, Matthew P. Aylett, for his loving care, patience and excellence guidance throughout these years, and also to our daughter Isabella for bringing so much joy and happiness into my life. I would also wish to express my gratitude to my family, including my in-laws, for their encouragement and support. Finally, I am grateful to my many friends both here in Scotland and at home in Spain for making this a wonderful journey.

This research was jointly funded by Nokia and EPSRC research grant EP/F023405 as part of the Gaime Project.

Dedication

In loving memory of my father Segundo. *A la memoria de mi padre, por ser motivo de inspiración y superación constante.*

Chapter 1

Introduction

1.1 Motivation of the Thesis

Ownership of smart mobile devices, be it a mobile phone or a media player, is growing very rapidly. Mobile devices are everywhere and every year they become more technologically advanced and increasingly support more complex functionality. Users are increasingly making use of this functionality when mobile. For example, when walking on the street users might receive a text message or they might want to check an appointment or phone number on their calendar or contacts. To access this information, users have to retrieve the device from where it is kept and start interacting with the screen, which is usually small-sized, with a tiny keyboard. Moreover, if users are walking on the street, they may have to stop walking, as these tasks will require their visual attention.

Auditory interfaces can offer a solution for non-visual, eyes-free, mobile interaction when visual attention is compromised or the mobile device is out of reach, especially if users wear a pair of headphones, something that is becoming increasingly common as the consumption of streamed information grows when the user is mobile. For example, a calendar application could be accessed using an audio menu. The user would interact with a hierarchical audio menu via key presses in order to find the required information from the calendar application. Instead of the user interacting with the information visually, the information is provided to the user as synthesised speech, symbolic audio or recorded prompts. This allows users to use their vision to concentrate fully on navigating the streets and avoid dangers such as walking into obstacles like signposts and traffic. However, although auditory interfaces can offer a solution for eyes-free interactions, they also face a number of challenges:

1. Multitasking – multitasking requires either interrupting your main activity or to focus on another one simultaneously. For instance, users engaged in a phone conversation might need to access a phone number in the address book application. Using an auditory interface to support multitasking would require the use of multiple auditory streams, one for each task. The challenge is how to efficiently present and manage these multiple audio streams without overloading the user.
2. Location-based information – location-based systems, which capture the user's location and then present context-sensitive information, are becoming increasingly popular with the widespread availability of Global Positioning Systems (GPS) units on handheld devices. Location-based information is usually presented spatially in current visual interfaces. The challenge for an auditory interface supporting access to location-based information is how to map auditory streams to context-sensitive information based on location, and how to discriminate spatially between them.

Spatial audio can be used to address these challenges. However, up till now very little work has focused on these issues, especially on mobile platforms.

Recent interest in 3D audio techniques has resulted in the availability of 3D audio APIs on various platforms such as Vodafone (VFX Specification), NTT Docomo, JAVA JSR-234 Advanced Multimedia Supplements (AMMS), and Open SL ES. A set of Head Related Transfer Functions (HRTFs) (Blauert, 1997) is typically used by 3D audio controls in these APIs to allow an accurate localisation of sound in 3D space. If a monaural sound signal is passed through these filters and heard through headphones, the listener will hear a sound that seems to come from a particular location in space.

Many perceptual studies have examined how well listeners can extract the content of one sound source (the target) in the presence of competing sound sources (maskers), a situation requiring selective attention, as in Cherry's Cocktail Party effect work. The Cocktail Party effect (Cherry, 1953) provides evidence that humans can, in fact, monitor several auditory streams simultaneously, selectively focusing attention on any one and placing the rest in the background. Thus, 3D audio could offer the means for discriminating between auditory streams when multitasking is necessary.

3D audio can also mirror the spatial organisation of a visual display. Therefore allowing the creation of a spatial mapping. In this way physical metaphors can be represented using 3D audio. For example, a radial menu around a user's head representing the time around a clock (Walker and Brewster, 2000), or perceiving the origin of a sound source advertising a restaurant nearby as coming from the actual physical direction of the restaurant. Thus, 3D audio could

offer the means for representing context-sensitive information spatially.

Previous research has investigated the use of 3D audio techniques to either passively browse multiple auditory streams such as news (Schmandt and Mullins, 1995), or to reinforce the cognitive mapping between sequential audio items and their spatial location in for instance a radial menu around the user's head (Zhao *et al.*, 2007). However, it is still unclear how 3D audio techniques might be implemented in an interactive environment, where we need to consider how to manage multiple auditory streams without overloading the user. This thesis investigates the challenges faced by auditory interfaces when designed for eyes-free interactions in mobile environments and will ultimately aim to put forward possible solutions to the problem of supporting multitasking and location-based information in such interfaces.

1.2 Aims of the Thesis

Thesis Statement: *This thesis asserts that 3D audio can be effectively incorporated into mobile auditory interfaces to offer users eyes-free interaction for both multitasking and accessing location-based information. Spatial minimisation and spatialised multi-level auditory displays offer an effective means of presenting and interacting with multiple auditory streams simultaneously in an eyes-free mobile interface, although the design of such interfaces is affected by attention demands, localisation error and subject preference.*

This thesis investigates the problem of designing eyes-free auditory interfaces supporting multitasking and location-based information for mobile interactive environments. The work in this thesis focuses on how current visual interface designs could be employed in audio-only interfaces to alter focus on multiple streams and spatially organise location-based information. These issues are addressed by focusing on the following thesis research questions:

RQ 1 To what extent can 3D audio techniques aid the user to maintain coherent attention on multiple auditory streams in a mobile eyes-free interface?

RQ 2 How can 3D audio techniques be used to disambiguate multiple auditory sources in order to access location-based information in a mobile eyes-free interface?

To address RQ 1, a number of 3D audio techniques are designed and tested for their efficiency and usability under varying amounts of cognitive load when supporting multiple and simultaneous auditory streams in an eyes-free auditory interface. A 3D audio technique (it will be referred to as *spatial minimisation*) is presented that allows the user to foreground and background an auditory stream by moving it to the side while a second stream is played from the front, thus

being able to alter the focus between auditory streams. This technique aims to mimic how items, for instance a window, are minimised in visual interfaces in order to enhance the interface.

RQ 2 is addressed by two experimental studies each evaluating different eyes-free auditory interfaces in a non-guided mobile audio-augmented reality environment. A first case study investigates the user experience of discovery in an outdoors audio-augmented reality environment including multiple and simultaneous sound sources. A systematic user experience evaluation is carried out focusing on user performance both quantitatively and qualitatively over a number of different auditory displays. In addition, a second study further investigates user interaction with multi-level auditory displays designed to support multiple simultaneous sound sources in an indoor audio-augmentation system. This is done by quantifying the impact of spatial context, level of immersion and cognitive load on user preference and performance when interacting with a multi-level auditory display.

1.3 Contributions

This work is novel in that it systematically evaluates the use of 3D audio techniques in the design of auditory interfaces supporting multiple auditory streams for eyes-free mobile interactions. A consistent evaluation framework for the analysis and description of user behaviour is provided to test user-driven interactions, the results of which will provide guidelines for designers when building eyes-free auditory interfaces for mobile applications. Also, a new approach is explored by taking two visual user interface metaphors and testing them in an auditory display using two 3D audio techniques (i.e. *spatial minimisation* and *spatialised multi-level auditory displays*) to mirror the spatial organisation of a visual display. The design of these spatialised multi-level auditory displays included a novel combination of egocentric and exocentric techniques within the same auditory interface that enables the user to access and manage location-based information.

The work presented here is directly relevant to the design of future auditory interfaces. For example: *Imagine a typical morning for David as he commutes to work: David is travelling on the New York City subway system on his way to work and listening to music on his phone, as he always does. While David is in transit, his boss Stephen tries to call him but, as there is no underground cell phone coverage, he leaves a voice mail. As David leaves the station at his usual destination, he listens to the voicemail while his favourite music is still playing. He finds out he has to reschedule a meeting with his boss. He then decides to pause his music and starts interacting with his calendar using an audio menu as he keeps on walking. As he starts browsing through his appointments to find a free slot, Stephen calls him back. While taking to Stephen,*

David browses his calendar and finds a suitable time later on the day for their meeting. The call ends and David continues listening to his music while making his way to the office.

An auditory interface like the one illustrated in this example makes eyes-free interactions possible while on the go when visual attention is compromised or the mobile device is out of reach, and enables the user to interact with his mobile devices purely through sound. The extent to which such an audio-only interface is desirable depends on how we deal with cognitive load and multiple streams of information. In this example, there are differing levels of both. From a notification of a voice mail, which might be delivered at the same time David listens to music, to a multitasking extreme of listening to music during a phone call while rearranging a meeting and walking the streets of New York. All of this without getting run over by a yellow cab. How should an auditory interface be designed to deal with these competing requirements? Chapter 5 investigates basic strategies for the presentation of multiple audio streams and their usability under varying cognitive load.

Consider the following scenario: *There is a new conceptual art exhibition at the Tate Modern in London and art lover David has arranged to visit the exhibition with his friend Rocio, who is not very much into art but agrees to go along. Before they enter the gallery, they download an audio application onto their mobile phone that will enable them to listen to information about the art pieces using their headphones while walking around the exhibition. As they get close to an audio-augmented location, different sounds allow users to browse the audio information available. This varies between comments left by visitors, the artist herself and a well known art critic who visited the exhibition the previous day. As they explore the exhibition, David is very interested in the comments left by the artist and critic and how the pieces formally implement the conceptual ideas they are based on. Rocio, in contrast, listens to a random selection of comments from other visitors and is amused to find that many of them, although enjoying the art, find the pieces hard to interpret. At one artifact, a multi-picture frame displaying vibrant woolen threads, Rocio browses the options using an audio menu and selects a comment left by a previous visitor that says the piece reminds him of veins and circulatory systems. This comment makes Rocio laugh. David selects a comment left by the artist, which describes how the frame squeeze wool of different colours to contrast the 2D nature of the photo frame with the 3D element of materials. David and Rocio have a lively discussion based on these comments. They agree that comments provided by the artist helped them appreciate the ideas in the work, while the opinions left by other visitors mentioned things they would never thought of themselves. Overall, the result is a personalised museum experience, which has responded to the individual user interests and encouraged them to appreciate and enjoy the art work in more depth in their own way. David and Rocio leave the exhibition happy and stimulated by the work they have*

seen.

As illustrated in this example, location-based information can be presented by augmenting a real environment with audio. When using such an eyes-free auditory interface, each location being augmented requires the use of an audio stream which means it may be necessary to discriminate between them, especially when they could overlap if locations are close to each other. Thus, when designing auditory displays for mobile audio-augmented environments, choices have to be made on both the audio presentation and the spatial arrangement of the audio streams.

Should information be provided sequentially or simultaneously? While simultaneous presentation is important to create a rich immersive audio environment, high levels of cognitive load may affect exploration and selection between different locations and also the exploration and selection of the various amounts of information provided at each location.

How do we structure concentrated areas of information in a location-based system? One approach is to use a multi-level auditory display. This design makes it possible to arrange and group information, but there are different advantages and disadvantages in using different auditory display techniques. While a homogeneous design across auditory displays follows the design principal of consistency, different information may require different types of auditory display. In Chapter 7 these practical questions are examined.

1.4 Thesis Structure

Chapter 2, *Human Spatial Audio Localisation of Real and Virtual Sound Sources*, reviews the literature on spatial audio perception in real and virtual environments. This chapter begins with a brief account of the characteristics of sound, followed by an overview of how the human auditory system is able to perceive and localise real sound sources, and finishing with a review of current audio techniques that are able to model human sound localisation in order to create virtual audio environments.

Chapter 3, *Review of Existing Research in Spatial Auditory Displays*, presents a review of auditory display design and spatial audio applications. In addition, the requirements for supporting multiple auditory streams in a spatial auditory interface are identified.

Chapter 4, *Experimental Groundwork: Evaluation of the Spatial Localisation Accuracy on the Nokia N95*, reports an evaluation that investigates the accuracy of the positional 3D audio controls available on the mobile device of choice to carry out the experimental work presented

in this thesis. This chapter is based on work published in CHI 2009 (Vazquez-Alvarez and Brewster, 2009b).

Chapter 5, *Designing Spatial Auditory Interfaces for Eyes-free Multitasking*, presents an experimental study comparing the efficiency and usability of spatial and non-spatial auditory interfaces in an interactive multitasking environment under varying cognitive load. An audio minimisation technique implemented using 3D audio is introduced and evaluated. The chapter concludes with a discussion of the experimental results, which indicates that 3D audio techniques can offer an effective means of presenting and interacting with multiple auditory streams simultaneously but highlight the need to control for cognitive load. This chapter is based on work published in HAID 2009 (Vazquez-Alvarez and Brewster, 2009a), MobileHCI 2010 (Vazquez-Alvarez and Brewster, 2010) and CHI 2011 (Vazquez-Alvarez and Brewster, 2011).

Chapter 6, *Case Study: Location-based Information in a Mobile Audio-Augmented Reality Environment*, presents an initial case study investigating user performance and interaction strategies with location-based information in a purely exploratory *outdoor* mobile audio-augmented environment that included multiple simultaneous sound sources. Quantitative and qualitative data collected in this study are presented, which provide an analysis of user exploration and interaction strategies. This chapter is based on work published in the *Personal and Ubiquitous Computing* Journal (Vazquez-Alvarez *et al.*, 2012).

Chapter 7, *Supporting Multi-Level Auditory Displays in Mobile Audio-Augmented Reality*, reports an in-depth experimental study that builds on the findings described in Chapter 6 and further explores the potential of different spatial audio configurations in an *indoor* mobile audio-augmented exploratory environment. A multi-level auditory display design is proposed to arrange and group multiple auditory streams of information. Quantitative and qualitative data collected while users were interacting with these multi-level auditory displays are presented that provide an analysis of user experience and performance.

Chapter 8, *Conclusions*, summarises the work presented in this thesis and relates the findings back to the research questions identified in Chapter 1. Also, based on this findings, a set of guidelines are identified. The limitations of this research are discussed and future work based on the issues raised by the thesis is suggested.

1.4.1 Relationship between the Experimental Studies

In order to address the aims of this thesis (as outlined in Section 1.2) a number of experimental studies were required that varied in the extent they were 'lab-only' (carried out in a controlled laboratory environment) or 'in the wild' (Outside a lab environment and closer to real application environment). In addition the scope varied from a basic groundwork study, a case study and two larger multi-user studies.

Chapter 4 begins with an evaluation of the 3D audio location controls used to implement the auditory displays tested in following studies. The controls were assessed in order to guarantee they offered *sufficient* accuracy for the auditory displays examined in this work. This was not an exhaustive evaluation but a requirement before any further spatial audio research is attempted. Without a firm understanding of a device's spatial audio capabilities it is impossible to investigate its use in an auditory display.

Chapter 5 builds on this work by evaluating the use of an egocentric spatial auditory display design (see Section 3.4 for a description of the different auditory display designs), evaluated statically in a lab environment. By carefully controlling the environment in this way, it was possible to compare a range of auditory display designs and produce conclusive results on their effect on auditory multi-tasking and cognitive load.

Rather than extending the work presented in Chapter 5 by re-evaluating the egocentric auditory designs in a a set of increasingly complex mobile lab experiments, attention was turned to exocentric auditory designs in order to compare the effect of combining both egocentric and exocentric display designs in the same application.

Full exocentric auditory displays require a real physical environment in order to locate exocentric audio sources. Thus in Chapter 6 and Chapter 7, the work in this thesis turned to 'in the wild' studies. The study reported in Chapter 6 was set in an audio-augmented 'sound garden' located in a municipal park in Madeira and the one in Chapter 7 in a lab-created art gallery environment. In order to experiment with both egocentric and exocentric designs a substantial infrastructure of engineering is required, i.e. user location tracking, head orientation tracking, and eyes-free input. Furthermore, very little experimental work has been carried out in this area, with previous studies typically focusing on single designs with very little quantitative evaluation (see Section 3.3). Therefore, the study in Chapter 6 was carried out as a case study to help develop appropriate quantitative evaluation, offer a clearer understanding of the engineering challenges, and broadly explore the use of non-speech audio within an exocentric auditory display design. In the study presented in Chapter 7, designs from the lab-based egocentric study (Chapter 5) were combined with exocentric displays (Chapter 6) in an indoor audio-augmented environment. This variety

of experimental approaches allowed the work presented in this thesis to address the key research questions, offering a wider scope at the expense of some experimental robustness.

Chapter 2

Human Spatial Audio Localisation of Real and Virtual Sound Sources

2.1 Introduction

The previous chapter defined the focus of this thesis, namely the investigation of 3D audio techniques for the design of eyes-free auditory interfaces supporting multitasking and location-based information for mobile interactive environments. In order to best understand how to design, implement and evaluate such techniques, it is important to review the literature on spatial audio localisation in both real and virtual environments. A more precise understanding into how humans locate sound sources in the physical world is necessary to compare the level of localisation accuracy that can be achieved in virtual environments and its limitations.

The aim of this chapter is to provide a background on sound and more specifically on spatial sound (or 3D audio). The chapter starts with a brief account of the characteristics of sound that is then followed by a literature review of the perception of sound localisation of both real and virtual sound sources.

2.2 From Sound to Human Auditory Perception

To understand how humans are able to perceive and localise sound, it is first necessary to understand some basic notions about the nature of sound itself. Research looking into the physics of sound and sound perception (Psychoacoustics) comprises a vast amount of work that would be beyond the scope of this thesis to cover (see (Handel, 1989) for a textbook review). Only a description of the concepts relevant to this thesis will be included in this chapter.

2.2.1 The Characteristics of the Sound Source

A sound wave is created by the mechanical movement of a physical object in a medium such as air. This movement generates vibrations in the atmosphere that will push air particles carrying the vibration through the air from the source to a listener (Moore, 2004). In this way, a sound wave will travel through the air as vibrations in air pressure.

A sound wave is typically described in acoustic terms as being either periodic (i.e. the wave pattern is repetitive) or aperiodic (i.e. the wave pattern is not repetitive). A simple periodic sound wave can be mathematically represented as a sine wave or sinusoid and it is usually described by its frequency, amplitude and phase. The *frequency* of a sound wave is measured in Hertz (abbreviated Hz). Hertz indicates the number of times per second the sound wave repeats itself (1 Hertz/Hz = 1 cycle or repetition per second or 1/sec) (Moore, 2004). The higher the frequency, the higher the perceived pitch and the lower the frequency, the lower the perceived pitch. The *amplitude* of a sound is the objective measurement of the amount of air pressure variation relative to atmospheric pressure caused by the sound wave (Moore, 2004). In this way, amplitude is directly related to the intensity of a sound and the amount of energy contained in the sound wave. The greater the intensity, the greater is the perceived loudness. However, as the distance from the source is increased, the intensity of the sound wave will also decrease exponentially. Hence, amplitude plays a key role in sound localisation (see Section 2.2.2).

The level of intensity our ears can perceive is usually measured using the decibel (dB) logarithmic scale in reference to Sound Pressure Level (SPL), which is the threshold of human hearing (hartmann, 1995). The faintest sound a human ear can perceive is known as the "threshold of hearing" and is equal to 0 dB SPL. A normal conversation is assigned a level of 60 dB SPL and the threshold of pain is usually established at 130 dB SPL. However, the amplitude in dB SPL does not directly equate to perceived amplitude for a number of reasons: 1) the frequency of the sound affects its perceived amplitude, i.e. sounds in the frequency range about 1 kHz to 4 kHz are perceived as louder than lower or higher frequency sounds, although the sound pressure is the same; 2) our sense of hearing is only roughly logarithmic; 3) individuals' hearing sensitivity varies. In the work presented in this thesis individual variation in the perception of the sound stimuli was not controlled for but the stimuli was always normalised to a conversational level at which individual variation is minimised. The third characteristic distinguishing a sine wave is called *phase*. Phase refers to a specific point in the cycle of a waveform as an angle, in degrees. This is a very important factor when sine waves interact with one another. For instance, if two identical sine waves are 180° out of phase they will cancel each other out in what is called phase cancellation.

Almost all of the periodic sound waves in the real world are complex periodic sound waves instead of simple sine waves. A complex periodic sound wave can be represented as a combination of multiple simple sine waves of different frequencies (spectral content) that form a harmonic series. Harmonics are vibrations at frequencies that are a multiple of the fundamental frequency. For example, if the fundamental frequency is 60 Hz, then the harmonic series is 120 Hz, 180 Hz, 240 Hz and so on. In order to determine the harmonic series of a sound wave, a Fourier analysis can be performed on the waveform using a fast Fourier transform (FFT) algorithm (Oppenheim and Schafer, 1989). By applying this algorithm, a spectral mathematical decomposition of the complex waveform can be obtained in the form of amplitude(dB)-versus-frequency over a single period of time. In other words, the time domain is transformed into the frequency domain. This algorithm is particularly important for the analysis of sound as it mathematically replicates the way our ears naturally decompose sound into frequencies before being sent to the brain. As different frequencies carry different information, this biological transformation allows humans to identify threatening sounds as well as being able to understand speech. Speech, however, is a good example of a waveform that can combine both periodic and aperiodic components. For example, the steady parts of vowels and voiced consonants correspond to the periodic part of the speech waveform, whereas the sounds produced without any vibration of the vocal cords, such as unvoiced fricatives and plosives, identify the aperiodic parts (Fry, 1979).

Unlike a periodic sound wave, noise is aperiodic (i.e. not periodic) and does not display a repeating pattern. Different types of noise can be identified depending on the random distribution of frequencies present in the sound. These different noises are usually labelled with colour names by analogy to the light wave frequencies, e.g. white, pink, blue, brown, grey, etc (Watson and Downey, 2009). White noise and pink noise are often used in psychoacoustic experiments to evoke audible resonances. Identifying and removing unwanted audible resonances improves the accuracy of audio systems and results in high-quality audio products. White noise is one of the best known noises and it is often reported to sound like the hiss of an untuned FM radio. It is characterised by a uniform mixture of random energy at every frequency of the audible spectrum (typically from 20 Hz to 20 kHz (Carlile, 1996)). In pink noise, on the other hand, white noise has been filtered to reduce the volume as frequency increases, with a power density of -3 dB per octave (Watson and Downey, 2009). In the experiments presented in this thesis, both aperiodic and periodic sounds will be used as stimuli.

Up till now in this chapter sound has been described as propagating in a free field without the presence of an obstacle such as a listener. The next section will focus on how the human auditory system is able to perceive and localise these sounds in an everyday natural environment with the ears uncovered and with the ability of moving the head.

2.2.2 Localisation of Real Sound Sources

Using only two ears the human auditory system is able to localise sounds from all directions. This phenomenon is generally known as "binaural hearing". Research on sound localisation investigates how humans are able to perceive the direction and distance of a sound source (Moore, 2004). The location of a sound source is usually specified by its azimuth (left to right position with respect to the facing direction of the head in the horizontal plane), elevation and distance in relation to the head of a listener (See Figure 2.1). In addition, the space around the listener's head is divided into three different areas that serve as a coordinate system for ease of reference. These areas are: the median (or sagittal) plane, on which any point is equidistant from the left and right ears; the horizontal plane, which is level with the listener's ears; and the frontal plane. See Figure 2.2 for an illustration of these different areas. Each coordinate is specified using azimuth, elevation and distance in a "spherical coordinate system", rather than an x, y, z system. All the positions of the sound sources included in the experimental part of this thesis will be specified with respect to the single pole spherical system. In this system the origin is the centre of the head and the azimuth and elevation are defined by lines of latitude and longitude respectively (Carlile, 1996).

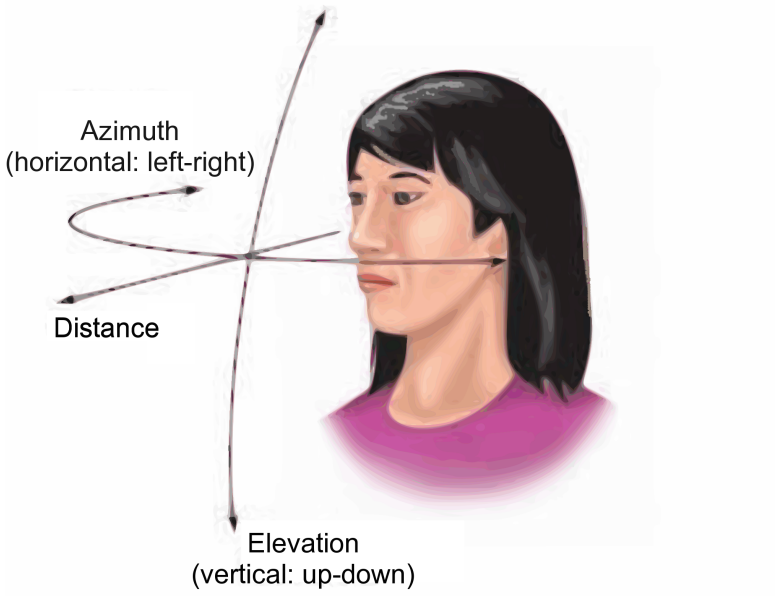


Figure 2.1: The three dimensions used to define the location of a sound relative to a listener's head: azimuth (horizontal: left-right); distance; and elevation (vertical: up-down). Adapted from (Goldstein, 2009).

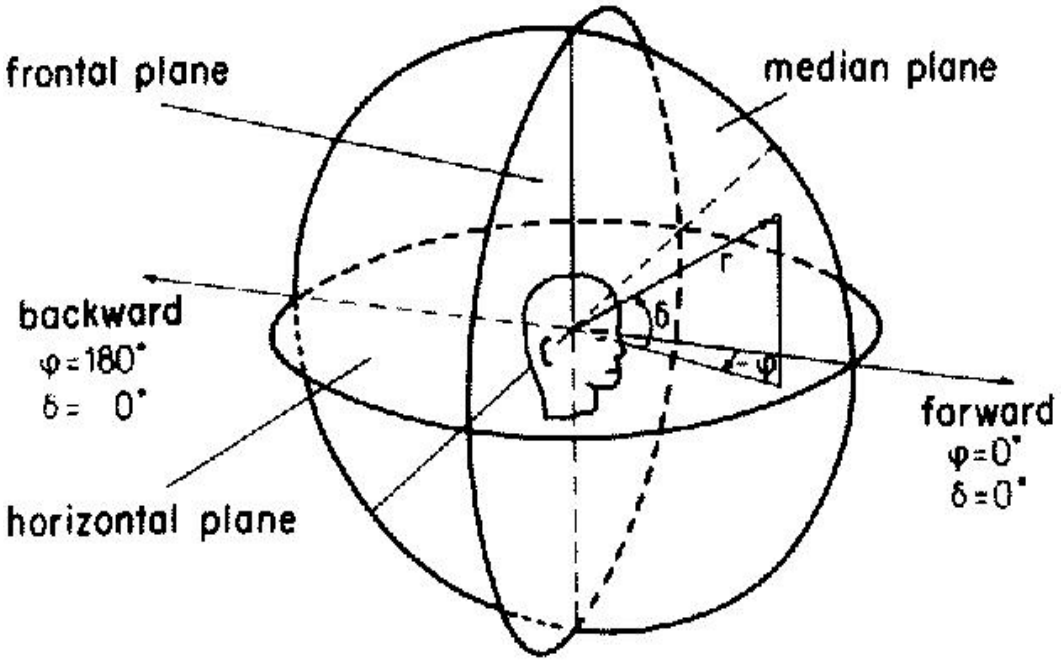


Figure 2.2: Coordinate system used to define the position of sound source relative to the listener's head. The horizontal plane defines the up/down dimension, the frontal plane defines the front/back dimension, and the median plane defines the right/left dimension. From (Blauert, 1997).

An explanation of sound localisation in real environments was provided by Lord Rayleigh's "duplex theory" (Rayleigh, 1907). He found that two physical cues dominate the perceived location (*azimuth*) of a sound source in the horizontal plane: an inter-aural time difference (ITD) and an inter-aural intensity difference (IID), also referred to as inter-aural level difference (ILD) when intensity is specified in decibels. These differences are also called binaural spatial cues. ITD refers to the difference in phase or time of sound waves reaching each ear. This cue states that unless a sound source is located directly in front or behind the head, sound arrives slightly earlier in time at the ear that is physically closer to the source, and with a perceived greater intensity (see Figure 2.3 for a schematic illustration). On the other hand, the ILD states that the 'shadowing' effect of the head prevents some of the incoming sound energy from reaching the ear that is turned away from the direction of the sound source. This shadow can attenuate the incoming sound by at least 6 dB between the two ears and up to 20 dB at higher frequencies. The ILD cue is generally considered ineffective for frequencies below approximately 1500 Hz (Blauert, 1997). At these low frequencies, sound waves wrap around the head minimising intensity differences. At frequencies above 3000 Hz, intensity differences are significant enough to act as a

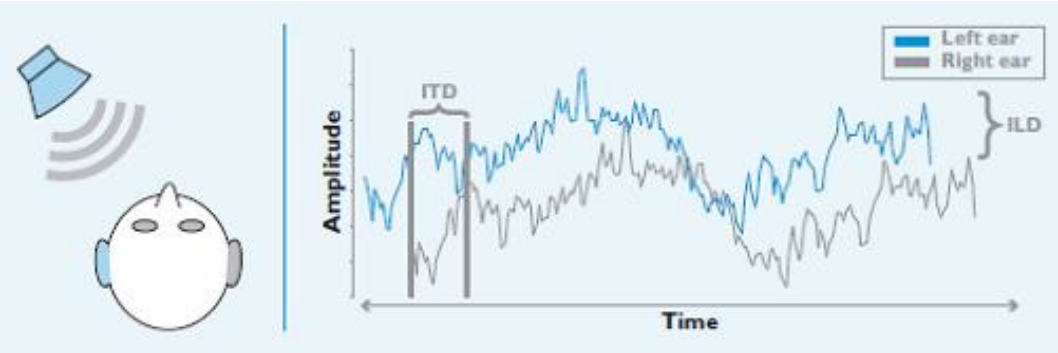


Figure 2.3: Schematic illustration of binaural cues, i.e. Inter-aural Time Difference (ITD) and Inter-aural Level Difference (ILD), created due to difference in how the sound arrives at the two ears. In this illustration sounds arrive first in the left ear with greater intensity. From (Litovsky, 2008).

cue to determine the sound source's position. Unlike the ILD cue, the ITD cue is effective for determining the position of low frequency signals of less than 1500 Hz. At a low frequency the phase of the sounds reaching the ear can be determined reliably because the waves are greater than the diameter of the head so the wave is not blocked but bends around the head instead.

However, Rayleigh's relatively simple ITD-ILD audio localisation model for the frontal horizontal plane does not provide a complete representation of our audio cues for localisation. When signals presented to the ears have only ILDs or ITDs, listeners can describe the extent to which the signals are to their left or right, but not whether they are in front of, behind, above, or below them. This phenomenon has been labelled as "the cone of confusion" because of the ambiguity of the ITDs and ILDs generated from these locations given that the distance between the sound source and each ear will be more or less the same (Mills, 1972) (see Figure 2.4). Due to the fact that identical ILD and ITD cues can be generated for multiple points in space, it is necessary for individuals to rotate their heads to accurately localise a sound. Moving the head not only improves localisation accuracy (Thurlow and Runge, 1967; Wallach, 1940; Begault *et al.*, 2001) but also reduces front-back ambiguities (Wightman and Kistler, 1999).

Rayleigh's model ignored the role of the the outer ear or pinnae in localising sound. When sound comes in contact with the pinnae, its frequency characteristics are modified (Batteau, 1967). These modifications vary depending on the position of the sound source thus providing an important directional cue and compensating for the limitations of the ITD and ILD cues. This cue is considered important for the localisation of *elevation* (vertical plane) (Middlebrooks, 1997; Musicant and Butler, 1985), as well as front-back discrimination and distance perception (Blauert, 1997).

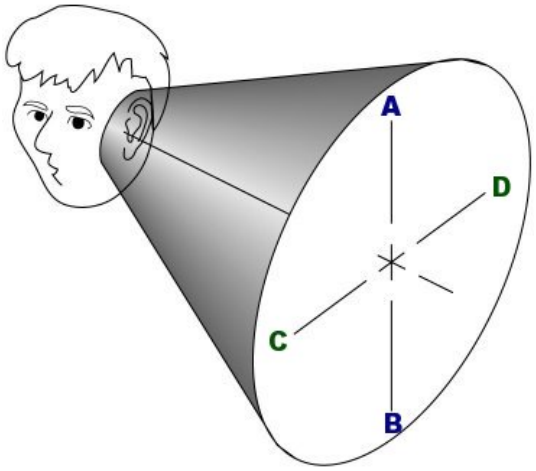


Figure 2.4: Illustration of the cone of confusion. Auditory sources A and B, and sources C and D share identical Inter-aural Time Difference and Inter-aural Level Difference, which could lead to confusion in localising these sounds in the absence of pinnae cues. Adapted from (Goldstein, 2009).

Research into localisation accuracy of human spatial hearing has focused on investigating how localisation error varies in a number of different directions around the human body. Part of this research has looked at the absolute accuracy of localisation by asking listeners to specify the perceived azimuth angle or direction of a single sound source (Makous and Middlebrooks, 1990; Oldfield and Parker, 1984). Others have focused their research on finding the "Minimum Audible Angle" (MAA) . MAA refers to the smallest noticeable difference in azimuth perceptible by a listener. Mills (1958) found that the MAA in the horizontal plane was approximately 1° when located straight ahead (0° azimuth) and could vary between 2° and 3° (Carlile *et al.*, 1997). It was also found that the MAA of a sound source located left or right of 0° was even larger being 3 to 10 times larger at 90° in either direction and twice as large at the rear showing a decrease in localisation accuracy (Blauert, 1997). These findings indicate that human audio localisation accuracy is better in the front hemisphere than in the back and worst to either side of a listener's auditory space.

Although it has not been studied as much as the other two cues for determining the location of a sound source, i.e. ITDs and ILDs, *distance* perception has been found to depend primarily on the sound pressure level at the position of the listener. Thus, loudness appears to be a crucial cue for determining distance. However, as mentioned in the previous Section 2.2.1, the varying

distance of a sound source from a listener will affect its perceived intensity or loudness. Given an anechoic environment, the "Inverse Square Law" can be used to predict how sound intensity is reduced with increasing distance. According to this law, an omnidirectional sound will have its intensity fall 6 dB as distance is doubled. However, the problem with this theory is that intensity, when measured in dB, is always measured against a reference level but loudness is based on the listener's subjective perception. A different approach investigated the reduction of intensity based on a loudness scale, such as a sone scale (Stevens, 1955). This approach was more successful at testing doubling of loudness by asking subjects to adjust sine waves that were twice as loud. This study showed that an increase of 10 dB equalled a doubling of loudness. However, both the inverse square law and the loudness scale techniques assume an anechoic environment which does not reflect how sounds are perceived in a real environment, in which reflection and reverberation are mixed in with the sound, especially in enclosed spaces. Also, loudness as a cue for distance is probably more important when the listener is not familiar with the sounds. Once the listener is familiar with a sound, the relationship between loudness and distance is learned from the listener's everyday exposure to a sound at different distances (Middlebrooks and Green, 1991). However, human auditory distance perception is still inaccurate even in real audio environments. Zahorik (2002) found that listeners would perceive a sound source located more than a meter away as closer and those located less than a meter away as further away. Moreover, a lot of variation was found even within one listener. For a more detailed review on cues for distance perception see (Coleman, 1963).

In summary, the most important perceptual cues for sound localisation include the intensity and the arrival time of the sound at each ear (to determine the azimuth); and the spectrum of the sound at each ear, i.e. the relative intensity of the sound at different frequencies (to determine the elevation). Based on these three cues an audio filter can be used to simulate the effect on a sound situated at a specific location. This is because an audio filter can be used to introduce a time lag, alter intensity and the spectrum of a free-field sound. Thus, one way of simulating the location of a sound is to use two audio filters, one for each ear. These filters are called head-related transfer functions (HRTFs) and will be discussed in more detail in Section 2.3.1.

2.2.3 Introduction to the Cocktail Party Effect

The human auditory system is also able to monitor several auditory streams simultaneously, selectively focusing attention on any one and placing the rest in the background. In 1953, Cherry (1953) investigated the problem of following only one conversation while many other conversations are going on around us. Cherry used shadowing tasks to study this problem, which involve playing two different auditory messages to a participant's left and right ears and instructing

them to attend to only one. The participant must then shadow or repeat this attended message. Cherry's work revealed that our ability to separate sounds from background noise was based on the characteristics of the sounds, such as the gender of the speaker, the spatial location of the sound source, the pitch, or the speaking speed.

Research by the psychologist Al Bregman on what he called Auditory Scene Analysis (ASA) (Bregman, 1990) tries to explain how the human auditory system takes multiple sounds from a complex natural environment (the auditory scene) and categorises them into separate streams. ASA specifies that a stream can be made up of a number of similar sounds and not just the one sound. Sounds that are similar to each other will be placed in the same stream and sounds that are different from each other will be placed in different streams. For example, in a music hall, the musical piece being played by the musicians would be placed in one stream and someone from the public coughing or sneezing during the performance would be placed in another stream. More importantly for the work presented in this thesis, Bregman found that two sounds originating from different locations are more easily segregated than two sounds originating from the same spatial location. For a more detailed overview of Auditory Scene Analysis, the reader is referred to Bregman (1990) and Deutsch (1999).

Many perceptual studies have examined how well listeners can extract the content of one sound source (the target) in the presence of competing sound sources (maskers), a situation requiring selective attention, as in Cherry's work (e.g., see (Devore and Shinn-Cunningham, 2003; Shinn-Cunningham, 2002; Hawley *et al.*, 2000)). Selective attention research investigates the extent to which we can focus on one task and ignore others. However, fewer studies have looked into how well listeners are able to understand the content of multiple, simultaneous sound sources (a situation requiring divided attention; e.g., see (Yost *et al.*, 1996)). Divided attention research investigates the extent to which we can do more than one task at the same time. Shinn-Cunningham and Ihlefeld (2004) have recently compared performance in selective- and divided-attention tasks. In the selective-attention task, subjects had to report content from one of the sources alone whereas in the divided-attention task, subjects had to report content from two sources. Results from this study showed that in cases where two independent sources of information are presented simultaneously to a listener, spatial acoustic cues can help a listener identify the content of multiple competing sources. Also, they appeared to use attention to spatial location and other features to modulate the salience of competing sound sources. However, the speech stimuli used in this work, i.e. two set phrases with varying contents, was far from the kind of auditory sources faced by real users in everyday applications.

2.2.4 Human Physiological Variation Effects on Sound Perception: Earedness and Handedness

Another important issue in spatial audio is that of physical human variation, such as head and pinnae size, that can alter the quality of the audio arriving in the eardrum. In addition, there is a possibility that differences in human perception and hemispherical organisation may also have an important effect in the way audio is perceived. These differences are of particular concern when employing audio techniques that expect users to monitor an auditory stream with only one ear.

The way in which most humans show a preference for one side of their body over the other is called "laterality". Many are right-sided, i.e. they prefer to use their right eye, right foot and right ear if forced to make a choice between the two. Handedness is the most obvious lateral preference and has been the most extensively studied (Hardyck and Petrinovich, 1977; Herron, 1980). The majority of humans are right-handed. The reasons for this are not fully understood, but it is thought that because the left cerebral hemisphere of the brain controls the right side of the body, the right side is generally stronger. It is suggested that the left cerebral hemisphere is dominant over the right in most humans because in 90-92% of all humans the left hemisphere is the language hemisphere (Porac and Coren, 1981).

The auditory system organisation also shows a right-hemisphere specialisation in spectral and spatial processing and a left-hemisphere dominance in temporal processing and speech perception (Zatorre, 2003). The equivalent of handedness in the auditory domain is termed "earedness". Earedness has been defined as "the preferential orientation of one ear toward a sound source or, alternatively, the preferential positioning of a sound source so that it stimulates one ear more than the other" (Noonan and Axelrod, 1981). This is comparable to handedness in that it is one of the hands that is dominant when performing a task but is not as strong as handedness (Hartmann *et al.*, 2001). Left- or right-ear dominance might affect the perception of the sound source location by increasing the localisation error in perception studies.

2.3 Localisation of Virtual Sound Sources

A great deal of sound localisation research has been devoted to investigating to what extent binaural cues can be replicated in a virtual audio environment. Audio techniques developed for this purpose include stereophonic sound, Wave Field Synthesis (WFS), ambisonics and binaural synthesis.

Stereophonic sound, or simply stereo sound, was invented by Alan Dower Blumlein in the 1930s. Stereo sound is based on the Duplex theory (described in Section 2.2.2) and it can be recorded in a number of different ways but in its simplest form it is recorded using two microphones separated by a distance that equals that of a human head. In this way, two separate monophonic (also monaural or simply "mono") signals are recorded with each microphone with their own specific level and phase relationship. The stereo recording can be then played back using a symmetrical configuration of at least two loudspeakers that will provide the perception of an image of the original source at some point between the loudspeakers. The main problem with this technique when reproduced over loudspeakers is that the optimal spatial impression is only achieve in a very small area or "sweet spot" between the two loudspeakers, where both the level differences and arrival time differences are small enough that the stereo image and localisation are both maintained. For instance, if both mono signals have equal intensity then the sound is perceived as being located in the centre, between the loudspeakers, provided they are at an equidistant from the listener. If the listener is closer to one loudspeaker than the other, then the stereo image will collapse and and the sound will appear to originate only from the loudspeaker that leads in time. This effect is explained by the Precedence effect (or "law of the first wavefront") that states that "at short delays the image location is dominated by the location of the leading source"(for a review on the Precedence effect see (Litovsky *et al.*, 1999)).

Other spatial audio reproduction techniques using loudspeakers include ambisonics (Malham and Myatt, 1995) or surround sound reproduction techniques and Wave Field Synthesis (Berkhout *et al.*, 1993). These techniques aim at reconstructing the acoustic sound field rendering the sounds as if they were physically present in the environment. The major drawback with loudspeaker-dependent techniques, when related to the topic of this thesis, is that they would be impossible to implement for supporting interactions in a mobile environment. For this reason, these techniques are unsuitable for the work intended for this thesis so no further literature review on these techniques will be provided in this chapter.

Spatial audio reproduction techniques that can be relayed over headphones include stereo and binaural synthesis techniques. Stereo sound can be played back over a pair of headphones, however, given that this technique is limited to left to right presentation of a sound source (usually referred to as "lateralisation"), the sound is most likely to be localised inside the head (Blauert, 1997) and not perceived externalised outside of the head as it would be desired when replicating a natural listening experience. Binaural synthesis, on the other hand, uses two separate channels, one for each ear, which contain time and spectral-shape cues generated by the shape of the pinnae and head width. In this way, both the characteristics of the source and the room can be acquired and used to provide a more natural-sounding experience. This is the basis of modern

virtual spatial hearing and it is implemented using a 3D sound system and stereo headphones. Due to the natural-sounding localisation experience of this audio technique and, most importantly, its portability, the experimental work presented in this thesis will use binaural synthesis over headphones throughout.

In the following section, the process to implement binaural synthesis over headphones will be described in more detail in order to highlight the strengths and limitations of this technique.

2.3.1 Binaural Synthesis: Head-Related Transfer Functions (HRTFs)

Two different approaches can be used when making a binaural recording. The first is to place a pair of microphones on a dummy's head and the second is to place the microphones on a human head. Both these techniques involve two microphones located at each ear on the dummy/human's head (William and Martin, 1995; Algazi *et al.*, 2001). In this way, a pair of Head-Related Transfer Functions (HRTFs), one for each ear, is recorded that can simulate how a sound changes on its way from the source to the listener's ear (see a schematic illustration of this process in Figure 2.5). These sounds are altered by diffraction caused by the torso, shadowing and boundary effects of the head, reflections off the pinnae and shoulders, and resonance in the ear canal (Martin, 2006). The signal that reaches the eardrum includes all the effects caused by the previously enumerated elements and more. The combination of these effects changes with different locations of the sound source and different orientations for the head. The end result is that your body creates different filter effects for different relationships between the location for the sound source and your two ears. Unless the sound source is on the median plane (see Figure 2.2), the signals arriving at your two ears will be different. This is why HRTFs are recorded for many locations of a given sound source relative to the head, resulting in a database of hundreds of HRTFs that describe the sound transformation characteristics of a particular head. HRTFs are then used to develop pairs of finite impulse response (FIR) filters for specific sound positions. Each sound position requires two filters, one for the left ear, and one for the right. To place a monaural sound at a certain position in virtual space, the set of FIR filters that correspond to the position is applied to the incoming sound to make it a spatial sound (see a schematic illustration of this process in Figure 2.6). Then, listeners wearing a pair of headphones will hear this sound as coming from a particular location in space. This filtering provides relatively good spatialisation in 3D virtual environments, however, the localisation error and the number of reversals or front-back confusions in a HRTF based system is still larger than in the real world and this is mainly due to human anatomical variability. Individuals differ in the shapes and sizes of their heads and pinnae. Thus, HRTFs also differ across individuals (Blauert, 1997).

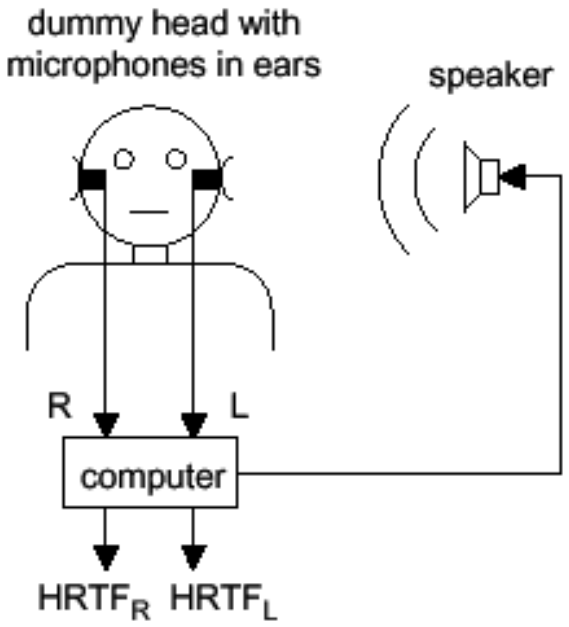


Figure 2.5: Illustration of how HRTFs are measured. From (Gardner, 1999).

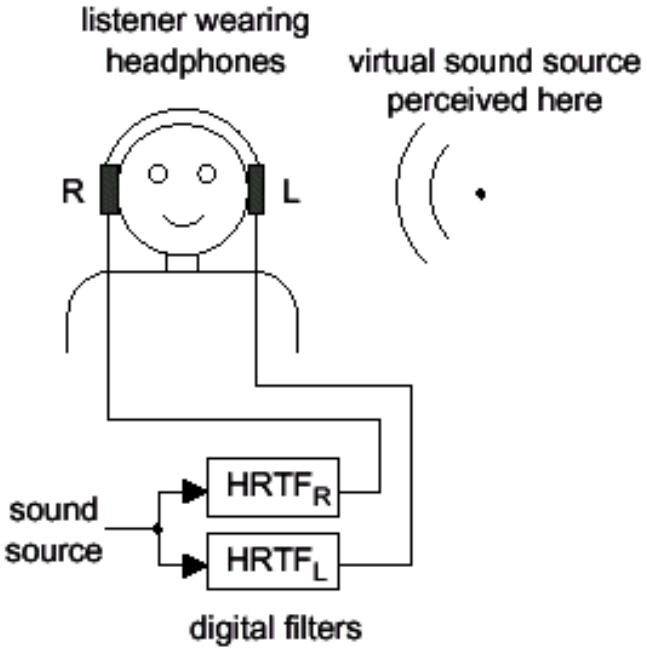


Figure 2.6: Illustration of binaural synthesis implementation using HRTFs. From (Gardner, 1999).

2.3.2 Sound Localisation using HRTFs

Wightman and Kistler (1989) compared localisation of real sound sources presented using loudspeakers to localisation of virtual (headphone-presented) sound sources synthesised using individualised HRTFs. Participants had to verbally report the perceived spatial location of the sound source. They found that, for the experienced listeners that took part in their study, the rate of front-back confusions increased from 6% for real sound sources to 11% for virtual sound sources. Otherwise, the results between real and virtual sound localisation were comparable. Mean absolute error in direction judgement (azimuth) ranged from 16° to 30° . In addition, large individual differences were found between the participants with some of them being significantly better than the rest at localising both real and virtual sources.

Bronkhorst (1995) also compared localisation of real sound sources and virtual sound sources synthesised using individualised HRTFs. Two tasks were used in this study. In the first task, participants had to turn their head while a sound was being played repeatedly and press a button when they thought they were facing the sound. In a second task, participants were ask to indicate in which quadrant of the horizontal plane a short sound was located by pressing a button. The stimuli were harmonic signals with a fundamental frequency of 250 Hz and a higher frequency from 4 to 15 kHz. Bronkhorst found that in the first task mean absolute errors decreased as the frequency range increased and were significantly greater for virtual than real sounds (virtual sources: 20° error for a cut-off frequency of 4 kHz and 14° error for a cut-off frequency of 16 kHz; real sources: 14° error for a cut-off frequency of 8 kHz and 10° error for a cut-off frequency of 18 kHz). Elevation errors were found to be the cause of the variation in these results as the errors in direction judgement were comparable between virtual and real sources. The second task tested confusion percentages relative to sound direction. The rate of confusions was significantly higher for virtual than real sound sources and to decrease significantly as cut-off frequency increased. However, the reason why small differences were found between virtual and real sources in the first task and larger differences were found in the second task seem to have been caused by localisation cues contained in high frequency regions above 7 kHz not appropriately simulated in the HRTFs synthesised for this study.

Wenzel *et al.* (1993) conducted a localisation experiment using non-individualised HRTFs. Localisation of real sound sources was compared to localisation using HRTFs measured on a 'good localiser' from the study by Wightman and Kistler (1989). The stimuli were broadband noise signals and the participants were non-experienced listeners. The average error angle in azimuth in low elevation was reported to be ~23°. In addition, confusion rates were found to be considerably higher for virtual sound sources (31% front-back confusions and 18% up-down confusions) than real sound sources (19% front-back confusions and 9% up-down confusions). These results

draw attention on higher front-back confusion rates when localising virtual sound sources using non-individualised HRTFs. Similar confusion rates of 29% were found by Begault and Wenzel (1993) when localising virtual sound sources (speech signals) using non-individualised HRTFs.

Head movement and sound source movement have been investigated as a way to reduce confusions. Wightman and Kistler (1999) conducted an experiment in which participants indicated the perceived positions of real and virtual sound sources in conditions in which head movements were restricted or encouraged. The sound source location was updated in real time based on the orientation of the participant's head. The results showed that front-back confusions nearly disappeared when head movement was encouraged. In a second experiment, Wightman and Kistler showed that, when the sound source was moved, front-back confusions disappeared only when the listener was the one in control of moving the sound source.

Begault *et al.* (2001) investigated the effect of head tracking, reverberation and individualised HRTFs on sound localisation when using virtual speech stimuli. The results showed that head tracking significantly reduced front-back confusions. Reverberation contributed to localisation errors and in this work HRTF type had no significant effect on localisation error. The authors attributed the null result for HRTF type to their experimental setup.

From the work presented in this section, it can be concluded that localisation error for individualised HRTFs is comparable to that of real sound sources. Using Non-individualised HRTFs result in slightly larger localisation error and an increase in front-back confusions. However, head tracking has been found to reduce these front-back confusions.

A generic head related transfer function that suits all individuals would be the perfect solution to localisation errors in spatial audio systems. However, due mainly to human anatomic variability, the development of a generic HRTF is highly unlikely. Individualised HRTFs provide better localisation results as they are custom generated for the individual but they are difficult to employ as the setup and equipment to acquire them is complex and very expensive. Non-individualised HRTFs provide worse localisation results but can be used by a much bigger number of users. This is the main reason why HRTF-based mobile phones use non-individualised HRTFs.

2.4 Conclusions

This chapter presented a review of the literature on sound perception and localisation that is relevant to the work presented in this thesis. The main aim of this review was to provide an understanding of the extent to which spatial audio can be simulated and the resources that can make this possible. This knowledge is required in order to understand the design of the 3D

audio techniques that will be tested in the experimental chapters of this thesis and thus makes a contribution to the thesis research questions included in Section 1.2: "To what extent can 3D audio techniques aid the user to maintain coherent attention on multiple auditory streams in a mobile eyes-free interface?" (RQ 1) and "How can 3D audio techniques be used to disambiguate multiple auditory sources in order to access location-based information in a mobile eyes-free interface?" (RQ 2).

In this review, a brief account of the characteristics of sound was first introduced that was then followed by a section highlighting the main research findings on how the human auditory system is able to perceive and localise real sound sources, including the ability to discriminate and separate multiple and simultaneous sound sources and the impact of human physiological variation on sound localisation accuracy. The final section of this chapter focused on current audio techniques that model human sound localisation in order to create virtual audio environments, paying special attention to those techniques that could enable mobile virtual audio environments.

The research included in this review has indicated that sound localisation is not one hundred per cent accurate even in real life. The main factors affecting accuracy are localisation errors and confusions, such as front-back ambiguities. However, research has found that the adverse effects of these two factors are minimised in the frontal plane. Thus, limiting the space of the auditory display to the frontal horizontal plane would help minimise these effects.

Currently, sound localisation in virtual environments is not as accurate as in the physical world and this is due in part to technological limitations. However, it has been shown that by synthesising a number of basic localisation cues, using HRTFs for instance, it is possible to create believable virtual audio environments. Although individualised HRTFs provide better localisation accuracy of virtual sound sources, it would be necessary to perform a costly and time-consuming calibration for each individual user that may not be feasible to perform in a mobile environment. Non-individualised HRTFs, on the other hand, although not as accurate due to human anatomical variability, can be used across a greater number of users without the need of an initial calibration and, most importantly, they are more widely supported on current mobile devices, which makes them more suitable for designing the spatial auditory displays to be tested in this thesis.

The next chapter will build on the knowledge of the research work presented in this chapter and will focus on the review of existing research in spatial auditory displays, paying special attention to design considerations when supporting multiple auditory streams.

Chapter 3

Review of Existing Research in Spatial Auditory Displays

3.1 Introduction

As discussed in Chapter 2, simulating natural spatial hearing in a virtual environment still faces many challenges. For this reason, the decisions made when designing an auditory display are of great importance, especially when designing eyes-free auditory interfaces for mobile environments. Interacting with audio information without any visual support can overload our auditory attention, especially when our visual attention is already engaged on a different task such as navigating the world. However, the spatial presentation of audio information provides orientation cues that aid segregation and attention switching between the auditory streams to maintain intelligibility when auditory information is being used (Stifelman, 1994; Ihlefeld and Shinn-Cunningham, 2008).

This chapter presents a review of the research literature on the design of spatial auditory displays, with special attention to those designed for eyes-free interaction. In addition, the design considerations for the implementation of the spatial auditory displays tested in the experimental chapters of this thesis will also be described. This review will further inform the thesis research questions put forwards in Chapter 1: "To what extent can 3D audio techniques aid the user to maintain coherent attention on multiple auditory streams in a mobile eyes-free interface?" (RQ 1) and "How can 3D audio techniques be used to disambiguate multiple auditory sources in order to access location-based information in a mobile eyes-free interface?" (RQ 2).

3.2 What is a Spatial Auditory Display?

In an auditory display, auditory means alone are used to display data, monitor systems, and provide enhanced user interfaces for computers and virtual reality systems. Although the term auditory display has not been formally defined yet, based on the body of research published as part of the International Conference on Auditory Display (ICAD), McGookin (2004) proposed the definition of an auditory display as "the use of sound to communicate information about the state of an application or computing device to a user". Such displays have also been called *auditory interfaces* (Gaver, 1997). In a spatial auditory display, spatial or 3D audio is used to enhance the auditory display by positioning information in the space around the user. This auditory display design enables the presentation of multiple and simultaneous auditory sources each situated at unique spatial locations.

3.2.1 Mapping Data to Sound in a Spatial Auditory Display

The effectiveness of an auditory display, including a spatial auditory display, will be influenced by the design of the display itself (see Section 3.4) but also the characteristics of the data being presented (Bly, 1985). A number of different semiotic mappings can be used for presenting information in the auditory display. These include Earcons (Blattner *et al.*, 1989), auditory icons (Gaver, 1989) and speech (Raman, 1997).

Earcons were originally developed by Blattner *et al.* (1989) and are structured non-verbal audio messages which use an abstract mapping between a music-like sound and the data. The main disadvantage of this audio technique is that the association between sound and event the Earcon is trying to represent must, at least initially, be explicitly learned.

Unlike Earcons, *Auditory Icons* make intuitive mappings between the sound and its intended function in a computer interface. They have been defined by Gaver (1997) as "Everyday sounds mapped to computer events by analogy with everyday sound-producing events". For instance, progress during a task may be represented by the sound of a bottle being filled with liquid. The disadvantage of Auditory Icons is that in some situations it might be difficult to find sounds suitable to represent certain abstract events. In addition, Auditory Icons can could be confused with actual environmental sounds (Cohen, 1994).

Earcons, used in menus for mobile interfaces, have been shown to have positive effects in decreasing performance times and errors (Brewster and Crease, 1999). In addition, previous work by McGookin and Brewster (2004) has shown that spatial location can be successfully used to improve the identification of simultaneously presented Earcons and suggest that spatial pre-

sentation should be used whenever practically possible when Earcons will be simultaneously presented.

Speech is the most popular technique for encoding data given that the way data is mapped to sound is more readily understood when compared to Earcons or Auditory Icons. Speech, synthesised or concatenated from audio recordings, has been used in many contexts, from screen readers for blind people, such as JAWS, to telephone enquiry systems and airline cockpits. However, speech might not be the best suited audio technique when simpler audio notifications could be used that would take less time to convey the message or limited processing is available. In these situations more abstract auditory techniques would be more appropriate.

3.3 A Review of Spatial Audio Applications

This section reviews earlier work employing spatial audio to design mobile interfaces, paying special attention to work including audio-only *eyes-free* spatial auditory interface designs supporting multiple auditory streams, which is the focus of this thesis.

3.3.1 Eyes-free 3D Auditory Interface Design

Pioneering work by Ludwig *et al.* (1990) and Cohen and Ludwig (1991) proposed the use of spatial audio to effectively present an eyes-free audio implementation of a visual window system (as displayed on a desktop interface). They called this audio management system 'Audio Windows'. In this system, a headphone-based spatial auditory display was used in which different applications were mapped to different parts of the audio space around the user. In the same way visual objects are distributed in a graphical user interface, audio objects are distributed in an auditory scene by using spatial audio. Handy Sound and MAW (Multidimensional Audio Windows) are two applications based on audio windows (Cohen, 1993). In the Handy Sound system, users interacted with spatial audio sounds using a set of hand gestures that allowed them to grab and move auditory sources in space to increase the separation between simultaneous sounds and improve their identification. MAW supported a teleconferencing environment in which users were given a position in space and their voices spatialised. In addition, hand gestures similar to the ones used in Handy Sound were enabled to allow users to interact with the sounds. Unfortunately, none of the systems proposed by Cohen were evaluated.

Schmandt's group at MIT focused on interactions with spoken information, such as news streams, documents, radio, phone calls and messaging. In their designs, spatial audio is used to assist simultaneous presentation of multiple auditory streams and gestures for interacting and navigating

the spatial audio environment. Audio Hallway (Schmandt, 1998) is a system in which spatial audio is used to create a computationally generated audio-only environment. The idea was to use each of the rooms situated off this virtual hallway to present detailed news excerpts while the user is travelling through the hallway. The user's head controlled movement in the virtual world.

AudioStreamer (Schmandt and Mullins, 1995) is another application that used spatial audio to improve the efficiency of browsing multiple simultaneous streams without a visual display. This application enabled users to listen to three spatialised simultaneous news streams while seated in a chair wearing a pair of stereo headphones allowing them to browse the audio, using speech/keyboard commands, in search of interesting segments and listen to it in detail. Gain or intensity increase was used to mark the beginning of a segment and changes in focus level. This application included head and hand gestures as part of the interface so the users could give a stream focus by using head rotation, hand motion or a combination of both.

Dynamic Soundscape (Kobayashi and Schmandt, 1997) is another application that used 3D audio techniques to enable browsing of content. As with the circular spatial audio progress bar designed by (Walker and Brewster, 2000) to facilitate the monitoring of background tasks, the Dynamic Soundscape used the same circular spatial mapping to represent the current time counter of a speech based audio recording. Dynamic Soundscape was inspired by AudioStreamer and was designed to simultaneously browse and monitor multiple parts of the same audio recording using synthetic speech and a speaker circling the user's head as it read out the audio data. This system used a maximum of four speakers simultaneously playing different portions of the same auditory stream and using loudness to focus on the speaker.

Whilst the previous systems have dealt with the use of spatial audio to interact with simultaneously presented auditory streams, these systems were not implemented for mobile environments or even run on a mobile device. Nomadic radio (Sawhney and Schmandt, 2000) is an early attempt to break away from the traditional desktop computing paradigm. This was not just an application but an audio-only wearable device that supported interaction with personal messages in a mobile environment. The output of the Nomadic Radio was spatialised and reproduced via shoulder-mounted loudspeakers, whereas the input was entered using spoken commands via a speech recognition interface. Messages were presented in the spatial position corresponding to the time of arrival, i.e. 12:00 in front of the user's nose, 3:00 and 9:00 to the right and left of the user, 6:00 right behind the user. The spatialised nature of this system enabled the presentation of multiple and simultaneous auditory streams that users could distinguish and separate from each other. Navigation allowed users to actively browse these messages via a synchronised combination of non-speech audio, synthetic speech and spatial audio techniques.

Unfortunately, although an informal evaluation was carried out for three of these applications (Schmandt and Mullins, 1995; Kobayashi and Schmandt, 1997; Sawhney and Schmandt, 2000), no formal usability evaluation studies of any of these applications were ever carried out. Thus, the usability and impact on user interaction of these 3D auditory interface designs is still unknown for an eyes-free mobile environment.

Other 3D auditory interface designs have used spatial separation to convey menu structure. Foogue (Dicke *et al.*, 2010) is an eyes-free spatial auditory interface purposely designed for state-of-the-art smartphones. Foogue allows the user to navigate, select or manipulate spatialised audio items from a hierarchical menu. All items are arranged in a 120 degree arc in front of the user and displayed in sequence. Unfortunately, no evaluation of this system was carried out to assess its effectiveness. Diary in the Sky by Walker *et al.* (2001) used a 3D audio radial pie menu, with the user's head in the middle of the pie, to encode the times of diary appointments. Using a desktop simulation, the diary entries were consecutively presented for selection according to their time of appointment, as in the Nomadic Radio system described earlier in this section. Although spatial audio significantly improved user performance in this system, its usability in a mobile environment is not known and in addition, it is unclear to what extent the presentation of audio information sequentially might have affected user interaction. Similarly, the earPod application (Zhao *et al.*, 2007) was used to evaluate the usability of a spatialised radial menu in which audio items were displayed sequentially. The efficiency of this audio menu was compared that of an equivalent visual menu display. User interaction was performed using a circular touchpad that reinforced the user's cognitive mapping between menu items and spatial locations on the touchpad. It was found that earPod was efficient to use, relatively easy to learn and comparable in both speed and accuracy with a visual menu selection technique. Unfortunately, only an informal evaluation of this system was carried out in a mobile environment and simultaneous presentation was not explored.

A spatial auditory environment for non-visual interaction with simultaneously presented auditory objects was developed by (Crispien *et al.*, 1996). Interaction with a maximum of three auditory objects simultaneously in a hierarchical menu was enabled using a virtual "auditory focus area" covering an angle of 90° in front of the user that was dynamically updated using a head-tracking device. However, this system was not a mobile implementation and an evaluation has never been reported.

Brewster *et al.* (2003) conducted a study to compare sequential and simultaneous sound presentation in a mobile radial audio pie menu interface. Three conditions were tested in this study in which sounds were presented sequentially in an egocentric or exocentric display and simultaneously in an exocentric display. In an egocentric display the position of the auditory sources

is perceived as relative to the user and in an exocentric display as relative to the world (egocentric and exocentric displays are described further in Section 3.4). A head gesture was used for selection. The results from this study showed that the egocentric display design was more effective, however the exocentric designs were only partially exocentric as they depended on head orientation but not on user position or user orientation. Marentakis and Brewster (2005) also investigated the usability of egocentric and exocentric auditory displays. Users were asked to select a target sound amongst a number of distracters using a physical pointing gesture while standing, with the help of a loudness cue, a timbre cue and an orientation update cue and combinations of these cues. The results showed that in the egocentric display participants were faster but less accurate, whereas in the exocentric display they were slower but more accurate. However, the exocentric display design used in this study was again, as in (Brewster *et al.*, 2003), only partially exocentric. Furthermore, the sounds in these exocentric displays did not relate to any targets physically located in the space.

The applications described in this section used 3D auditory interface designs to present auditory information in various settings, lab, desktop, mobile. However, very little formal evaluation was carried out, and apart from Brewster *et al.* (2003); Marentakis and Brewster (2005), only one auditory display configuration was tested in each application. Furthermore, the majority of these work was not tested on a mobile device or in a mobile environment. In the following section, a review of mobile 3D auditory interface design for audio-augmentation of physical spaces is presented.

3.3.2 Audio-Augmented Reality

In a mobile audio-augmented reality application auditory information is situated in the physical space and presented based on the user's physical movement around that space. This section presents previous work on 3D auditory interface design for user interaction with location-based information in both outdoor and indoor mobile audio-augmented reality applications.

Outdoor Mobile Audio-Augmented Reality Applications

A number of spatial audio applications designed for outdoor environments have used global positioning system (GPS) receivers to identify personal locations and to sonify interesting and relevant points of interest nearby (Mariette, 2007). These applications make use of spatial audio to create an Audio-Augmented Reality (AAR). Cohen *et al.* (1993) referred to AAR as the action of superimposing virtual sound sources upon real world objects. Early applications demonstrating the concept of AAR for mobile environments include Here&There (Rozier *et al.*, 2000). The

Hear&There system was able to determine the location and head position of the user using the information from GPS and a digital compass. This system used 'audio imprints' at the points of interest. Audio imprints were "customizable collections of sounds that [could] be placed in the space" and consisted of "a single primary sound, with other audio braided in the periphery. These braids overlap the imprint, with each braid of audio shifting into and out of prominence". Users could listen to these imprints by walking into the area the imprint occupied which was triggered by proximity. However, no further details on how these imprints were implemented or formal evaluation was provided for this work. *Riot! 1831* (Reid *et al.*, 2005) included similar techniques in a more sophisticated application that recreated the Bristol riots of 1831 as a location-based audio drama in the streets of modern day Bristol. Users walked around one of the squares in the city equipped with a small backpack containing an iPAQ PDA, a GPS receiver and a pair of headphones; user position was used to trigger a variety of non-overlapping sound effects and script files based on real events that took place in the square. A quantitative and qualitative evaluation of the Riot! 1831 system showed a deep level of immersion for users in this exploratory experience.

Route finding applications such as the AudioGPS system (Holland *et al.*, 2002), Mediascapes (Cater *et al.*, 2007), Audio Bubbles (McGookin *et al.*, 2009) and Soundcrumbs (Magnusson *et al.*, 2009), have used non-speech sounds as an auditory beacon to support navigation tasks and guide users to points of interest. These beacons alert users of their proximity to a location of interest through a brief repeating sound such as an Earcon or an Auditory Icon (see Section 3.2.1). Generally, two concentric levels of beaconing feedback are implemented around a landmark, the first in a wide proximity zone and the second in a narrower activation zone (Stahl, 2007). The goal of audio cues in the proximity zone is to provide unobtrusive audio guidance, which enables a user to move towards the activation zone. Once this inner zone is successfully reached, additional content is made available to the user, either to indicate that a landmark has been found or to provide structured information describing it. The implementation of these proximity and activation zones vary across applications. For instance, the proximity zone had a radius of 250 meters in the Audio Bubbles study, 55 meters in the Mediascapes, 20 meters in the Soundcrumbs study and not reported for the AudioGPS. The activation zone had a radius of 10 meters in the Audio Bubbles study, 5m in the Mediascape implementation, not used in the Soundcrumbs study and no information was provided for the AudioGPS application.

Other applications like Roaring Navigator (Stahl, 2007), which used sound to indicate the location of an animal enclosure in a zoo, estimated the position and orientation of the listener's head by means of a GPS receiver and magnetometer, and also used stereo panning to indicate the direction of a navigational goal, i.e. animal sounds, located at the various enclosures in a

zoo both in a navigational and an exploratory scenario. This implementation is similar to AudioGPS and Mediascapes in that the landmarks were spatialised using stereo panning and more complex than the Audio Bubbles and Soundcrumbs implementations. Audio Bubbles did not spatialise the landmarks and only used distance mapped to the repetition rate and volume of a short 'click' sound to indicate that the user was near a point of interest (replicated the Geiger counter principle implemented in AudioGPS). Similarly, in the Soundcrumbs system, which allowed users to sonify their GPS position as they walked creating a trail of sonified locations or "soundcrumbs" that would enable them to retrace their steps, the proximity of a soundcrumb while navigating the sound trail was mapped to a linear increase of the sound's volume. In addition, Stahl's (2007) system allowed for the simultaneous playback of five spatial sound sources but no detailed investigation was carried out into how this affected the user experience.

Apart from the use of non-speech audio, such as abstract or animal sounds for navigational tasks, other studies have explored the use of music in a similar manner. Examples include the Tactical Sound Garden (TSG) (Shepard, 2006), Mobile Immersive Music (Lemordant and Guerraz, 2007), the Melodius Walkabout project (Etter and Specht, 2005), ONTRACK (Jones *et al.*, 2008), and gpsTunes (Strachan *et al.*, 2005), a system in which a user's own music was spatialised through the panning of the sound across the stereo sound stage to aid users navigate towards a landmark. Except for the TSG application and the Melodius Walkabout project, all the other systems logged heading data using magnetometer sensors supported on the mobile device. However, no heading data analysis was provided in the ONTRACK or the Mobile Immersive Music study, and in the gpsTunes system, heading data were used to identify at what point users were trying to locate the direction of targets by rotating around and pointing the device at each target.

Other applications, such as Lyons *et al.* (2000) and more recently Heller *et al.* (2009), made use of ambient sound and narration to construct their sound environments. Interestingly, Heller *et al.* tracked head orientation in a non-realistic Wizard-of-Oz experience by mounting a compass sensor on the headphones worn by the user and, although no user experience evaluation was carried out, they observed that turning the head was the key to navigation by ear in this kind of mobile AAR environment. The importance of head-turning data was also highlighted in Mariette's (Mariette, 2010) experimental work on outdoor navigation performance.

Spatial audio applications designed for outdoor environments have been generally used for supporting navigation and little work has supported casual exploration. Although in this body of work similar designs have been used to advertise locations of interest, the way the auditory displays were implemented varied greatly from one system to another. Furthermore, very little comparative and evaluation work has been carried out, especially taking into account head ori-

entation data which could offer a valuable insight into how users interact with these mobile AAR systems.

Indoor Mobile Audio-Augmented Reality Applications

Fewer mobile AAR applications have been implemented for indoor environments than outdoors. This is mainly due to greater technological complexities when tracking user location indoors when compared to GPS technology which is ready available on most mobile devices. The majority of these indoor systems have been developed for museums, exhibitions or historic sites in order to replace linear keypad-based audio tour guides that constrained users to a linear access to information, and which could pull the visitor's attention away from the actual exhibits and disturb the overall user experience. As with outdoor AAR systems, the key idea is that visitors are able to explore a virtual environment augmenting a physical space solely by listening as they walk. In Eckel's (2001) words: "by moving through real space, users automatically navigate an attached acoustic information space designed as a complement or extension of the real space". An early example of such system is Bederson's automated tour guide (Bederson, 1995). This prototype system relied on a non-linear playback system and codes locally broadcasted by small infrared transmitters installed above every exhibit. The visitor had to carry a random access audio device, a modified Sony MiniDisc player, and a custom infrared receiver that would track the location of the visitor. As the visitor came close to an exhibit, the associated comment would automatically start and then stop if the visitor walked away. Similarly, the Audio Aura system (Mynatt *et al.*, 1998) was designed to provide serendipitous information via audio cues based on the motion of the user in the workplace. The location of the user was tracked using an active badge system that triggered audio delivery in the Audio Aura system. The design of these audio cues combined speech, music and sound effects to provide peripheral information such as calendar reminders, email status and information of activities of other colleagues. This information was relevant not only in the general context of the receiver, but also semantically connected to the physical space. Unfortunately, no formal evaluation of these early prototypes was carried out to determine their effectiveness.

In indoor AAR applications audio-augmented locations are advertised using user proximity, however the increase of information available for the user at these audio-augmented locations has made the design of the auditory displays within the activation zones more complex. The ec(h)o system (Wakkary and Hatala, 2007) presented a dynamic audio experience of the Canadian Museum of Nature in Ottawa. This system consisted of a pair of headphones connected to a wireless receiver for audio and a combination of RFID-based and optical tracking of the visitor's position. The audio experience included an ambient soundscape and short audio sequences in

the form of audio prefaces. The audio prefaces were like "teasers" that acted as multiple-choice options for the audio objects, which contained a greater depth of information. As the visitor walked through the exhibition space, different ambient sounds faded in and out according to the artefacts they related to in the visitor's proximity, to invite the visitor to take a closer look. However, as the visitor's orientation was not tracked by the system, the ambient sounds were not spatialised. Using a tangible object in the form of a cube, the visitor was able to interact with the paired prefaces and audio objects using an auditory display when in front of a display of artifacts. The design of the auditory display was simple using the left channel audio for the left, right channel for the right and both channels for the centre, presenting the prefaces to the user from left to right in a sequential order. The spatial arrangement of the auditory display was mapped to the tangible interface for selection. Only a preliminary overall qualitative evaluation of the ec(h)o system was carried out that consisted of a questionnaire and a semi-structured interview limited to six different participants that tested the system. This evaluation, however, did not include objective measures of user performance or behaviour that would support the qualitative results.

More sophisticated systems like the LISTEN project (Eckel, 2001) have been implemented to deliver a tailored audio-augmented user experience in a museum environment. The LISTEN system was deployed in the August Macke art exhibition at the Kunstmuseum Bonn (Terrenghi and Zimmermann, 2004). This system consisted of a pair of wireless headphones equipped with a wireless navigation transmitter used to determine the user's position and head orientation and eight receivers deployed in the exhibition room to obtain a maximum accuracy of the head position of approximately 10 cm. The information acquired by the tracking system made it possible to spatialise the virtual sound sources precisely and divide the physical space into virtual zones. The virtual acoustic space created by the LISTEN system used two models: the world model and the location model. The world model contained the detailed geometric information of the exhibition space and its objects, whereas the location model defined areas within the world model that the visitor could interact with. These interactive areas were referred to as object zones and near fields (Goßmann and Specht, 2002). The object zone was defined using a square shape that acted as a type of proximity zone establishing a connection between the user's position and the corresponding physical object, whereas the near field was defined using a circular shape that acted as a type of activation zone connected to smaller parts of the physical object containing more detailed audio comments. This detailed information was only audible if the visitor was located at a specific angle and distance away from the physical object and the visitor was facing the object. Again, as with the ec(h)o system, a qualitative evaluation of the general acceptance of the LISTEN system was carried out that comprised a questionnaire and handwritten statements

from the visitors but no objective measures of user performance or behaviour data was presented.

Mobile audio-augmented applications running the audio engine directly on a mobile platform, such as a smart phone, are a fairly recent development driven mostly by the increasing advances in mobile phone technology. The CORONA system (Heller and Borchers, 2011) created an interactive audio experience in the Coronation Hall in Aachen, Germany, simulating a virtual audio space rendered on an Apple iPhone and presented over a pair of headphones. The audio space was rendered using simple stereo panning and then extended by adding a low-pass filtered sample played back if the virtual sound source was behind the listener and a reverberated sample was used as a distance cue. The audio space included a background atmosphere and ten source areas where information was presented to the visitor. Each of these source areas was surrounded by a circular proximity and activation zone which were triggered by the visitors as they explore the space. The user's position was tracked in real time and the head orientation measured using a digital tilt-compensated compass in order to spatialise the source areas. The usability and user acceptance of the interactive audio space of the CORONA system has not yet been evaluated.

Early indoor AAR systems relied on triggering mechanisms to access location-based information. Indoor tracking systems, a requirement for fully spatialised audio augmentation, are still in a very early stage. In addition, as with outdoor mobile AAR systems, very little comparative evaluation has been carried out.

3.3.3 Summary

The spatial auditory displays described in this review section have been implemented using a wide range of different 3D audio techniques in order to present multiple sources of information. These differences make it difficult to compare the efficiency and usability of these types of interfaces across the different applications. In addition, the design of these spatial auditory displays vary enormously but little systematic user evaluation has been carried out on how these differences in design may affect their usability and as a consequence the user experience. Thus, a more detailed and controlled investigation into the effects of spatial auditory display design on user interaction would make a timely contribution to the field.

In this thesis, a number of different choices for designing spatial auditory displays that support multiple auditory streams will be tested both quantitatively and qualitatively within an auditory multitasking scenario as well as a mobile AAR environment.

In the following section design choices for spatial auditory displays will be described and an explanation of how the experimental approach of this thesis addresses and guides these choices will be offered.

3.4 Design Choices for Spatial Auditory Displays Supporting Multiple Auditory Streams

Spatial audio is able to mirror the spatial organisation of a visual display, thus allowing the creation of a spatial mapping. In this way physical metaphors can be represented using 3D audio techniques. For example, a radial or pie menu around a user's head representing the time around a clock (Brewster *et al.*, 2003; Walker *et al.*, 2001), or exhibits being advertised from their actual physical locations to attract visitors in an AAR application (Goßmann and Specht, 2002). Invariably, these spatial mappings involve the presentation of more than one item of information.

When designing spatial auditory displays that support the presentation of multiple sources of information, choices have to be made on both the presentation and the spatial arrangement of the auditory sources. Namely:

1. Continuous *versus* non-continuous auditory streams – the auditory streams that form part of the auditory display could be continuous or non-continuous. In other words, audio could be streamed like a podcast or radio program or just short audio prompts such as menu items in an audio menu.
2. Degree of active *versus* passive attention – the degree of attention required from the user when listening to an auditory stream needs to be quantified. Different degrees would be expected when the user actively listens to a set of instructions to reach a specific location than when listening to a piece of music in the background.
3. Sequential *versus* simultaneous sound presentation – presenting auditory streams sequentially will prevent information sources from competing with each other but this could result in a more lengthy interaction when switching between sources, poorer recall of earlier information, and irritation caused by continuous interruption. The human auditory system allows humans to monitor several auditory streams simultaneously, selectively focusing attention on any one and placing the rest in the background (for more information on the Cocktail Party effect see Section 2.2.3).
4. 3D audio techniques to place sounds in different spatial locations *versus* a single point of presentation – although audio is often seen as a single stream coming from a fixed point, if users are wearing headphones, 3D audio techniques can create the perception that a sound is coming from a specific spatial location (Begault, 1994).
5. Egocentric *versus* exocentric location – the sound source position in an auditory display can be perceived as relative to the user (egocentric design) or relative to the world (exocentric design). In an egocentric display, elements are always in a fixed position relative to the user, which can be particularly useful for mobile users as changes in orientation when

moving are inevitable. Egocentric displays are more suitable for interactions that exhibit a repeatable pattern, such as interactions with lists or menus because display elements are always in a fixed position relative to the user so they are easy to remember. In an exocentric display, on the other hand, display element positions have to be updated in real-time according to the user orientation as they appear to be fixed to the world. This is usually implemented using a head-tracking device that provides the orientation of the user's head that is then delivered to the spatial audio engine, which updates the sound positions. As a result, a sound that is located to the right of the user will be perceived as originating from the front when the user's head is turned to the right. Head tracking has been found to help greatly in resolving front-back confusions (Begault *et al.*, 2001). Exocentric displays are well suited for navigation in virtual or real worlds but they are more computationally intensive than egocentric displays, as the sound scene needs to be rendered at a fast rate for convincing results.

6. Dynamic movement *versus* fixed locations of auditory sources – the use of 3D audio also raises the issue of how streams are presented in the auditory display over time. Not only can audio appear to come from a specific position, this position can be dynamically moved. The movement of items in visual interfaces is commonly used to enhance the interface. For example, animating a window as it is minimised. Such techniques can also be used in an auditory interface. For example, moving an auditory stream to the side while a second stream is played from the front.
7. Speech *versus* non-speech sounds – in an auditory display, information is conveyed using speech or non-speech sounds. Non-speech sounds (Earcons and Auditory icons, as described in Section 3.2.1) are good at conveying structured information and providing rapid feedback, whereas speech sounds (as described in Section 3.2.1) can convey more complex information such as absolute values and instructions (Brewster, 2002).

All of these different features relate to decisions that have to be made when designing a spatial auditory display. In this thesis, a series of experiments will be carried out to evaluate these features.

3.4.1 Auditory Multitasking

When making design choices in an auditory display, it is critical to consider how to manage multiple auditory streams without overloading the user. In a visual display users can move their gaze or look away from an interface if required. However, in an auditory display users cannot *shut* their ears.

When multiple tasks are supported purely by audio, users must be able to direct their attention selectively to each individual auditory stream representing a task. However, when performing multiple tasks at once and conveying them through audio simultaneously, masking of information (i.e. both auditory streams are audible and easily confused with each other) occurs as the auditory streams overlap.

Auditory scene analysis (Bregman, 1990) (see Section 2.2.3) has shown that two sounds originating from different locations are more easily segregated than two sounds originating from the same spatial location. This effect can be duplicated using spatial audio and has been used in previous research (Schmandt and Mullins, 1995; Walker and Brewster, 2000; Brungart *et al.*, 2002) as a successful technique to segregate multiple auditory streams by placing each auditory stream at a different location around the user's head, mirroring how humans perceive sounds in real life (Bronkhorst, 2000).

What is not yet clear is what spatial audio design might be the most effective for supporting multiple auditory streams and how much, if at all, the spatialisation of the streams might contribute to an increase or decrease of the user's cognitive load when engaged in a number of simultaneous tasks.

3.4.2 The Importance of Cognitive Load

Different attention demands impose different amounts of cognitive load on the user. A mobile user can listen to a voicemail left by a friend while monitoring his music (a selective-attention task). However, if the same mobile user is talking to a friend while interacting with the calendar using an audio menu to find a suitable time for a meeting, this user is dividing his/her attention between both auditory streams (a divided-attention task). The first task results in less cognitive load, and the second in higher cognitive load. Cognitive load has been described as the amount of mental resources needed to perform a given task (Draycott and Kline, 1996). As tasks add up, the mental resources needed increase and cognitive load rises. Previous research by Marentakis and Brewster (2005) investigating pointing efficiency in deictic spatial auditory displays, showed that increased cognitive load resulted in reduced pointing efficiency. Shinn-Cunningham and Ihlefeld (2004) and Best *et al.* (2005) have also investigated how perceived spatial separation of sources and consistency in source locations influences performance on selective- and divided-attention tasks. They found that performance was better when sources were perceived at different locations instead of the same location. However, she adds that "further experiments are necessary to determine whether spatial attention influences performance differently when competing sources differ from one another in more natural ways".

When designing auditory interfaces it is critical to consider the attention demands expected from the user. This affects the attention required to monitor the information being relayed by the stream and also the attention required to monitor the spatial location of the stream. Spatial audio offers the ability to foreground and background auditory streams, for example, moving an auditory stream to the side (*spatial minimisation*), while a second stream is played from the front. Spatial minimisation could help users alter focus between streams. Buxton (1995) differentiated between foreground and background tasks and defined the former as "activities which are in the fore of human consciousness intentional activities" and the latter as "tasks that take place in the periphery 'behind' those in the foreground". In this thesis, the focus will be on a traditional view of foreground and background perception where two auditory streams, one offering a user-driven audio menu, and a second providing continuous streamed audio information, compete for attention.

3.4.3 Access to Location-based Information

Designing spatial auditory displays that enable interaction with a wide range of information without overloading the user becomes even more problematic in an AAR system. The main concern in an audio-augmented environment is that the way information is presented could unnecessarily divert the user's attention from the task at hand, in this case exploring and discovering a particular physical location that this information is augmenting. The end result being that the user will stop interacting with the audio content.

A location-aware audio-augmented space usually consists of a virtual audio environment superimposed on a *real* physical space featuring a set of precisely situated sounds surrounding the user. A sound garden (Shepard, 2006) and an audio-augmented art exhibition would be examples of such spaces. Advances in mobile technology enable users to interact with these virtual audio environments when on the move. In contrast to a simulated virtual reality environment in which participants are abstracted from the reality they are interacting with, in a mobile AAR environment participants interact with the virtual audio mixed with *real* vision and motion.

Virtual audio environments such as a sound garden or an audio-augmented art exhibition are usually intended for users to explore and experience casually rather than navigate via predefined paths. The unstructured nature of this activity presents unique challenges for the design of an auditory display to support exploration. Fundamentally, individual audio-augmented locations need to advertise themselves both to attract the user's attention and support subsequent targeting. This is typically achieved through a combination of user tracking technology (e.g. Global Positioning System (GPS) or Infrared (IR) sensors) and auditory beacons that activate when a

user is within a specific distance from an audio-augmented location, typically within a circular proximity and activation zone (see Section 3.3.2). Any error provided by the positioning system used will tend to require an increase in the size of these zones. Furthermore, the more unstructured and exploratory the environment, the more important the proximity zone becomes as a means of advertising locations. In a real environment, there is a likelihood that proximity zones may overlap if audio-augmented locations are situated close to each other.

One way to manage the presentation of overlapping audio-augmented locations is using spatial auditory displays. Although much work has examined the use of spatial audio for AAR (see Section 3.3.2), less work has compared different auditory feedback strategies (Mynatt *et al.*, 1998; Marentakis and Brewster, 2006), and no work has investigated the use of 3D audio HRTF techniques (described further in Section 2.3.1) in a non-guided mobile AAR environment, especially dealing with the problem of overlapping audio-augmented locations.

3.5 Scope

The work presented in this thesis touches on a variety of different areas: Psychology, Engineering, Psychoacoustics and Human-computer interaction. As such, it is important to make clear the methodology applied in this work and, furthermore, to delineate what this thesis does not intend to address.

User Experience In some cases the term *user experience* has "...become a catchphrase, calling for a holistic perspective and an enrichment of traditional quality models with non-utilitarian concepts..." (Hassenzahl, 2005). In contrast, the studies presented in the experimental chapters of this thesis, the term *user experience* is used to describe the formal and informal feedback collected from users through interviews and questionnaires after an experiment. For example, did a user respond to the question 'Did you enjoy using the application?' positively or negatively. This qualitative data was then used to support the analysis of the quantitative usability data.

Psychology Although some of the work presented here focuses on the issue of cognitive load, it is not a psychological study. The interest is purely on how differences in psychological load affect positive or negative response and performance to auditory display designs. The subtleties of how multimodal stimuli affect and modify cognitive load are not the subject of this work.

Psychoacoustics A classic psychoacoustic approach might be to modify the audio sources (typically the intensity, frequency content and duration) in order to investigate the effect on the

user within the same auditory display design (Pressnitzer *et al.*, 2006). In this work the audio sources are not modified except *in terms* of the auditory display design (for example, presented from a single point or spatially). As such, the psychoacoustic effects of spatial audio outwith the auditory display designs explored in this thesis are not considered.

3.6 Conclusions

This chapter has offered a review of previous literature covering the use of spatial auditory displays for the support of multiple auditory streams in an eyes-free audio environment. This review showed that the design of such auditory displays has varied enormously but little systematic user evaluation has been carried out to determine how these differences in design may affect their usability and the user experience, especially on mobile interfaces.

This chapter has identified the different design choices for presenting multiple sources of audio information as:

- Continuous *versus* non-continuous auditory streams.
- Degree of active- non-active attention.
- Sequential *versus* simultaneous sound presentation.
- 3D audio techniques to place sounds in different spatial locations *versus* a single point of presentation.
- Egocentric *versus* exocentric location.
- Dynamic movement *versus* fixed locations of auditory sources.
- Speech *versus* non-speech sounds.

While the majority of these design choices have been suggested in previous research, they have not been evaluated formally against each other. This thesis proposes a systematic evaluation of these design choices to determine their usability in an interactive mobile environment where we need to consider how to manage multiple auditory streams without overloading the user.

The next chapter presents a baseline evaluation of the positional 3D audio controls supported by the mobile device of choice for this work, i.e. Nokia N95 8GB. Significant differences in the implementation of spatial audio amongst different mobile devices result in unknown levels of localisation accuracy and for this reason a calibration of the 3D audio system is a requirement before spatial auditory interfaces can be implemented and tested. In the remaining chapters of this thesis, how spatial auditory display design affects auditory multitasking and cognitive load will be investigated in a lab environment as well as in a more naturalistic application setting.

Chapter 4

Experimental Groundwork: Evaluation of the Spatial Localisation Accuracy on the Nokia N95

4.1 Introduction

Chapter 2 and Chapter 3 have discussed how spatial audio, which allows us to localise a sound source in a 3D space, can offer a means of altering focus between auditory streams as well as increasing the richness and differentiation of audio cues. However, the inclusion of spatial audio on mobile phones is a recent development and significant differences in the implementation of spatial audio can be found amongst different mobile devices resulting in unknown levels of localisation accuracy. Thus, a calibration of this new technology is a requirement for any further spatial audio research and will lay the necessary experimental groundwork required before attempting to answer the research questions posed by this thesis.

This chapter reports an evaluation of the JSR-234 Advanced Multimedia Supplements API (AMMS) 3D audio location controls on the Nokia N95 8GB (2009) in order to investigate what level of localisation accuracy listeners could achieve. The HRTFs and API of the AMMS 3D audio location controls were used to position sounds at arbitrary points around the user. Each participant was required to adjust an auditory pointer to the same direction of a static auditory source. This method helped determine to what extent listeners were able to discriminate the auditory sources as originating from different locations. In addition, differences between pink noise, speech and auditory dominance were also controlled for, as this could have a critical effect on spatial audio perception.

4.2 Experimental Study

4.2.1 Design of the Experiment

An auditory pointer adjustment program was developed using the 3D audio capabilities offered in the AMMS API. The methodology used in this study replicated the one from Pulkki and Hirvonen (2005) to evaluate an apparatus for auditory pointer adjustment and its localisation accuracy in an eight-channel and 5.1 loudspeaker setup. This method was used to test to what extent listeners were able to discriminate the auditory sources as originating from different locations. It has been found that humans generate errors and bias when interpreting auditory perception with any method (Blauert, 1997). However, when listeners are comparing two auditory perceptions, and adjusting the auditory pointer direction until there is no perceived difference in the direction between the pointer and the static sources, fewer errors and biases occur.

4.2.2 Evaluation Setup

Twelve listeners matched the auditory pointer direction with single static sources in directions [0° (directly in front of the nose), 45°, 90°, -45°, -90°] and elevation 0°. All static sources were placed in the front 180°, as it has been found to be the area of most accurate perception of direction (Marentakis and Brewster, 2005). All five directions used in this evaluation formed part of Pulkki and Hirvonen's study and so a comparison of the results from both studies will be possible. The experiment consisted of a training session followed by two different conditions. In one of the conditions, the static sources emitted pink-noise (a noise signal that contains all frequencies with equal energy per octave, commonly used to test loudspeakers (D'Appolito, 1998)). The pink-noise source was 500ms with a 50ms fade-in and fade-out. In the other condition, the static sources emitted recorded speech, using the phrase "One head-line in Britain today", taken from a BBC podcast. Speech will be one of the audio source types that will be used to present information in the following experimental studies reported in this thesis. The speech source was 1500ms long. Both pink-noise and speech static sources were mono, 16-bit and sampled at 16 kHz (see Appendix A 1), as required by the AMMS API. The order of the conditions was randomised per participant to control for ordering effects. The acoustic pointer was a source placed closer to the listener (85mm) and the static sources were placed further away (100mm), as in Pulkki and Hirvonen's study. The acoustic pointer was always identical to the corresponding static source per trial for the given condition, be it pink-noise or speech. A 250ms gap was inserted between the target sound and the pointer sound. The participants were able to move the pointer in 15° increments by using the left or right keys on a Nokia N95 8GB mobile phone (see Figure 4.1).

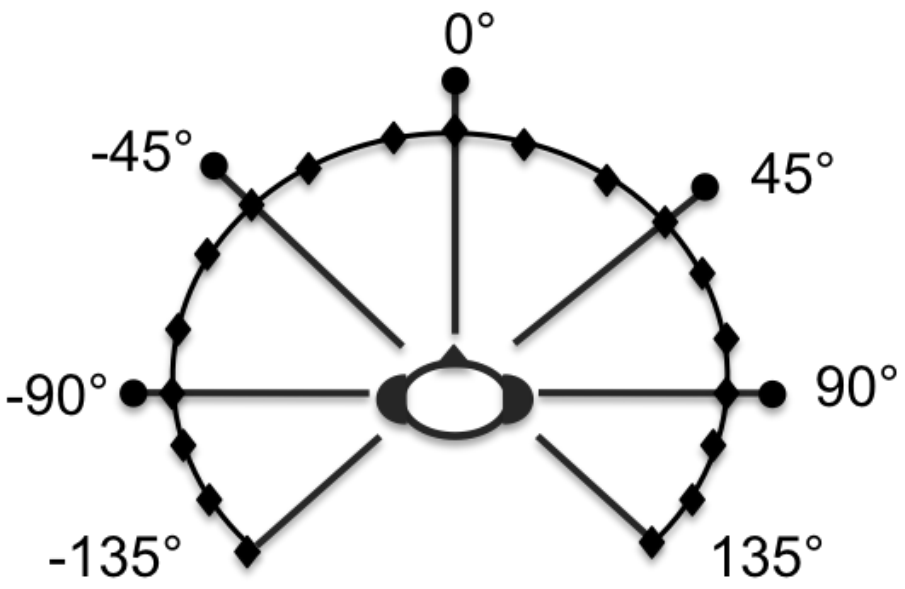


Figure 4.1: Evaluation setup. The black filled circles represent the different azimuth locations of the static sources placed at 100mm from the listener. The inner circle with diamonds shows the trajectory of the acoustic pointer placed at 85mm from the listener.

4.2.3 Evaluation Procedure

Participants were seated on a chair, holding the mobile phone in an upright position and wearing a pair of DT770 PRO 250 OHM Beyerdynamic headphones (see Figure 4.2). The participants were mostly students at Glasgow University, ten males and two females aged between 23 and 35 who were paid for their participation. All participants were informally asked to report their dominant hand and ear, either right or left, answering simple questions based on Handedness Earedness Questionnaires (2009): *Which hand do you prefer to use for writing? (right or left) When you receive a phone call. Which ear do you use to listen? which ear do you put the phone receiver next to? (right or left) The same when you can hardly hear something, which ear do you put forwards closer to the sound source? (right or left).* When asked about hand dominance, only one participant reported a dominant left hand. Only one participant reported a dominant left ear and another four a mixed right/left ear dominance. None of the participants was excluded based on handedness or earedness results and they all reported normal hearing.

The static source and the pointer signal were played once, one after another, every time a key was pressed to move the acoustic pointer left or right. Once the listener adjusted the pointer to the same direction as the static source, the central navigation key on the phone was pressed

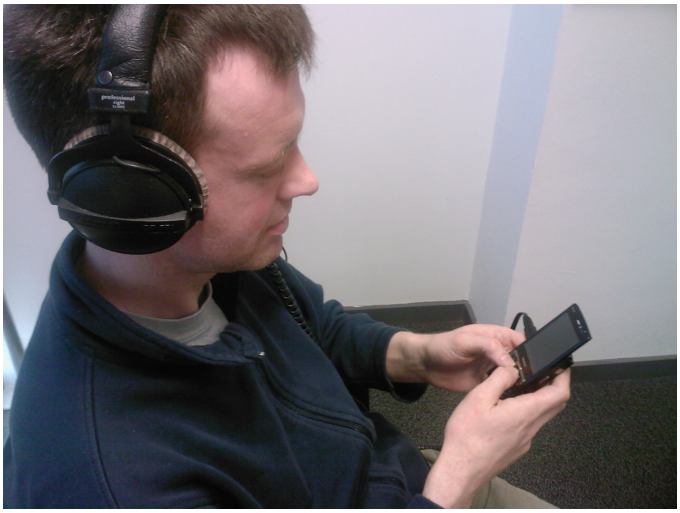


Figure 4.2: Experimental setup.

to indicate the adjustment was complete. After this, the location of the auditory pointer was recorded and a spoken prompt saying 'next' was played to introduce the next stimulus.

The test was organised so that both the pink-noise and the speech condition contained a total of 15 trials (five azimuth directions x 3 repetitions of each stimulus type) with 3 trials of each stimulus type in the training session. Each trial took approximately one minute. Sessions took less than 30 minutes in total and participants were allowed to rest between conditions. The trials were presented in randomised order for each session. The full set of instructions provided to participants can be found in Appendix B.

4.2.4 Results

The deviation of the acoustic pointer adjustment from the direction of the target source was recorded. A three-way between-subjects ANOVA was performed comparing the different static source azimuth directions, type of stimuli and earedness. The results showed a significant main effect for the different static source azimuth directions ($F(4,340) = 317.753$, $p<0.001$). *Post hoc* Tukey HSD comparisons indicated that static source azimuth direction -90° (mean=-81.00), -45° (mean=-51.55), 0° (mean=1.07), 45° (mean=53.00) and 90° (mean=85.71) were all perceived as being significantly different locations, ($p<0.001$). Figure 4.3 presents the acoustic pointer data across participants.

There was a main effect for the different stimuli type: speech and pink-noise ($F(1,340) = 4.065$, $p<0.05$) showing that participants were better at localising pink-noise than speech, especially

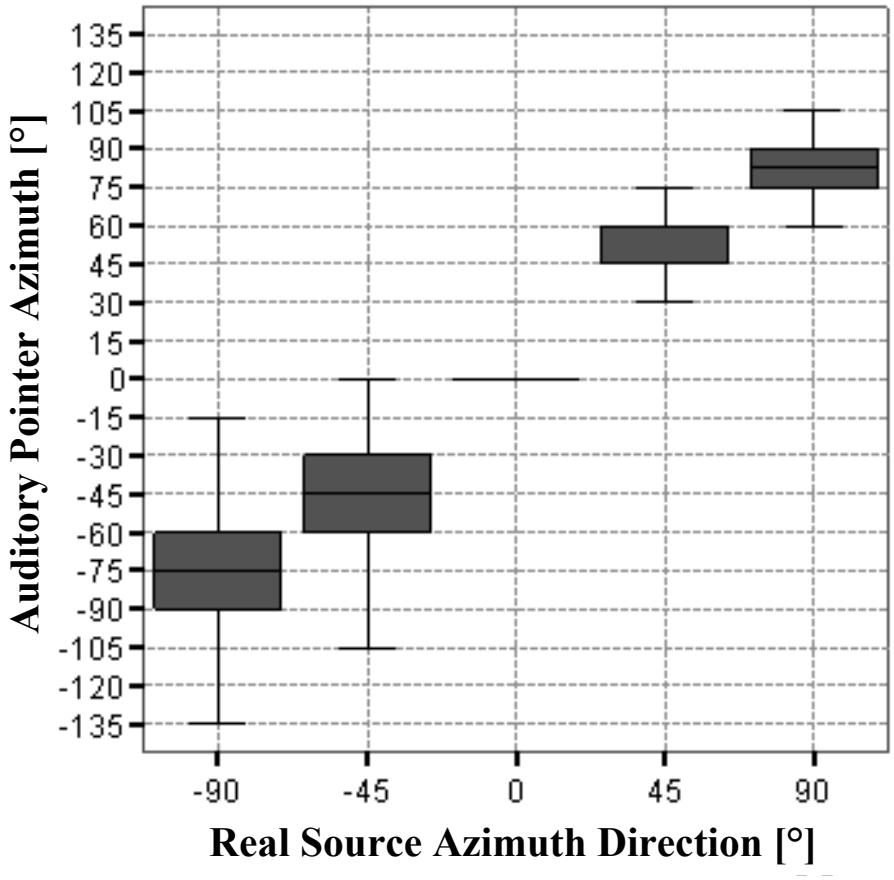


Figure 4.3: Box plots showing the accuracy of the method of adjustment applied in the evaluation study. The boxes contain the middle 50% of the data and the horizontal bold black lines show the median.

on the left side of 0°; earedness, i.e. right *versus* non-right ear dominance ($F(1,34) = 3.889$, $p<0.05$) showing that right-eared participants were more accurate than nonright-eared ones; and a two-way interaction between earedness and the different static source azimuth directions ($F(4,340)=5.469$, $p<0.001$). It could be concluded from these results that both ear dominance and the type of stimuli would be important factors influencing spatial audio localisation. However, the results from our only left-handed participant contained a high number of outliers. If this subject is removed earedness, stimuli type and the interaction, stop being significant. Without more data it is not possible to say if these results were caused by left-handedness alone.

Figure 4.4 shows the different signed error means by earedness grouped by left (azimuths -90°,

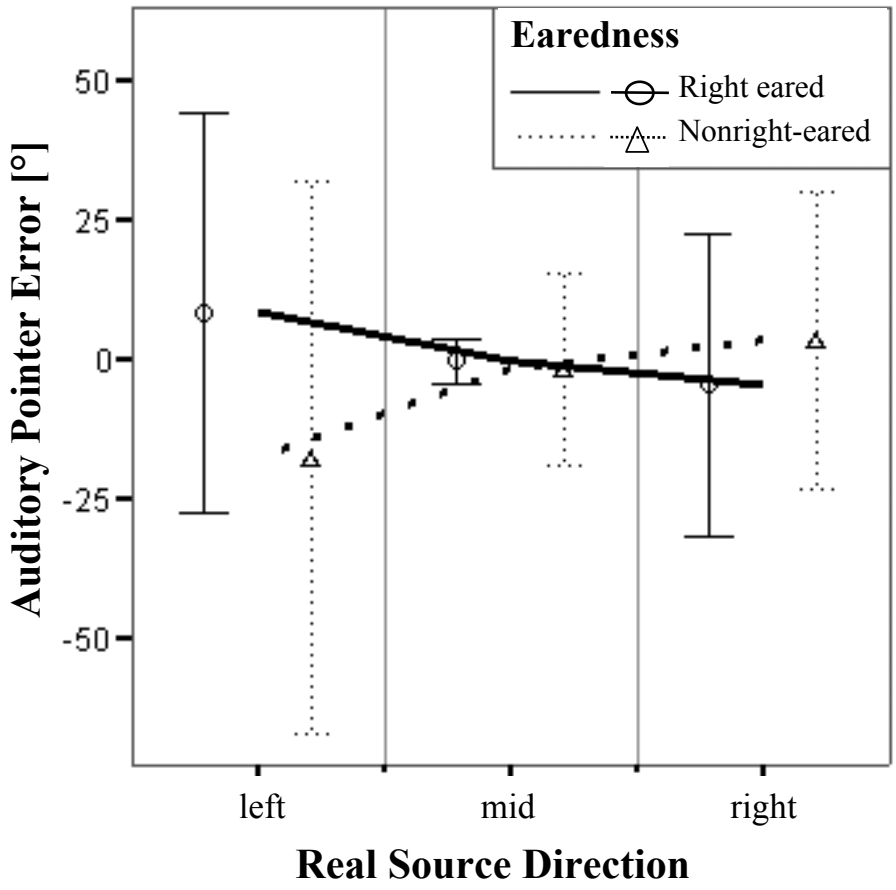


Figure 4.4: Signed error per static source direction and ear dominance (Right eared: N=7; Nonright eared: N=5). Error bars show ± 1.0 SD.

-45°), mid (azimuths 0°), left (azimuths 90°, 45°). The data suggest that right-eared participants tend to perceive sources as being more central than participants whose right ear is not dominant.

4.2.5 Discussion

The participants in this evaluation were successfully able to use the 3D audio system on the Nokia phone to identify unique targets at 45° intervals. As in Pulkki and Hirvonen's study, the deviation was considerably larger on the left than on the right side of azimuth 0°, but based on the results from this study it can be assumed that the AMMS 3D audio location controls will be appropriate for a 3D auditory interface. However, discriminative locations greater than five

seem unlikely unless head tracking is used to allow more 'active' listening. On this basis, the Nokia N95 will be used in the following experimental studies presented in this thesis.

Results concerning earedness were inconclusive but there is a suggestion that left or right ear dominance might affect the perception of the relative positioning of azimuths without affecting discriminative ability. The effect of centralising the sources for right-eared participants might be connected to the right hemisphere dominance in spatialisation (the right ear is more strongly connected to the left hemisphere). However, our single left-handed participant performed very differently from the rest but without more data it is not possible to say if this was caused by left-handedness alone. In the light of these results it was decided that, for the remaining experiments presented in this thesis, participants would be screened for hand dominance and only right-handed participants would be allowed to take part in the experiments.

4.3 Conclusions

This chapter presented the results of an evaluation of the Java AMMS 3D audio location controls supported on the Nokia N95 8GB mobile phone. Results showed that the spatial audio system on this device provided clear location discrimination for five sources at 45° intervals in the front 180°, so these location controls were considered appropriate for developing a 3D auditory interface. This suggests that the audio capabilities of mobile phones are now capable of running 3D auditory interfaces that were previously only possible on laptops, allowing the design and evaluation of more practical and effective mobile spatial audio interactions.

The next chapter presents a detailed research study, using the 3D audio controls evaluated in this chapter, which focuses on the interaction between cognitive load and eyes-free spatial audio interfaces supporting multiple auditory streams.

Chapter 5

Designing Spatial Auditory Interfaces for Eyes-Free Multitasking

5.1 Introduction

In the previous chapter, the 3D audio location controls supported on the Nokia N95 8GB were shown to be appropriate for developing and implementing spatial auditory interfaces on this platform.

In this thesis, it is hypothesised that 3D audio techniques offer a means of designing effective auditory interfaces to support eyes-free mobile multitasking. This is addressed in this chapter by investigating how to manage multiple auditory streams without overloading the user. The 3D audio controls evaluated in Chapter 4 are used to design and implement spatial and non-spatial auditory interfaces, which are then evaluated in an interactive multitasking environment under varying cognitive load. The outcome of this investigation provides further answers to the first research question posed by this thesis: "To what extent can 3D audio techniques aid the user to maintain coherent attention on multiple auditory streams in a mobile eyes-free interface?" (RQ 1).

This chapter starts by introducing the design of an audio minimisation technique implemented using spatial audio. Then, this technique is evaluated and its efficiency and usability is reported together with guidelines for designers building eyes-free auditory interfaces for mobile applications.

5.2 Audio Minimisation

The work by Ludwig *et al.* (1990) and Cohen and Ludwig (1991) in Audio Windows used the visual metaphor of a window-based graphical user interface in their spatial auditory display design. The work presented in this chapter extends this approach by considering the visual metaphor of *minimisation* in an auditory display.

In a visual display, minimisation has been used to present concurrent information. For example, current TV graphical interfaces deal with the issue of presenting concurrent visual streams by minimising the TV image when the user interacts with the television menu to change channels or just browse what is available in the different channels. In the same way, in a rich auditory interface we could minimise auditory streams when we are busy and need to focus on something else. We could also minimise the current sounds to allow interaction with the audio menus controlling our user interface.

Audio minimisation, as with minimisation in visual systems, could act as an important component in any audio interface. Audio minimisation could be achieved by reducing the loudness of an audio stream (similar to reducing the size of the TV image on the TV screen), and moving the perceived location of the audio stream from a central position (similar to moving the reduced TV image to the side of the TV screen). Once such minimisation has occurred, a second audio stream could be played at normal loudness with a perceived location directly in front of the listener (for example, a menu to control the interface). This audio minimisation technique presents a novel solution to the problem of presenting concurrent audio streams in a spatial auditory display.

However, using minimisation as a strategy for alternating focus between auditory streams in a multitasking environment will affect the attention demand required from the users (as discussed in Section 3.4.2). Previous research has not only looked into how well listeners are able to focus on one audio stream while ignoring the others (a situation known as selective attention, see Shinn-Cunningham and Ihlefeld (2004)), but also on how well they are able to understand the content of multiple, simultaneous sources (a situation requiring divided attention, see Yost *et al.* (1996) and Brungart *et al.* (2001)). However, it is unclear what the most useful implementation of audio minimisation might be when supporting multiple competing sources and whether 3D audio techniques are effective enough to support audio minimisation.

5.3 Experimental Study

The experimental study presented in this chapter quantifies the effect of cognitive load on a novel audio minimisation technique implemented using 3D audio (this technique will be referred to as *spatial minimisation*) and a number of non-spatial audio techniques during selective- and divided-attention tasks involving user interaction.

5.3.1 Design of the Experiment

Participants

Forty-eight participants (26 males, 22 females, aged 18 to 53 years) were recruited. All were native speakers of British English, reported normal hearing and were right-handed. Participants were split equally into two groups: divided- and selective-attention in a between-subjects design.

Stimuli

Participants listened to two different streams: one continuous and the other user activated. In the selective-attention group, the continuous stream was a piece of classical music taken from Mozart's Sonata for two pianos K448 in D Major. This specific music piece has been frequently used in spatial-temporal reasoning research (Rauscher *et al.*, 1993). The sonata was divided into different fragments: one for the training session and four others were used in the four different conditions. These fragments were all mono, 16-bit and sampled at 16kHz, and approximately 1.5 minutes long (see Appendix A 2). The participants were told they would have to answer a question on the audio menu tasks to ensure selective attention. In the divided-attention group, the continuous stream was a podcast selected from the BBC Radio 4 programme 'From our own correspondent'. Five different podcasts with a similar journalistic format were chosen. One podcast was used for training the participants and the rest were used in four different test conditions. They were all mono, 16-bit and sampled at 16kHz, and narrated by a male speaker (see Appendix A 2). In order to ensure divided attention, participants were asked to monitor the podcast and told they would have to answer questions on content *as well* as a question on the audio menu tasks. In order to retain coherence, and to allow enough audio material to pose content questions before, during and after the audio menu tasks, the podcasts were longer than the classical musical streams (3 minutes). Our aim was to generate a low cognitive load for the selective-attention group by using classical music, and a high cognitive load in the divided-attention group by using speech.

The user-activated audio stream was a hierarchical audio menu with synthesised audio items. It consisted of a three-item top level: music, appointments and current time. The 'music' item included three items in two different sub-levels: 1) previous track, current track and next track, 2) the song title for each of the items. The 'appointments' item included three items in three different sub-levels: 1) Monday, Tuesday and Wednesday, 2) the times for the appointments, 3) appointment information. The 'current time' item only had one sub-level with time information, (for more details see Figure 5.1). The song titles, appointment information and current time were different for the different conditions. The audio menu items were synthesised using Cereproc's (Aylett and Pidcock, 2007) British English female RP voice. All the audio items were mono, 16-bit, sampled at 16kHz (see Appendix A 2). The Amplify filter in Audacity (2010) was used to normalise the volume of both the continuous and user-activated streams were to 70% of the audio dynamic range, which equals to a normal conversation typically 60-70dB (Kryter, 1972).

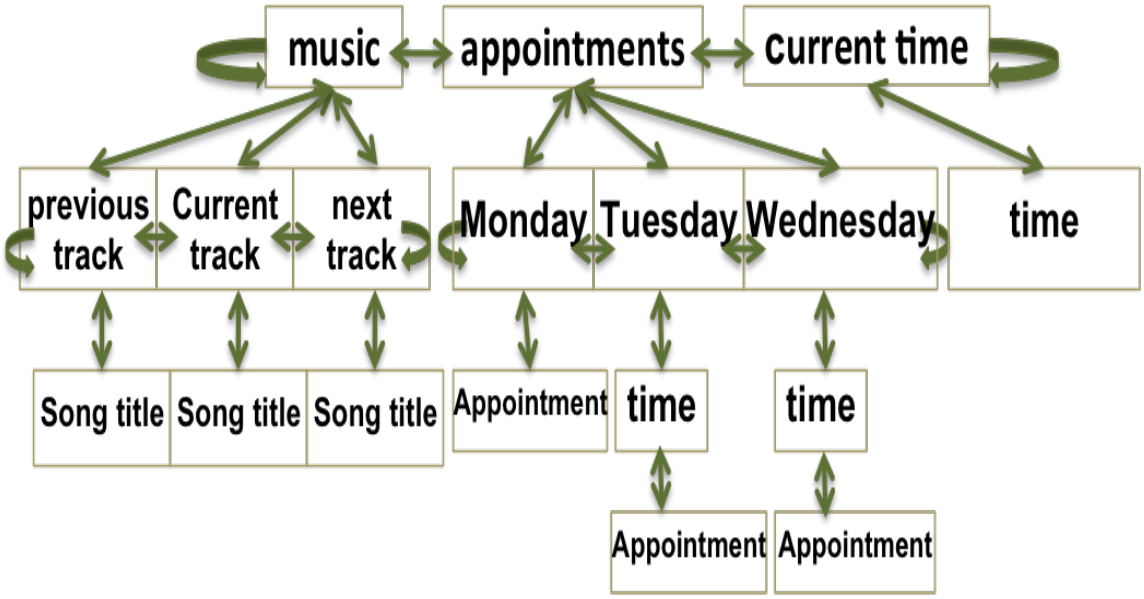


Figure 5.1: Audio-menu structure.

Procedure

Each group was tested in four different conditions:

1. *Baseline*: The continuous stream was paused or interrupted while the participant carried out the audio menu tasks and then resumed after the tasks were completed. Both the continuous stream and the audio menu were located at the origin (0° azimuth) and at a distance of 1m in the frontal horizontal plane.

2. *Concurrent*: The continuous stream played while the participant carried out the audio menu tasks. Podcast and audio menu were located at the origin (0° azimuth) and at a distance of 1m in the frontal horizontal plane.
3. *User-activated spatial minimisation*: The continuous stream was located at the origin (0° azimuth) 1m away from the listener in the frontal horizontal plane (see Figure 5.2a), and moved to the right hand-side (90° azimuth) only when the participant was engaged in the audio menu tasks (see Figure 5.2b). The decision to move the podcast from the origin to the right hand-side was based on the evaluation results presented previously in Chapter 4. This specific location showed less variation in the localisation perception by listeners. The volume level of the podcast was attenuated by approximately -10dB by moving the source to the right hand-side (-3dB intensity drop) and doubling the perceived distance by placing it 2m away from the listener. Listeners have been shown to perform best when monitoring an audio stream at -10dB, compared to lower levels (Ihlefeld and Shinn-Cunningham, 2008).
4. *Fixed spatial minimisation*: The continuous stream was fixed to 90° azimuth and 2m away from the listener for the entire duration of this condition. The audio menu was located at 0° azimuth. Both streams were presented concurrently.

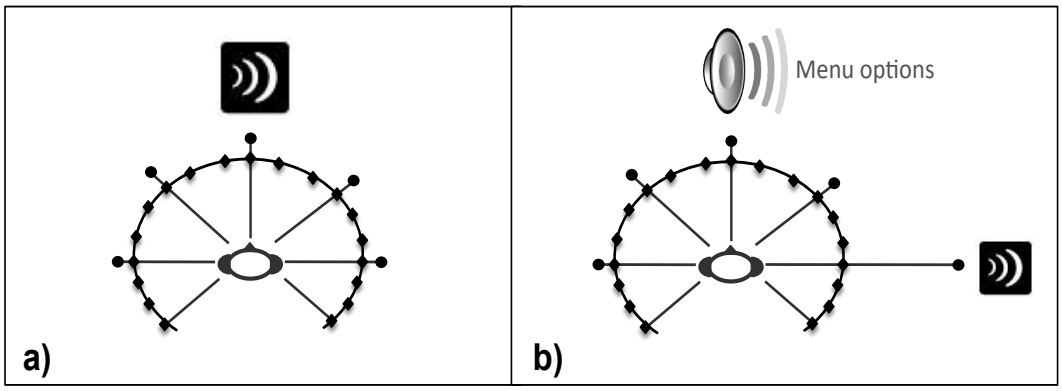


Figure 5.2: (a) Single continuous stream. Black filled circles show different azimuth locations. (b) User-activated spatial minimisation: stream moved from front to right.

In all four conditions, participants performed three tasks using a hierarchical audio menu:

- Finding the next music track title.
- Checking an appointment for Tuesday.
- Finding the current time.

The audio menu was always presented at 0° azimuth (in front of the user's nose) and always 1m away in the frontal horizontal plane. All four conditions were tested in a randomised order in both the divided- and selective-attention groups. Participants were tested in a static lab environment seated on a chair holding the mobile phone in an upright position wearing a pair of headphones (see Figure 4.2). The study was run on a Nokia N95 8GB using the built-in HRTFs and the JAVA JSR-234 Advanced Multimedia Supplements API to position the auditory sources. The audio was played over a pair of DT770 PRO 250 OHM Beyerdynamic headphones.

Participants completed two training sessions before the test conditions. First, a training session was devoted to familiarising the participants with the audio menu structure in their own time. The second training session used the concurrent condition to familiarise the participants with listening to a continuous audio stream while interacting with the audio menu. For each test condition, participants listened to a continuous audio stream and after approximately 1 minute, the user was prompted with a 25 ms sine wave beep at 1500Hz to start interacting with the menu and complete the three tasks described previously in any order. To initiate this interaction, the participant pressed the central navigation key on the phone. The arrow keys on the phone were used to browse the menu items. Once the tasks were completed and the audio menu was exited by pressing the central navigation key again, the user continued listening to the continuous audio stream until it was over. Participants in the divided-attention group were instructed to monitor the continuous podcast (the instructions provided to participants can be found in Appendix C.1.1). After the end of the podcast, for each condition, participants were asked to answer a set of six questions as in Stifelman's study (Stifelman, 1994). Five of the questions requested information that was located at evenly spaced points in time over the length of the podcast to confirm the participant had paid attention to it. The last question requested information about one of the menu tasks. The full list of questions per condition can be found in Appendix C.1.2. Participants in the selective-attention group were only required to recall information about one of the menu tasks as there was no content to be recalled from the classical music piece (the instructions for participants can be found in Appendix C.2.1 and the recall question per condition in Appendix C.2.2). Following the recall questions, participants were asked to complete a NASA-TLX subjective workload assessment (Hart and Staveland, 1988). NASA-TLX is a well validated multi-dimensional rating scale designed to obtain workload estimates from one or more operators while they are performing a task or immediately afterwards (see Appendix D for a sample of the NASA-TLX form provided to both the divided- and selective-attention groups). After all four conditions were completed, participants were instructed to rank them in order of preference: '1' being most preferred and '4' the least (see Appendix C.1.3 for the preference form provided to the divided-attention group and Appendix C.2.3 for the preference

form provided to the selective-attention group). The experiment took 30-45 minutes in total. Participants were allowed to rest between conditions.

Metrics

In this evaluation user preference and workload metrics together with performance indicators were combined to assess the effectiveness and usability of a spatial minimisation technique and a number of non-spatial audio techniques. The independent variable (IV) was the type of condition (the *Baseline* condition, the *Concurrent* condition, the *User-activated spatial minimisation* condition and the *Fixed spatial minimisation* condition) per attention group (divided-attention and selective-attention), and the dependent variables (DVs) were a combination of subjective (user preference and perceived subjective workload) and objective measures (recalled information and time taken to complete the audio menu tasks).

5.3.2 Results

Ranked Preferences

Figure 5.3 shows a stacked count for the order of preference for the four conditions compared in the divided- and the selective-attention groups. A non-parametric Kruskal-Wallis test (Kruskal and Wallis, 1952) for different conditions per attention group showed there was a significant difference in the medians ($\chi 2$ (7, N=192)=61.810, p<0.001). Mann-Whitney tests (Mann and Whitney, 1947) for independent samples with Bonferroni correction showed a significant difference between the interrupted conditions by group, and also between the user-activated spatial minimisation conditions by group. Users' preference for interrupting the continuous stream significantly decreased (two-tailed p<0.0001, total 1st preferences dropped from 20 to 4) and preference for spatially minimising the continuous stream significantly increased (two-tailed p =0.008, total 1st preferences increased from 2 to 8) when the streamed source was classical music.

Overall Workload

Raw overall workload was calculated from the NASA-TLX questionnaires completed after each condition. A repeated-measures ANOVA with condition type as a within-subjects factor and attention group as a between-subjects factor showed a significant main effect for condition type

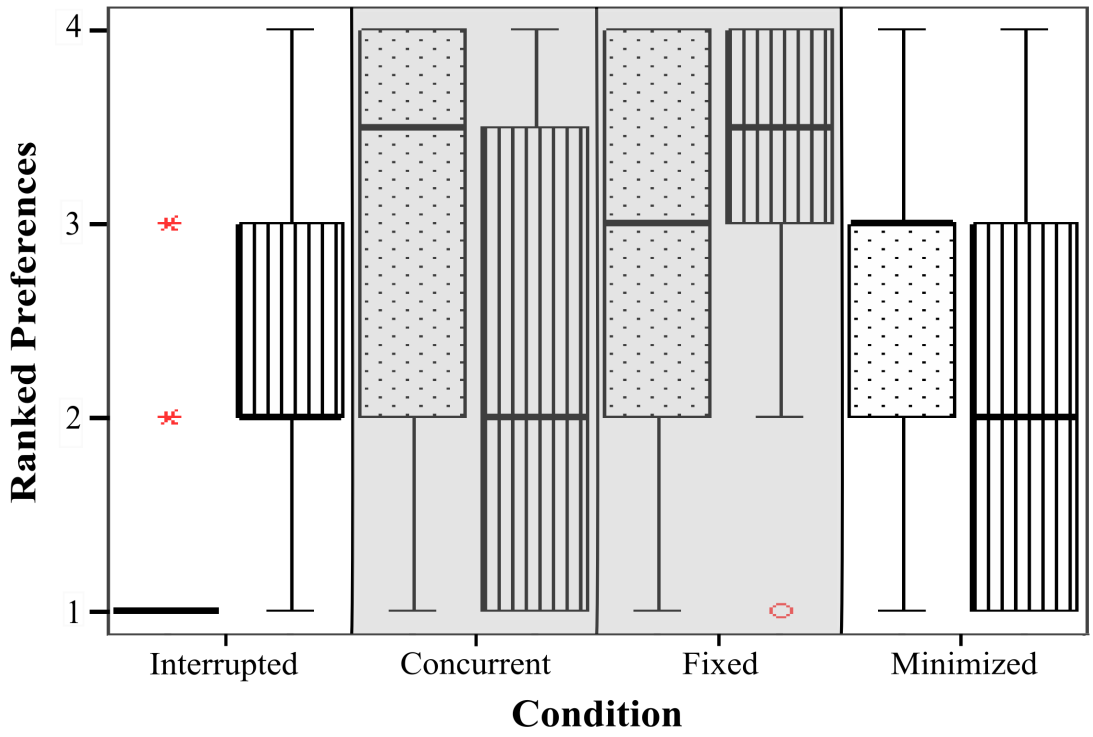


Figure 5.3: Box plots present ranked preferences per condition and attention group: divided-attention (dotted) and selective-attention (striped). The boxes contain the middle 50% of the data, the horizontal bold black lines show the median and the red points outside are suspected outliers. The grey shaded conditions showed no significance.

($F(3,96)=9.786$, $p<0.001$) and attention type ($F(1,32)= 48.284$, $p<0.001$). There was also an interaction between attention and condition type ($F(3,96)= 4.34$, $p<0.01$). As expected, perceived overall workload was significantly higher in the divided-attention group (mean=51.71) than in the selective-attention group (mean=21.33). *Post hoc* Pairwise Comparisons with Bonferroni correction for condition type showed that perceived overall workload during the interrupted condition was significantly lower ($p<0.015$) than in the rest of conditions for the divided-attention group. No significant differences were found between conditions for the selective-attention group (see Figure 5.4).

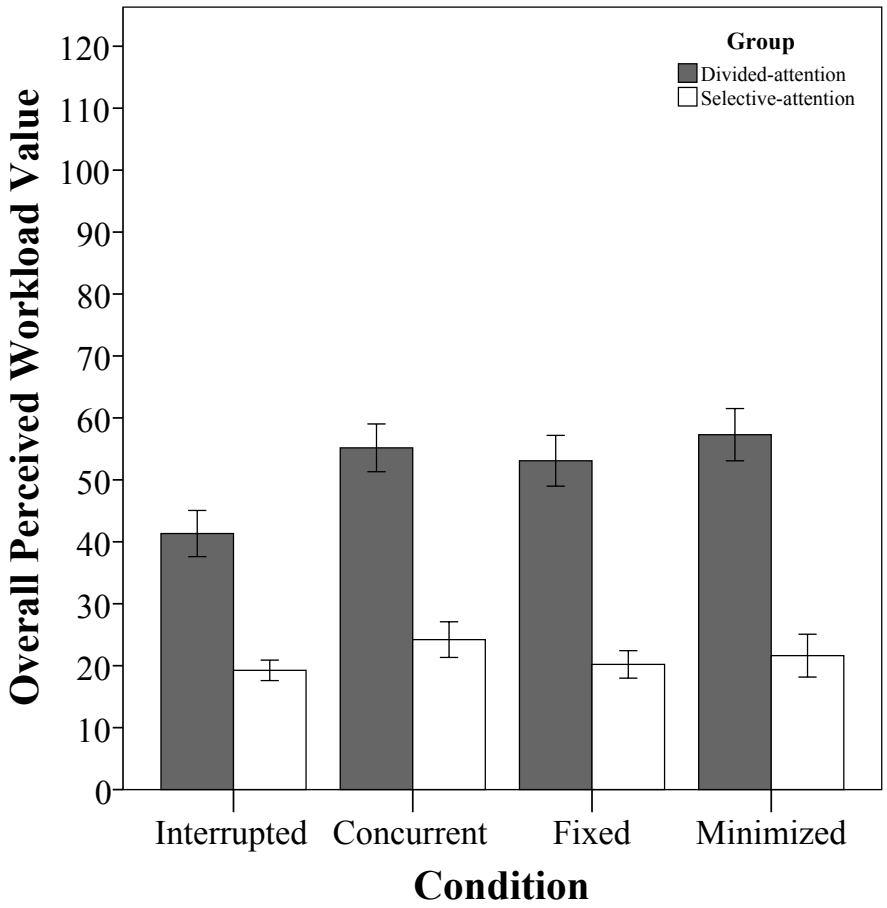


Figure 5.4: Overall perceived workload per condition and attention group. Error bars show Standard Error of Mean ± 1.0.

Performance

Recall performance Recall performance was calculated using the percentage correct of answers in each condition per different attention group. These results are presented per attention group first given that the number of recall questions for the divided-attention group was six (see Figure 5.5), whereas for the selective-attention group was only one (see Figure 5.6). Condition type per attention group is treated as a within-subjects factor. Then, results on the recall question of the menu task alone is presented across attention groups. The menu task question was the only one shared by both attention groups.

A repeated-measures ANOVA on recall performance means with condition type as a within-subjects factor showed a significant main effect for condition type ($F(3,48)=5.109$, $p<0.010$) in

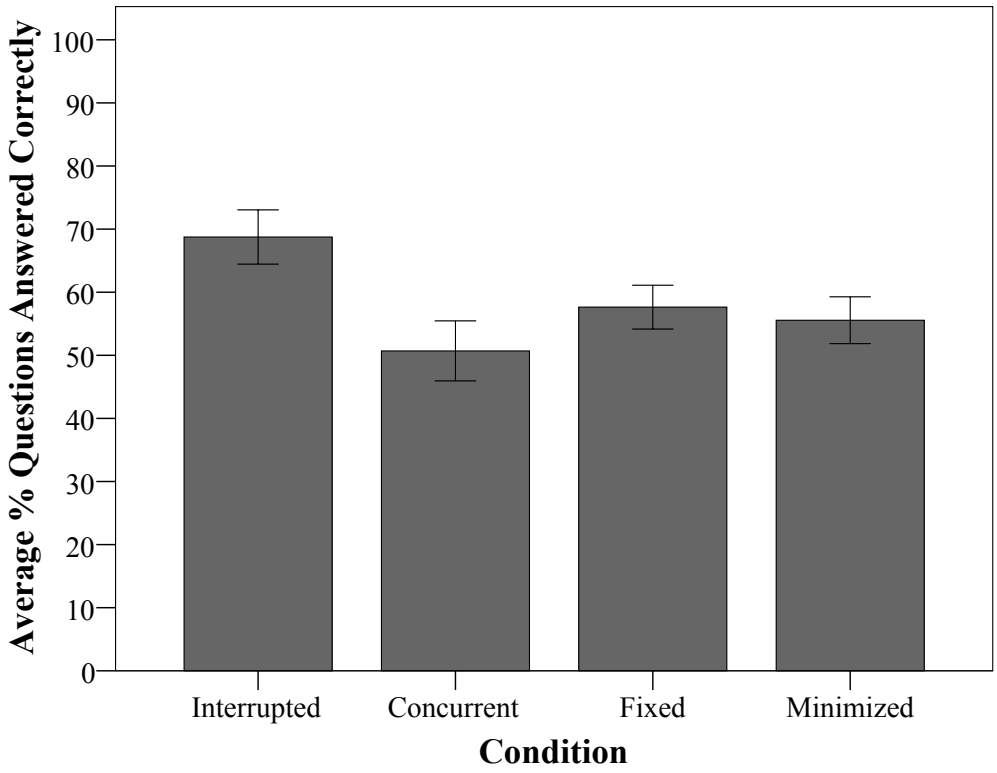


Figure 5.5: Average percentage of correct answers for the divided-attention group (Number of recall questions (N) = 6). Error bars show Standard Error of Mean ± 1.0.

the divided-attention group. *Post hoc* Tukey HSD tests with Bonferroni correction for condition type in the divided-attention group showed that the spatially fixed ($p<0.050$) and concurrent conditions ($p<0.050$) showed a significant performance drop caused by the cognitive load (from 70% recall in the interrupted condition to 50% recall in the spatially fixed and concurrent conditions). There was no significant effect in the selective-attention group for condition type.

A non-parametric Kruskal-Wallis test (Kruskal and Wallis, 1952) on the recall of the menu task showed there was a significant main effect per attention group ($\chi 2$ (1, N=192)= 4.159, $p<0.05$). Total recall of the menu task across conditions was significantly lower for the divided-attention group (80% (SD=6.25%)), due to higher cognitive load, than for the selective-attention group (91% (SD=5.24%)). There were no interactions. In addition, a non-parametric Friedman test (Friedman, 1937) for related samples showed no significant main effect for condition type for either attention group.

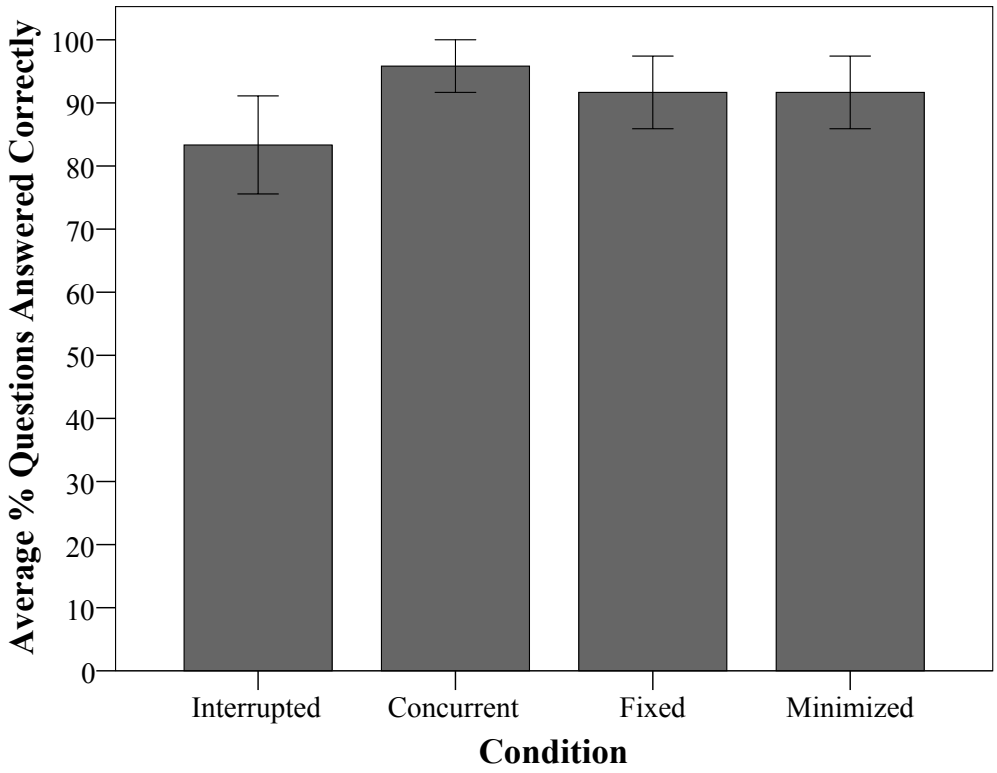


Figure 5.6: Average percentage of correct answers for the selective-attention group (Number of recall questions (N) = 1). Error bars show Standard Error of Mean ± 1.0.

Task completion times Total time taken to complete the audio menu tasks was also computed (see Figure 5.7). A repeated-measures ANOVA with condition type as a within-subjects factor and attention group as a between-subjects factor showed a significant main effect for condition type ($F(3,96)=5.45$, $p<0.005$) and attention type ($F(1,32)=7.21$, $p<0.015$). There was also an interaction between attention an condition type ($F(3,96)= 2.89$, $p<0.050$). Task completion times were significantly higher (mean= 41.76 secs) for the divided-attention group than for the selective-attention group (mean=32.03 secs). *Post hoc* Pairwise Comparisons with Bonferroni correction for condition type showed that task completion times for the interrupted condition were significantly lower ($p<0.05$). Also, *Post hoc* Independent samples t-tests across attention type showed that task completion times were significantly higher for the concurrent condition ($t(46) =2.640$, $p<0.050$) (mean=43.95 secs) and the minimised condition ($t(46)=2.73$, $p<0.05$) (mean= 47.42 secs) for the divided-attention group than for the same conditions for the selective-attention group (concurrent: mean=32.31 secs; minimised: mean=33.53 secs).

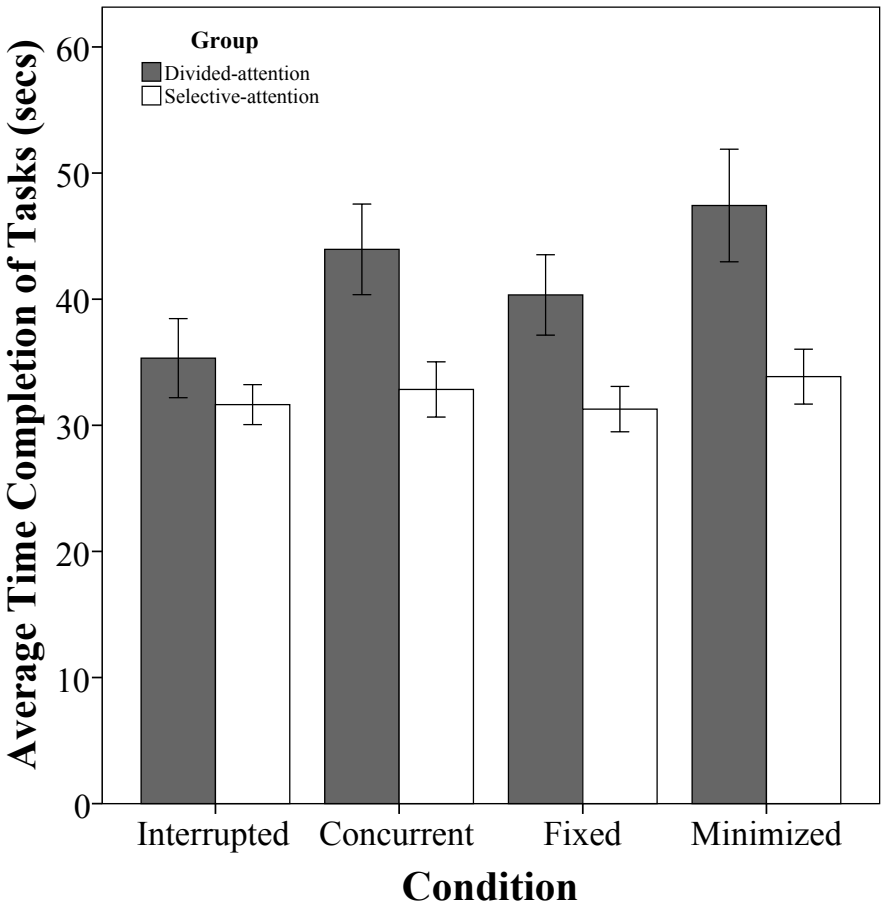


Figure 5.7: Mean task completion times per condition and attention group. Error bars show Standard Error of Mean ± 1.0.

5.3.3 Discussion

As we might expect, users do not like being put under cognitive load. Participants in this study reported higher workload and took longer to carry out tasks in the divided-attention group. Furthermore, listening to concurrent audio increased the effect of the load. In this situation, one could expect a drop of performance (in this study from 70% to 50% recall and an increase in task time from 35.32 to 47.43 secs). The use of spatial techniques had a negligible impact on reducing this effect (although there was a tendency to prefer minimisation compared to the other simultaneous conditions). There is a tendency for spatial audio to improve recall against single point presentation, for the user-activated spatial minimisation to increase workload (possibly because participants found the movement distracting “and felt it was more of an adjustment to

pay attention to the continuous stream when it moved"), and a trend for participants preferring the spatially fixed and minimised conditions over the other simultaneous condition. Overall, interrupting the continuous stream was significantly preferred over the simultaneous conditions but, even under extreme cognitive load, participants were able to carry out the tasks although at a reduced ability.

In the selective-attention group however, preference results were significantly reversed for interrupted and minimisation conditions. This shows that users disliked having the continuous stream interrupted when not under load, and using spatial techniques to separate the continuous stream from the user-activated menu were preferred.

5.4 Conclusions

In this chapter, a number of techniques for simultaneously presenting multiple auditory streams when mobile were implemented that included spatial minimisation. These techniques have been suggested in previous research but have never been evaluated formally against each other. The efficiency and usability of these techniques was evaluated and reported, which showed that the spatial minimisation technique offers an effective means of presenting and interacting with multiple auditory streams simultaneously in a selective-attention scenario. However, spatialisation techniques were not as effective in the divided-attention task, in which the interaction benefited significantly from the interruption of the continuous stream. In a mobile spatial audio interface design in which interrupting the streams is not an option, it should be noted that, even in this extreme cognitive load scenario, participants were able to carry out the tasks, although at a reduced ability. As such, this chapter proposed a 3D audio technique that enabled users to maintain coherent attention on multiple auditory streams in a mobile interface thus informing the first research question: " *To what extent can 3D audio techniques aid the user to maintain coherent attention on multiple auditory streams in a mobile eyes-free interface?*".

The results presented in Section 5.3.2 suggest a mixed design of audio techniques would be required when designing eyes-free auditory interfaces. When in a selective-attention task concurrent streams should be used with spatial audio being used to help separate the information streams. Also, spatial minimisation could be used as this will not disrupt his task efficiency. However, if the decrease in performance is unacceptable (and the interaction allows it), the recommendation would be to avoid the simultaneous presentation of auditory streams. It would be nonetheless important to allow users to interrupt their eyes-free interaction with a mobile device at anytime while navigating a space.

The next chapter reports an initial case study in which a number of eyes-free mobile auditory interfaces, including spatial auditory interfaces, are designed to support multiple presentation of location-based information. This study presents an 'in the wild' evaluation of the proposed auditory interfaces and the user experience they delivered in a mobile audio-augmented reality environment.

Chapter 6

Case Study: Location-based Information in a Mobile Audio-Augmented Reality Environment

6.1 Introduction

Chapter 5 described the design, implementation and evaluation of a number of auditory interfaces, i.e. interrupted, concurrent, fixed, minimised, supporting eyes-free mobile multitasking in an interactive environment under varying cognitive load. Results showed that, given an appropriate task structure, 3D audio techniques offer a means of designing effective auditory interfaces, to support eyes-free mobile multitasking.

The spatial auditory displays tested in Chapter 5 exhibited an egocentric design. In other words, the multiple information streams supported in the auditory displays were always placed in a fixed position relative to the user. This design is ideal for user interactions when on the go as changes in orientation have no effect on the position of the information streams around the user, making them easy to remember. An exocentric spatial auditory display, where sounds are placed relative to the real world, offer an alternative approach to supporting multiple information streams in a mobile environment. Exocentric displays are particularly suited for augmenting spaces, where audio is situated in the real world (see Section 3.3.2). However, unlike egocentric displays, exocentric displays present an additional technical challenge as the head orientation of the user needs to be tracked in real time in order to update locations in the physical space. Furthermore, if multiple information streams are presented in such environments, as with egocentric designs,

the effects of cognitive load on the user need to be taken into account.

Exocentric spatial auditory displays for mobile audio-augmented environments can be designed to support navigational tasks but also more exploratory or wandering situations. A navigational system can be assessed by the user's success or failure at reaching a navigational goal, but this can also result in a system which prioritises technology and efficiency over the exploratory and playful nature of the user experience (McCarthy and Wright, 2004; Morrison *et al.*, 2007). On the other hand, evaluating *exploratory behaviour* in an audio-augmented environment presents challenges due to the implicit open-ended nature of exploration. In this thesis, the work presented in this chapter and Chapter 7 will use an exploratory mobile audio-augmented environment to test a number of different design choices for implementing spatial auditory displays that support multiple auditory streams. It was felt that such a design made less prior assumptions concerning user behaviour, allowing users more freedom in their interaction with the auditory displays.

This chapter reports a case study comparing four different auditory interfaces varying in the use of non-speech audio (Earcons, as discussed in 3.4) to advertise audio-augmented locations and 3D audio spatialisation in an interactive and exploratory *exocentric* eyes-free mobile environment. This study was designed to contribute towards answering the second research question posed by this thesis: "How can 3D audio techniques be used to disambiguate multiple auditory sources in order to access location-based information in a mobile eyes-free interface?" (RQ 2).

In order to answer Research Question 2, a quantitative and qualitative analysis of user exploration and interaction strategies from this initial case study was carried out. Although this case study included a limited user sample, it validated the technology used, provided a valuable framework for evaluation and offered an insight into user behaviour in mobile audio-augmented environments. Due to the small user sample (N= 8), results from this case study are not conclusive. A small sample size will tend to be unduly affected by outliers and care must be taken in generalising significant results across users. However, this case study forms a key initial investigation into the design of auditory interfaces able to support multiple location-based information streams in a purely exploratory audio-augmented reality environment. The research work reported in Chapter 7 takes this investigation further with a much larger sample size.

6.2 Experimental Study

The experimental case study presented in this chapter carried out an initial investigation into the effect of non-speech audio (Earcons), spatial auditory feedback and concurrent presentation

of multiple location-based information on the user experience in an interactive and exploratory mobile audio-augmented reality environment. A sound garden (see Section 3.4.3) was the setting for this study. This is a space that consists of a virtual audio environment superimposed on a *real* urban park featuring a set of precisely situated sounds surrounding the user. Such a space is dedicated to encourage visitors to explore and casually experience the space around. For this reason, a sound garden is an ideal space to investigate exploratory user behaviour.

6.2.1 Sound Garden Implementation

The case study presented in this chapter took place in a sound garden set in the Municipal Gardens in Funchal, Madeira. The sound garden ran on a Nokia N95 8GB mobile phone using software adapted from the Mobile Trail Explorer (Mobile Trail Explorer, 2010) application together with the HRTFs in the JAVA JSR-234 Advanced Multimedia Supplements API to position the auditory sources. The location of the user was determined using an external Qstarz BT-Q1000X Travel Recorder GPS receiver (Qstarz GPS Receiver, 2010) connected to the mobile phone via Bluetooth. The head orientation (compass heading) of the user was determined using a JAKE Sensor Pack (2010) also connected via Bluetooth. No pre-determined route or visual aids were provided, but users held the N95 in their hands in order to press keys and make system input. They listened to the sounds planted in the garden using a pair of Beyerdynamic DT231 headphones. The GPS receiver was placed on the headphone's left ear-cup and the JAKE on the crown of the head, in the middle of the headphone's headband. Both sensors were mounted using Velcro tape. Figure 6.1 shows the final system setup.

User Location Tracking Reliability

Location inaccuracy is always a concern in studies relying on GPS user tracking. Therefore, it was ensured that, at all times the GPS data were as accurate as possible. In the design phase of the sound garden it was noted that the sensitivity and reliability of the in-built GPS receiver on the Nokia N95-8GB were not good enough for the requirements of this study, at least on the island of Madeira. Hence, the Qstarz BT-Q1000X external GPS receiver was tested and found more reliable and consistent for the purpose of this study. Also, before the start of each trial and as a training exercise for each participant, the GPS accuracy was checked by asking users to find a virtual audio landmark situated outside the park. The application running the sound garden logged the GPS signal accuracy and printed it to the screen so the experimenter could confirm the GPS signal was good enough before asking the user to enter the park and start the experiment. During each trial, the experimenter closely shadowed the participant at all times.

Figure 6.1: Experimental setup. 1) JAKE sensor, 2) GPS receiver (both mounted on headphones) and 3) mobile device.

As all participants had been instructed beforehand to 'think aloud' while they walked through the park, the experimenter was able to detect whether the GPS had stopped tracking the user location. The GPS resolution proved to be sufficient as participants were demonstrably able to find the virtual audio landmarks. However, if at any point the GPS stopped updating and it was not recoverable, the experimenter made a note of it, restarted the application, the participant was asked to go back to the last landmark they had successfully discovered and the data were discarded from the analysis. After the study was completed and while analysing the GPS data, all the trajectories recorded for each participant were plotted and confirmed GPS tracking reliability.

Audio Content and System Configuration

Five different Earcons in the form of recordings of animal sounds (an owl, goose, cricket, nightingale and frog) were created to alert the user of the presence of five physical landmarks: the Rua Sao Francisco; a Coat of arms of the Saint Francis convent; the Statue of Joao Reis Gomes; the café and the pond. An illustrative map of the garden is shown in Figure 6.2. Animal sounds were used to identify landmarks because they seemed a good fit to the natural environment. Otherwise, the mapping between sounds and landmarks was abstract and symbolic; there was no pre-existing relationship between the sounds and the information they were representing. Furthermore, for each landmark brief speech audio clips were synthesised using Cereproc's

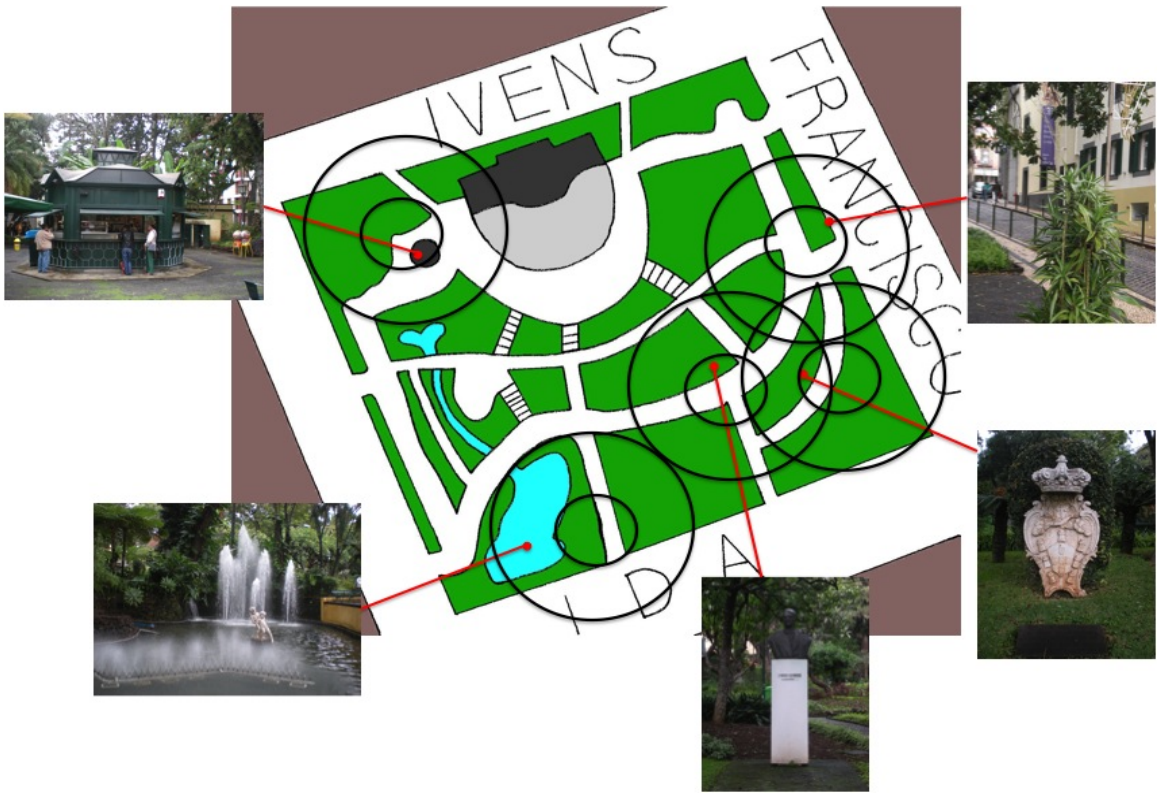


Figure 6.2: Municipal Gardens in Funchal, Madeira. Still images of the landmarks and illustration of proximity and activation zone per landmark.

British English male RP voice. These clips provided basic factual information about the sites. Synthesis made the setup of the sound garden easier by offering consistent and well-enunciated recorded speech without the need for a voice talent and a studio. Both the animal sounds and the audio clips were mono, 16-bit and sampled at 16 kHz (see Appendix A 3). They were adjusted to a conversational volume (approx. 60-70dB). Two circular zones surrounded each landmark: activation (radius 10m) and proximity (radius 25m) zones, in which different auditory feedback could be enabled. Due to the size of the garden (82m x 109m), only three landmarks had overlapping proximity zones while the other two were isolated. Figure 6.2 shows the audio landmark configuration.

6.2.2 Design of the Experiment

Conditions

In order to contribute to the research question addressed in this chapter, an evaluation was carried out to evaluate the absence or presence of the following auditory display features: non-speech sounds (Earcons), proximity zone and spatial 3D audio. Some combinations of these features are inappropriate. Without a proximity zone spatial 3D audio cannot be used, as there would be no area for spatialisation. Given that the investigation of overlapping proximity zones is central to this study and the work from Chapter 5 has shown that users can find concurrent speech streams frustrating and difficult to understand, Earcons are a requirement for these conditions. These restrictions result in four separate conditions, which vary in their complexity (see Table 6.1 for a summary):

	Earcons	Proximity zone	Spatial 3D audio
Baseline	✘	✘	✘
Earcons	✔	✘	✘
Spatial	✔	✔	✘
Spatial3D	✔	✔	✔

Table 6.1: Summary of auditory display features per condition.

1. *Baseline*. No Earcons or auditory spatialisation: When the user entered the activation zone, only the audio clip with information corresponding to that landmark was triggered and played once. The proximity zone was not used.
2. *Earcons*. Earcons but no auditory spatialisation: Whilst the user was within the activation zone, the Earcon (animal sound) corresponding to that location played continuously. The audio clip containing information about the location could be played (and the animal sound stopped) by pressing the central navigation button on the mobile phone. The proximity zone was not used.
3. *Spatial*. Basic proximity zone with Earcons and limited auditory spatialisation (distance): When the user entered the proximity zone, the Earcon, i.e. animal sound, corresponding to the location was triggered to alert the user of its presence (see Figure 6.3). The animal sound increased in loudness as the user walked closer to the physical landmark. The original sound level of the animal sound (60-70dB) dropped normally over distance (approx.

6dB per doubling of the distance to the sound source) making the quietest sound at the edge of the proximity zone 36dB. Once the user entered the activation zone, the audio clip could be played (and the animal sound stopped) by pressing the central navigation button on the mobile phone.

4. *Spatial3D*. Earcons and auditory spatialisation: Behaviour similar to Condition 3, with the difference that the animal sounds in the proximity zone were played using full spatialisation (in an exocentric spatial auditory display), varying not only in amplitude but also by direction of the sources.

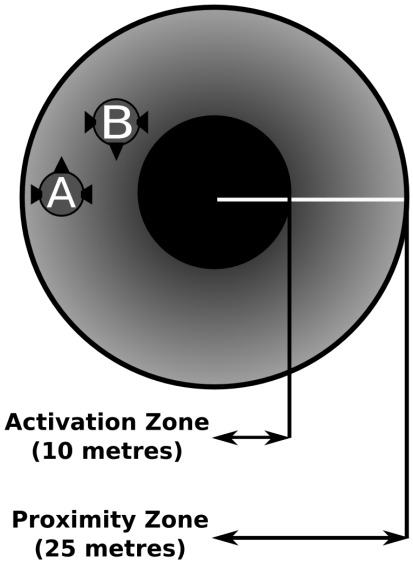


Figure 6.3: Audio landmark - gradient indicates volume. In the Spatial3D condition, User A (looking up in figure) hears a quiet sound to the right; User B (looking down) hears a louder sound front left.

Methodology and Procedure

Very little previous work has carried out systematic and repeatable user experience evaluations in mobile audio-augmented reality. In addition, there is a lack of formal methodology on how to analyse and interpret user data that is not just qualitative, especially in an outdoor mobile audio-augmented exploratory environment. Thus, in this initial case study these issues are addressed by focusing on user performance both quantitatively and qualitatively over a number of different auditory displays. This resulted in a between-subjects design that offered rich and detailed results by participant but at the expense of controlling for cross-subject variation.

Eight users (6 male, 2 female, from 24 to 39 years in age), who were familiar to the experimenter, participated in the study. They were all students and members of staff at the University of Madeira and were familiar with the Municipal Gardens in Funchal. They all reported normal hearing and were right-handed. Five of these users had used GPS-based systems before. None were paid for their participation. Two different participants tested each of the four auditory display conditions described in the previous section. The experiment lasted no more than half an hour.

First, users were asked to familiarise themselves with the system by finding a landmark situated outside the park. This procedure served to check the system had GPS signal prior to starting the test and also provided participants with the chance to ask questions. They were then asked to enter the park and explore it freely whilst looking for the audio landmarks. They were all given a maximum of thirty minutes to walk around the garden (the instructions provided to participants can be found in Appendix E.1). Half were directed to start at the part of the park with the isolated landmarks, while the others started where the landmarks were clustered together. Participants were instructed to verbalise their thinking process (a 'think aloud') while they walked through the park, and this information was noted down (see Appendix E.2 for a sample of the think aloud note taking sheet used by the experimenter). As they encountered each audio landmark, the users were asked to listen to the corresponding audio clip before continuing their search. At the end of each trial for each different condition, participants filled in a questionnaire and provided informal feedback about their experience (see Appendix E.3). In addition to participants' comments and opinions, detailed logs (including distance covered, time spent, user location coordinates and head orientation) were collected on the mobile device to later perform an in-depth analysis of participant behaviour.

Metrics

In this evaluation quantitative and qualitative data were collected and analysed to investigate the impact of different auditory displays on the user experience of an exploratory mobile audio-augmented environment. The independent variable (IV) was the type of condition (the *Baseline* condition, the *Earcons* condition, the *Spatial* condition and the *Spatial3D* condition), and the dependent variables (DVs) were time taken to complete the sound garden experience, the distance walked in meters, walking speed in meters per second, time spent stationary and head-turning data collected from participants exposed to spatial auditory feedback. In addition to feedback from the participant questionnaire, user location coordinates and head orientation data were also included in an in-depth analysis of participant behaviour.

6.2.3 Results

Quantitative Analysis

The logged data showed that participants completed the experiment on average in 16 minutes and 15 seconds and the average distance covered by each subject was 692 meters (see Figure 6.4 and Figure 6.5 for more details per participant). The inclusion of spatialisation in the auditory feedback resulted in participants spending more time walking through the park and covering more distance.

In addition, participants' average speed dropped with increasing audio feedback complexity (Figure 6.6). The distribution of speed by non-spatial and spatial conditions (Figure 6.7) showed a significant main effect for condition type (t-test on log10 transform, to reduce skew, of speed values: $t(2874)=13.662$, $p<0.001$). Participants walked at a significantly lower speed during the spatial conditions (mean= 0.62 m/sec., SD= 0.51) than during the non-spatial conditions (mean= 0.90 m/sec., SD= 0.79). Looking more closely at the distributions, it can be seen that this drop in average speed was caused less by the participants walking more slowly but rather by an increase of the time they spent stationary (note the peak at 0 for spatial conditions compared to non-spatial conditions).

In this study a threshold of less than 0.25 m/sec. (0.9 km/h) was used to identify stationary periods to allow for error in GPS readings. Error from the GPS readings means that subsequent positions are rarely identical even when the participant is completely stationary. Thus, in order to quantify stationary periods, the threshold was set based on the observation of the distributions in Figure 6.7. Histograms for both the spatial conditions show a bimodal log distribution. As a participant was regarded to be either stationary or moving, these two distributions were fitted to these two states. Given an average human walking speed is 4.3 km/h, it is reasonable to regard 0.9 km/h as slow enough to be stationary. Using this threshold, Figure 6.8 shows the differences in the percentage of time participants were stationary. A Chi-square test showed that the percentage of time participants remained stationary significantly differed by condition ($\chi 2(3, N= 3025) = 85.565$, $p<0.001$). The effect of providing proximity information and full spatial auditory feedback was that participants appeared to stop more often.

The number of overlapping proximity zones for audio landmarks also had an effect on the percentage of time participants stayed stationary. Figure 6.9 shows percentage of time participants were stationary per number of nearby audio landmarks for the spatial conditions[1].

[1]Only data from within the proximity zone were considered and data points while in the activation zone were excluded as we were only interested in user behaviour while exploring and not once they had reached the activation zone.

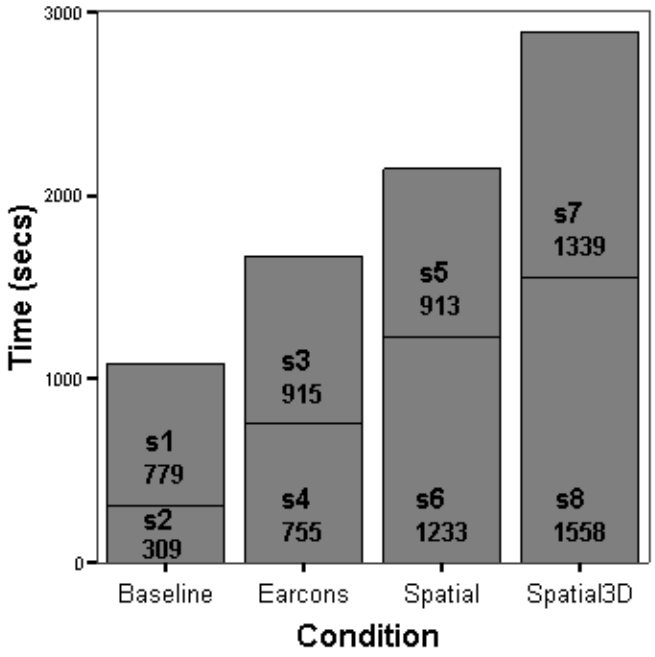


Figure 6.4: Time spent exploring for each participant (s1-8), stacked to show time spent per condition.

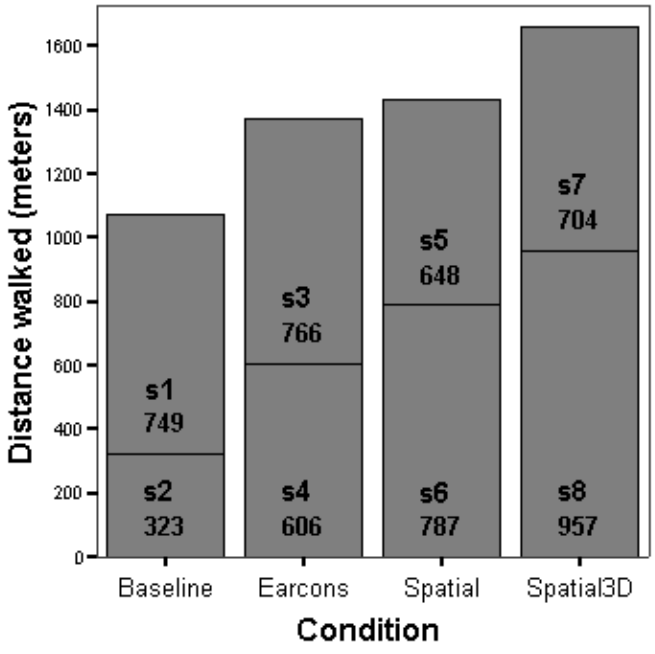


Figure 6.5: Distance walked for each participant (s1-8), stacked to show distance walked per condition.

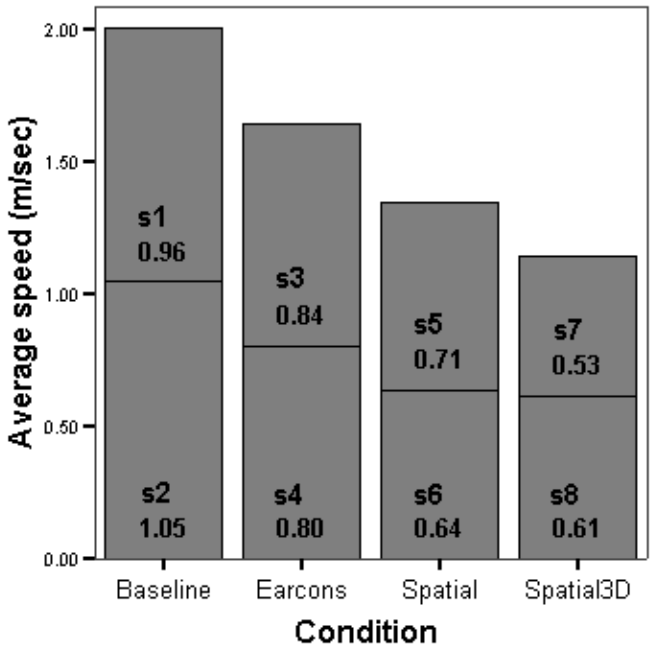


Figure 6.6: Average walking speed for each participant (s1-8), stacked to show walking speeds per condition.

A Chi-square test showed that the percentage of time participants remained stationary significantly differed by number of overlapping proximity zones for audio landmarks ($\chi 2(7, N= 842) = 100.273$, $p<0.001$). Participants exposed to full 3D audio feedback (Spatial3D condition) stopped more often as more proximity zones for the audio landmarks overlapped. In contrast, participants in the Spatial condition show a constant percentage of stopping as overlapping increased (see Figure 6.10a and Figure 6.10b in the *User behaviour* section immediately after the *User feedback* section for an illustrated example of user behaviour).

User Feedback

Based on the user feedback, the extra time spent stationary and the extra distance covered when auditory spatialisation was used, did not lead to frustration, rather it appears to be related to the enjoyment and sense of discovery of the participants. In contrast, for the conditions lacking auditory spatialisation, participants behaved more like in a navigation environment setting themselves the task of finding all the landmarks by systematically walking through the park. This behaviour emerged despite participants in all conditions being given the same set of in-

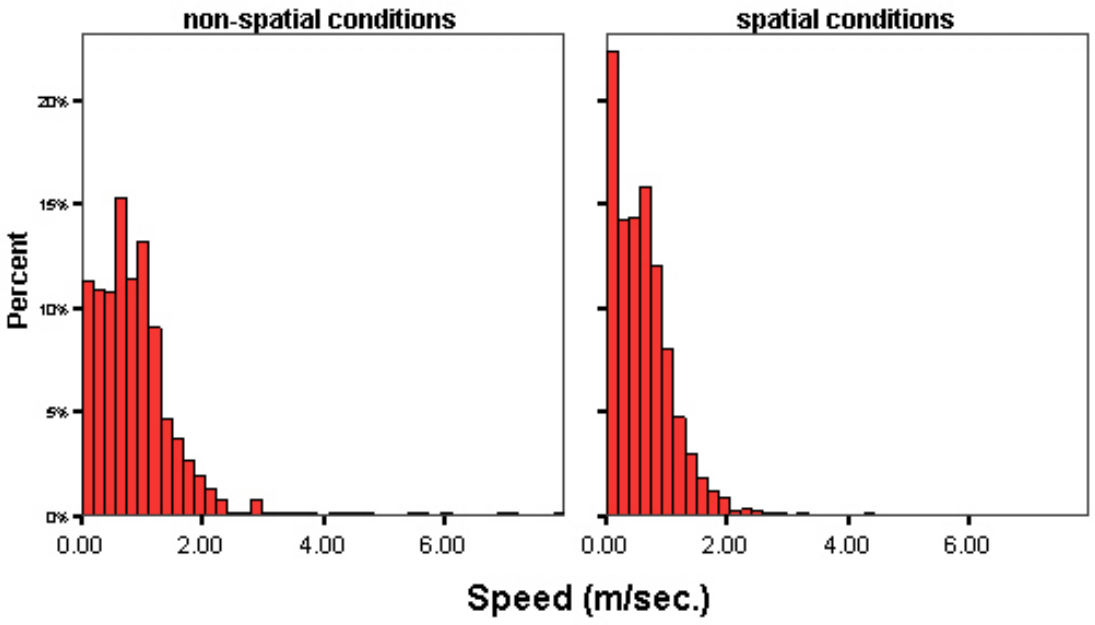


Figure 6.7: Histograms showing the distribution of walking speed by non-spatial (Baseline and Earcons) and spatial (Spatial and Spatial3D) conditions. Speed was calculated by dividing the distance walked by the time taken between each data point logged approximately every 2 seconds (mean = 2.28secs, SD = 0.29).

structions before starting the exploration of the garden (see Appendix E.1). They were all told to walk through the park in their own time and without rushing or walking too fast and that audio landmarks would be triggered as they got closer to them. Overall, sound levels were reported to be appropriate and the speech was clear and intelligible. Informal user feedback is presented for each of the four auditory display conditions.

Baseline In the first auditory display condition (no Earcons or spatialisation), the audio clips were simply triggered when users entered the activation zone. Consequently, the users tended to systematically explore to find the audio clips. Once they were located, users reported being pleased with locating the landmark but remarked the sound was "a bit abrupt when triggered". The value of the information in the audio clips was found to be appropriate, but especially directed towards tourists. The material in the audio clips was found "appropriate and informative" mainly due to the physical landmarks and because "if you were walking around the garden you wouldn't like to read it". One user suggested that the content of these audio clips "would potentially trigger a conversation" if walking with a friend or partner. The users highlighted that

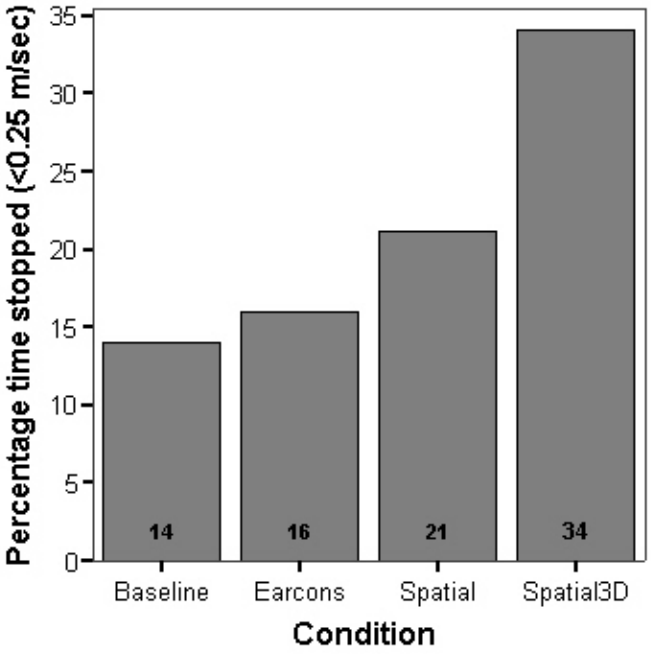


Figure 6.8: Percentage of time stopped for each condition. A threshold of less than 0.25m/sec was used to process user data identifying stationary periods.

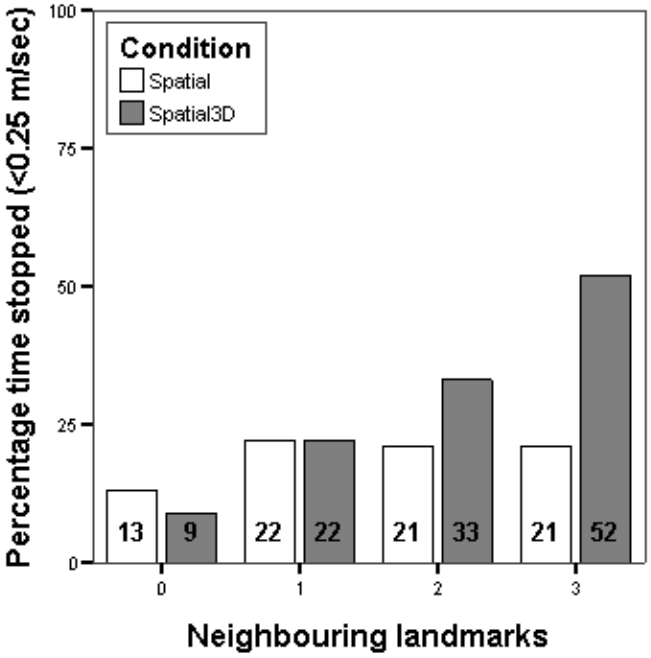


Figure 6.9: Percentage of time stopped for different numbers of overlapping proximity zones for audio landmarks. A threshold of less than 0.25m/sec was used to process user data identifying stationary periods.

"the experience of moving around to get the information was good" and the "novelty of coming across the message like stumbling across something on your way. Serendipity and wonder". Users found navigating the park to find the audio clips "very easy, just walking around" as it was "not a big space". However, the instability of the GPS information sometimes resulted in the user overshooting the physical landmark by the time the audio clip was triggered. Users sometimes found that "the sound was triggered after walking past" or "it was quiet and thought I was on the wrong path". One of the participants failed to find one of the audio clips, reflecting the difficulty of successfully exploring such a sound environment.

Earcons Earcons were present in three of the conditions. Participants reported they "liked the sound of the animals" and described them as "lively", "clear", "natural", "crisp" and "interesting". They reported enjoying the fact that "you just walk around and the sounds get triggered". Despite the background noises in the park, the animal sounds successfully indicated the presence of information at particular locations. One user remarked: "I liked that I realized that it [the animal sound] was prompting me to press the button. Maybe if it had been too realistic I would have missed that". In the third condition (which adjusted volume based on distance to the landmark in the proximity zone) one of the participants reported that the animal sounds "blended very well. Made it more seamless". The other participant felt that the echoing (reverb) in the animal sounds made him feel "like being in a quiet place in the forest. Reminded me of a place close to home". A participant in the fourth, fully spatialised condition suggested that it "helped that they [the sounds] were different from the ones already in the park". Both participants in this final condition enjoyed the animal sounds, stating they were the "best part" and "especially nice for a garden like this one". They did not expect these animal sounds to blend so well and also found them "just playful in themselves".

Auditory spatialisation *Spatial*

Participants experiencing the Spatial condition, in which the amplitude of the Earcons varied with distance to target, reported this to be useful and appropriate. The intensity was reported to remain at a comfortable level throughout. However, users experienced difficulty determining the distance to particular landmarks. One stated: "guessing how close I was from a location was based on distance travelled when I first heard it and intensity combined. Not proportional" and reported that the alterations to volume were not physically accurate. The other user noted that the variations in volume were a bit "jumpy", something probably due to noise in the GPS position sensing. He also noted, that "it took time to get used to the distance distinction near/far. Once I found the first one [landmark] it was easier to find the others because I already knew what I was looking for".

Auditory spatialisation *Spatial3D*

During the Spatial3D condition, participants reported a sense of "discovery" and that the sound garden was "quite immersive". The participants in this condition liked the experience because "you rely only on your hearing" and often closed their eyes in order to listen to the Earcons. They found the system curious because "you know sounds come from headphones but it sounds like it is coming from the outside". The variation in loudness used to represent distance away from the landmark gave "a good indication of distance" but it was also reported that "going from far away to closer was too quick". One participant stated that even in situations with multiple sound sources "overall the localisation was easy" but became harder in the area of the park where three animal sounds overlapped. However, when the user walked away from this area and only two animal sounds overlapped, heading helped. This was echoed by the opinion that while two overlapping sounds were understandable, three were "a bit chaotic". Overlapping sounds also conveyed benefits as "hearing sounds at a distance that [I] have already heard gave familiarization with the surroundings". One of the users admitted: "it would be difficult to find them [landmarks] without spatialisation. If it doesn't point you in the right direction it would be harder".

User Behaviour

A more detailed analysis of the logged data for each participant revealed a tendency for participants in the Baseline condition to walk at a steadier pace, in straighter lines, while looking in the direction they were going, when compared to participants in the Spatial3D condition. Figure 6.10a shows an example of subject 1 in the Baseline condition walking from the stone coat of arms to the statue of Joao Reis Gomez. The solid line is the direction of travel and the short splines illustrate the participant's head orientation approximately every two seconds. Figure 6.10b shows a contrasting path from a participant in the Spatial3D condition.

The grey rings 1&2 highlight two points where the participant stopped and began looking around, probably trying to ascertain the direction of the audio being played in the proximity zone. This type of behaviour was typical of the Spatial3D condition where the head movement while stationary appears to characterise a 'searching behaviour'. If we examine the distributions of head orientation change for the spatial conditions (see Figure 6.11), it can be observed that the Spatial3D condition encourages this type of head movement (lower percentage of 0° data points and broader distribution) compared to the Spatial condition showing a more peaked distribution, i.e. a different kurtosis[2].

[2] *Kurtosis* is the name of a statistical measure used to describe the distribution of observed data around the mean. A normal distribution has a kurtosis 0, a peaked (tall and skinny) distribution has a positive or high kurtosis and a flat distribution has a negative or low kurtosis.

The mean and SD of both distributions are similar (Spatial: mean= 0.038, SD= 42.77; Spatial3D: mean= -1.612, SD= 50.630), however the kurtosis is quite different (Spatial: Kurtosis= 3.470; Spatial3D: Kurtosis= 1.648). This means that head change within the regions 36° to 108° contains more data than angles closer to 0° and wider angles. Wider angles are likely to be caused by changes in body position.

This would fit the observation that participants moved their heads from side to side in the 3D spatial condition to gauge the direction of sounds heard. Although there is no formal statistic test to compare Kurtosis, a Chi-square test on observed counts across five bins (as shown in Figure 6.11), showed that observed counts from the Spatial3D condition significantly differed from expected counts matched based on likelihoods calculated on observed values in the Spatial condition ($\chi 2(4, N= 1160) = 73.764, p<0.001$).

If we compare logged information from participants with limited spatial information, it can be seen that they did stop as in the Spatial3D condition, but they seemed to keep their head much closer to their direction of travel (Figure 6.12a). Finally, Figure 6.12b shows one of the participants in the Spatial 3D condition within the three overlapping proximity zones. This participant shows an extreme case example of amount of head-turning to ascertain direction, which frequently occurred in the spatial conditions. This user in particular spent a substantial amount of time walking and altering his head position in order to determine the direction of one

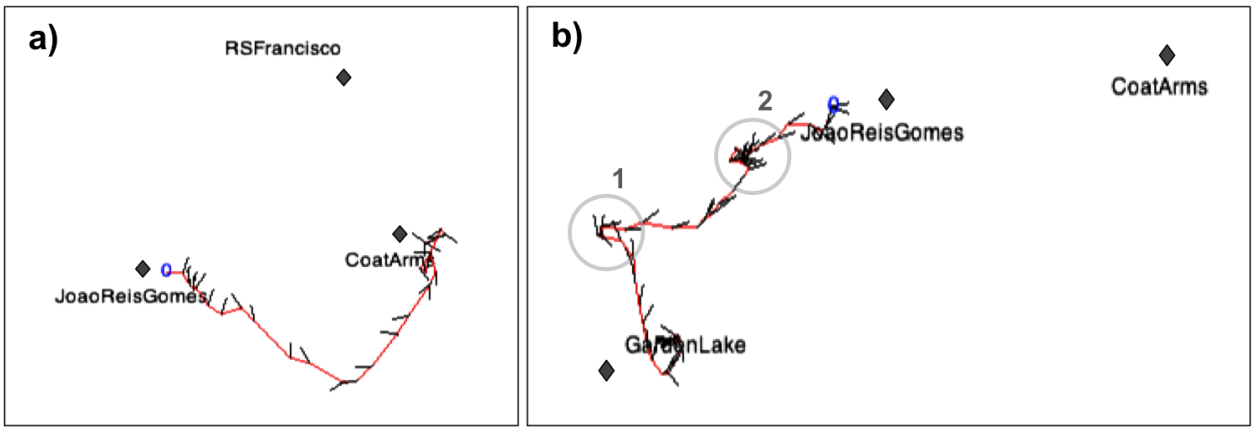


Figure 6.10: **a)** Route taken by one user from the stone coat of arms to the statue of Joao Reis Gomez during the Baseline condition. **b)** Route taken by one user from the Garden Lake to the statue of Joao Reis Gomez during the full 3D audio spatialisation (Spatial3D) condition. Gray circles indicate stationary periods along the route with greater amounts of head-turning. Short splines illustrate the user head direction approx. every 2 seconds (mean= 2.28 secs, SD= 0.29).

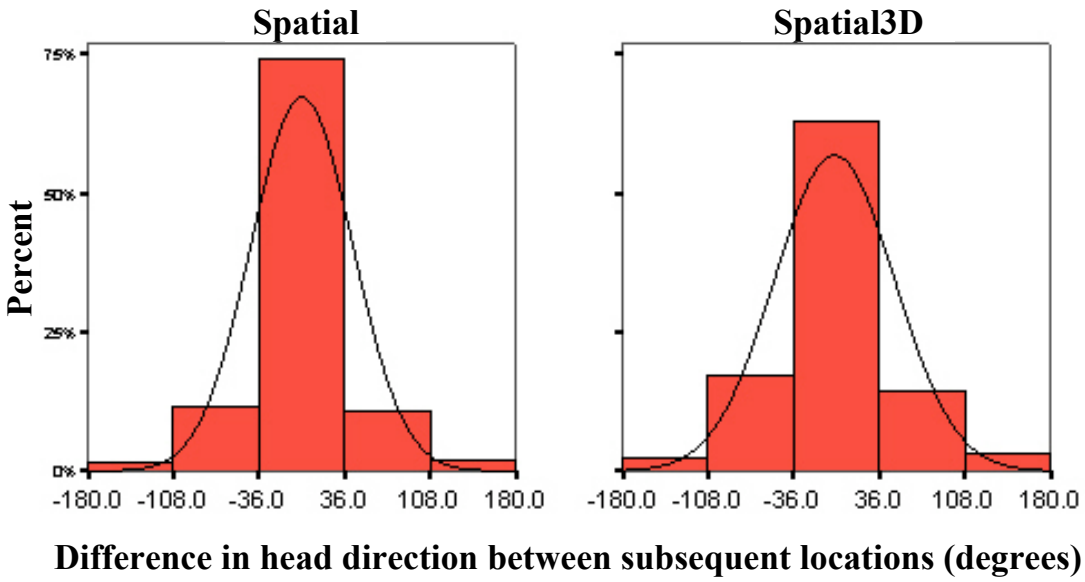


Figure 6.11: Histograms showing the distribution of the total amount of head-turning for the spatial conditions. Head-turning auditory feedback was only provided in the Spatial and Spatial3D conditions.

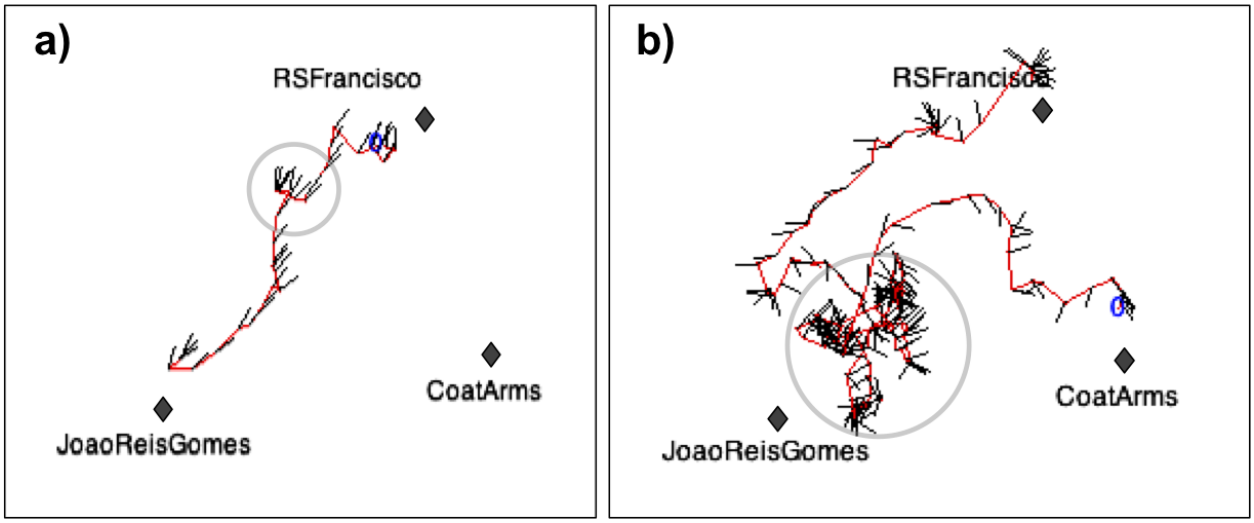


Figure 6.12: **a)** Route taken by one user from the statue of Joao Reis Gomez to the Rua Sao Francisco during the limited auditory spatialisation (Spatial) condition. Head direction fits much closer to the direction of travel (short splines illustrate the user head direction). **b)** Route taken by one user from the Rua Sao Francisco to the stone coat of arms during the full 3D audio spatialisation (Spatial3D) condition. Head direction changes greatly in order to determine the direction of one of the landmarks as illustrated by the route data within the grey circle.

of the landmarks. Far from frustrating, as user feedback showed, this searching process was enjoyable and added to the sound garden experience.

6.2.4 Discussion

In this initial case study, four different auditory displays were evaluated in a mobile audio-augmented reality environment (a sound garden). User behaviour was quantified and qualitatively described and found that it varied when exposed to the different configurations of proximity zones, non-speech sound and spatial 3D audio available in our sound garden. In addition, head-turning data and its relation to user behaviour in exploratory environments was also analysed, with particular attention devoted to situations when multiple audio landmarks overlap. Although this study did not examine a large user sample, the wide range of measurements recorded were able to support a rich, detailed and informative analysis.

Results from this study showed that when users were provided with spatial auditory feedback within the proximity zone, they spent more time in the park, walked more and spent more time stationary while turning their heads searching for landmarks. When distance away from the landmark was the only spatial audio cue available, some participants reported it to be useful while others were confused by the relationship between virtual and physical distance. GPS error also appeared to interfere with the overall experience. However, in the Spatial3D condition, participants reported that the auditory feedback gave a good indication of distance and was more immersive.

Users reported they were able to resolve two overlapping sounds easily but when three sounds overlapped, although heading information was a great help, localising the sounds became harder. As a result, participants' average speed dropped more when they were provided with spatial 3D audio feedback, as they had to stop to search and ascertain the direction of the audio, which was not the case when distance was the only cue available. However, far from frustrating users, they appeared to take their time to enjoy the sense of discovery (on average 21 minutes when spatialised compared to 11.49 when not spatialised), and immerse themselves in this mobile audio-augmented reality environment. Earcons also played an important role as a playful element successfully indicating the presence of information at a particular location. In contrast, when users were not provided with spatial auditory feedback, they systematically explored (as in a navigation task) at a steadier pace in straight lines mainly looking in the direction they were walking. In the non-spatial conditions, only the animal sounds were reported to provide a touch of playfulness to the exploration. However, users remarked on the abruptness of walking right into the audio clips and GPS error had a worse effect on the user experience in this case.

A number of technical limitations affected this study. Firstly, as from previous work, GPS can be problematic when seeking to situate audio precisely in space. However, GPS technology is increasingly present in smartphones and becoming ever more popular in mobile applications making use of geo-tagged data. Participants in this study did complain that the sound garden was jerky and unpredictable at times due to variance in the position reported by the GPS unit. Despite this system limitation, a high level of immersion was reported by users when exposed to spatial 3D audio and the combination of proximity and activation zones around the landmarks helped minimise GPS error.

Four separate devices were used in the system: a GPS unit, a magnetometer unit, a mobile phone and a pair of headphones. This was a somewhat overwhelming collection of devices and there would be many benefits to creating a more integrated solution. However, as the sensors were all situated on the headphones, one key advantage of this solution is that it enabled true 3D audio interaction based on head position and orientation. It is not clear the sound garden would be as compelling if all sensing was integrated into a handheld device, but further work is required to explore this issue.

These results build on previous work by extending and evaluating the complexity of the auditory spaces used previously for exploration in audio-augmented environments (Rozier *et al.*, 2000; Reid *et al.*, 2005; Holland *et al.*, 2002; Cater *et al.*, 2007; McGookin *et al.*, 2009; Magnusson *et al.*, 2009). Moreover, this study offers an initial qualitative and quantitative insight into overlapping spatialised sounds in a realistic environment, a design feature first implemented by Stahl (2007) but never evaluated. In particular, the findings on the critical importance of head position data in spatialised mobile audio-augmented environments confirm and complement those by Heller *et al.* (2009) and Mariette (2010). Ultimately, this work follows up on recent studies describing the design of purely exploratory audio-augmented reality systems such as CORONA (Heller *et al.*, 2009) and Soundcrumbs (Magnusson *et al.*, 2009), rather than on navigational tasks e.g. (Holland *et al.*, 2002; Cater *et al.*, 2007; McGookin *et al.*, 2009). As in Heller and Magnusson's work, the non-speech sounds used to identify the landmarks created an enjoyable and playful experience, despite increasing the auditory feedback complexity due to their spatially overlapping nature.

A number of practical lessons were also learned regarding the creation of audio-driven sound gardens. For example, although the circular activation zones used in this work are simple and easy to understand, they are a poor fit for the complexities of a space with paths, hedges and trees. There is a clear tension between situating sounds at the correct geographical location and situating them at a place where it is possible to ensure that users can observe the target item. With activation radii of 10 or more meters, users can easily encounter sounds from behind barriers

such as walls or dense plants, a potentially confusing situation. One clear way to address this is through developing non-circular activation regions, but this may also cause problems, as the realism of the metaphor connecting the virtual sounds to physical spaces may break down. Other solutions may include dynamically adjusting activation zones or calculating optimal solutions, which maximise the size of all zones, as in the bubble cursor (Grossman and Blakrishnan, 2005). Exploring richer interactions with the sound sources would also be beneficial. In this work, users were able to explore a physical space and press a button to start an audio clip. By allowing other interactions such as silencing, moving, adjusting or otherwise interacting with audio in a sound garden, it may be possible to create denser audio environments, which remain simple, effective and engaging.

6.3 Conclusions

This chapter presented the design and evaluation of four different auditory interfaces in a sound garden, in which user exploration and interaction strategies with location-based information were investigated. The initial findings and methods presented in this case study provide a valuable framework for the analysis and description of user behaviour in exploratory mobile audio-augmented reality environments.

The use of 3D audio was proposed as an effective technique to disambiguate multiple sound sources in mobile audio-augmented reality environments. A quantitative and qualitative analysis of the data gathered from the initial case study described in this chapter showed that the combination of 3D audio techniques together with Earcons was the most effective auditory display when dealing with audio landmarks that are very close together or overlapping, either due to concentration of information, or positional error. The quantitative data presented in this study aimed at describing users' exploratory behaviour. As discussed in the introduction of this chapter, assessment of exploratory behaviour presents a significant challenge. In a standard task-based assessment carrying out the task quicker is better. For example, if the task is navigation, finding the shortest distance in the fastest time is a positive performance indicator. For an exploratory system, in contrast, taking more time and covering more area may be a positive performance indicator, but only if this extra time and energy is not seen by users as detrimental to their overall user experience. In this study, a detailed analysis of user performance and behaviour together with positive informal user feedback, supported the hypothesis that taking more time was better. However, a formal evaluation of user experience using indicators such as perceived workload and user satisfaction could offer a stronger qualitative method for assessing a user's reaction to such a system. If a performance indicator, such as time, increases but perceived workload and

user satisfaction do not change then this supports the hypothesis that time spent was enjoyable and exploratory. However, if a significant increase in time is coupled with a significant increase in perceived workload and a decrease in user satisfaction, then the time spent can be associated with frustration and a poor user experience. In the following chapter, such formal assessment will be included as part of the evaluation framework.

The results presented in this chapter partly contribute to answering the second research question posed by this thesis: "How can 3D audio techniques be used to disambiguate multiple auditory sources in order to access location-based information in a mobile eyes-free interface?". This contribution is, however, not complete without an investigation of denser audio environments incorporating richer interactions with sound sources. The next chapter presents a further study on more complex interaction techniques that support multiple auditory sources in an exploratory environment, not only when audio landmarks overlap, but also when these audio landmarks enable access to more than one auditory information item.

Chapter 7

Supporting Multi-Level Auditory Displays in Mobile Audio-Augmented Reality

7.1 Introduction

This chapter reports a study in which complex interaction techniques are investigated in a mobile audio-augmented reality environment. This study contributes to answering Research Question 2, "How can 3D audio techniques be used to disambiguate multiple auditory sources in order to access location-based information in a mobile eyes-free interface?" by investigating the efficiency and usability of complex spatial auditory displays designed to enable user interactions with concentrated areas of information in a location-based system.

In previous chapters, a number of spatial auditory display designs were evaluated. In Chapter 5, a 3D audio technique referred to as *spatial minimisation* was successfully used to aid user multitasking. This spatial minimisation technique was implemented using an egocentic auditory display in which a number of auditory streams, each representing a task, were placed at fixed positions relative to the user's head. This kind of egocentric design facilitates user interaction with the auditory streams as their position is always fixed with respect to the user, even when the user orientation has changed. On the other hand, Chapter 6 investigated the use of an exocentric auditory display to discover information at particular physical locations in an exploratory mobile audio-augmented reality environment. The design of exocentric displays for applications such as audio-augmented reality environments require a multi-layered or *multi-levelled* approach. A top level functions as a sonification layer, with a proximity zone used to identify audio-augmented locations and an activation zone used to access a secondary level containing the information the

user can interact with (see Section 3.3.2). The research work presented in Chapter 6 showed that a spatialised top level delivered a more engaging and immersive user experience than a number of other non-spatialised alternatives and aided users' exploration when audio-augmented locations overlapped. However, the secondary level only allowed for interactions with one audio information item and interactions with multiple information items were not investigated. If multiple information items must be supported in the secondary level of the auditory display, how should such an interface be designed?

In order to design an interactive secondary level containing multiple location-based information items in a multi-level auditory display, we again need to consider how to present the auditory streams without overloading the user. Should we mirror the presentation arrangement displayed in the top layer or would other designs, such as a combination of exocentric top level and an egocentric secondary level, be more usable and efficient? A homogeneous design across levels in the auditory display would follow the design principle of consistency. Consistency is a widely used principle in user interface design (Helander *et al.*, 1997; Shneiderman, 1998; Nielsen, 1994) and it has been found to impact both usability and cognitive load (Lund, 1997). In visual interfaces "consistency allows users to transfer existing knowledge to new tasks, learn new things more quickly, and focus more on tasks because they need not spend time trying to remember the differences in interaction. By providing a sense of stability, consistency makes the interface familiar and predictable" (Microsoft, 1995). When designing visual interfaces that display large amounts of data, multiple visual levels have been suggested in order to improve usability and reduce cognitive load (Lam and Munzner, 2010). A Zoomable User Interface (ZUI) is an example of such an interface. ZUIs have been defined as *"systems that support the multi-scale and spatial organisation of and magnification-based navigation among multiple documents or visual objects"*(Bederson, 2011). In a ZUI, multiple visual information can be presented simultaneously using a *consistent* multi-level layout in which the user can navigate to different zoom levels by zooming up close to interact with detailed content or zooming out for an overview. In an auditory display, this overview+detail structure could be supported in a multi-level auditory display implemented using 3D audio techniques on both levels in order to present information simultaneously. In this way, zooming between the sonification top (overview) level and the interactive secondary (detail) level would remain consistent.

The research work described in this chapter investigates the potential of different spatial audio configurations to enable mobile user interaction with multiple location-based information items in a multi-level auditory display. By establishing the efficiency and usability of these different configurations, this work is contributing further to answering the second research question in this thesis, "How can 3D audio techniques be used to disambiguate multiple auditory sources in

order to access location-based information in a mobile eyes-free interface?".

7.2 Experimental Study

The aim of the work reported in this chapter was to establish the efficiency and usability of a number of spatial audio configurations in a multi-level auditory display, and to understand how these different configurations affected the user experience. As with the case study reported in Chapter 6, this work was carried out in a non-guided exploratory environment. However, in order to address the limitations of the outdoor user tracking technique used in Chapter 6 (as discussed in Section 6.2.1), the study reported in this chapter was carried out indoors to improve user tracking accuracy. Minimising user tracking inaccuracies improves the system implementation and most importantly aids users to concentrate solely on the audio experience.

7.2.1 Audio-augmented Art Exhibition Implementation

A conceptual art exhibition was used as the setting for this study. A variety of different mobile auditory interfaces designed to provide access to multiple location-based information were implemented and tested in this exhibition space, always aiming at a full eyes-free interaction between the user and the mobile device running the auditory interfaces so the user's visual attention can be focused on the interaction with the object being audio augmented.

The virtual audio environment superimposed on the art exhibition was run on a Nokia N95 8GB and the built-in HRTFs and the JAVA JSR-234 Advanced Multimedia Supplements API were used to position the auditory sources. User position was determined using an Infrared (IR) camera tracking an IR tag powered by a 9V battery (see Figure 7.1a) and mounted on top of a pair of headphones. Coordinate information was fed to the mobile phone over a WiFi connection and was used to activate the zones associated with the art pieces in the exhibition space. User orientation (compass heading) was determined using a JAKE Sensor Pack (2010) (see Figure 7.1b) connected to the mobile phone via Bluetooth. No visual aids were provided on the screen of the mobile device and, to ensure a full eyes-free experience, the phone was placed on a lanyard around the user's neck (see Figure 7.2 (right)). The navigation switch on a SHAKE SK6 sensor pack (2010), also connected via Bluetooth, was used to feed user input into the system while users were holding it in their hands. This navigation switch allowed users to activate and deactivate audio content by pressing the switch and also to browse the content by pushing the switch left or right (see Figure 7.2 (left)). The audio was played over a pair of DT431 Beyerdynamic open-back headphones with the aim to reduce the isolation of the listener

Figure 7.1: (a) IR tag with 9V battery attached, (b) JAKE sensor pack shown with a five cent euro piece.

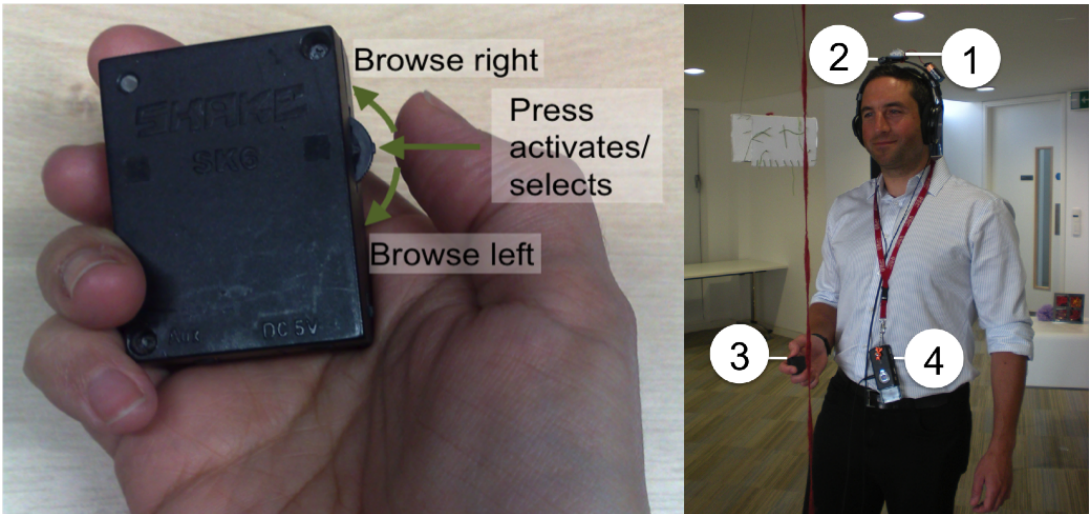


Figure 7.2: Experimental setup (right): 1) IR tag and 2) JAKE sensor (as shown in Figure 7.1, both mounted on headphones), 3) SHAKE SK6 sensor pack and 4) mobile device; and interaction technique using the navigation switch on the SHAKE sensor pack (left).

from the surrounding environment. The IR tag and JAKE sensor were placed in the middle of the headphone's headband and both mounted using Velcro tape. Figure 7.2 (right) shows the final system setup and 7.3 shows the system architecture.

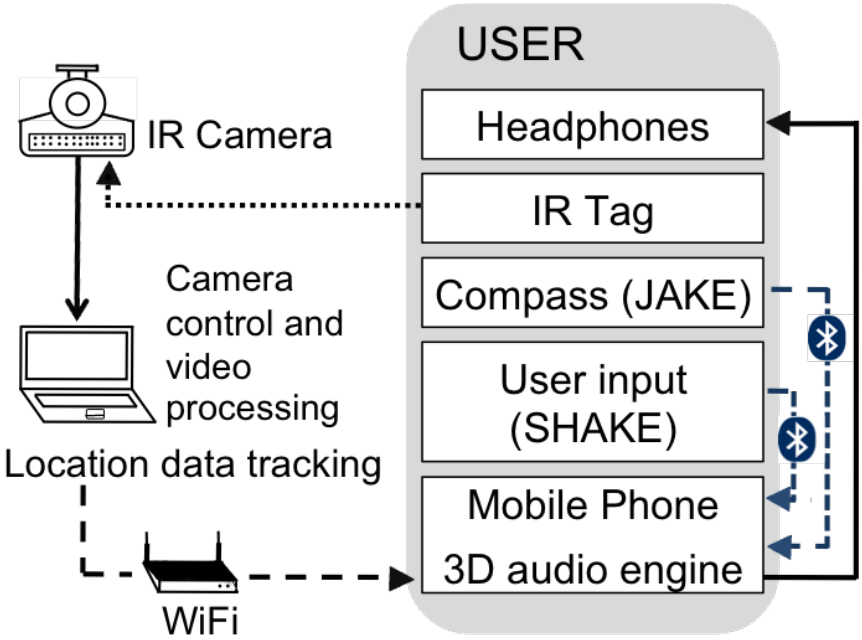


Figure 7.3: Schematic representation of the system architecture.

Conceptual Art Exhibition Space

The exhibition consisted of six different art pieces from the *Weaving the City* project (`www.weavingthecity.eu`) kindly donated by Rocio Von Jungenfeld, a PhD student at the Edinburgh College of Art. Four art pieces were made of woollen threads and paper and exhibited in a space that measured 3m wide(x) x 3.85m long(y). They were complemented by another two media pieces placed outside the exhibition space. One media piece captured the participants' image via a webcam as they walked past and, after being processed using a Max/MSP patch running on a Mac mini, projected on the wall. A second media piece was a movie about the Weaving the City project playing in a loop on an iBook G4. The media pieces were not audio-augmented, i.e. no audio information was offered about these pieces, and their purpose was to make the exhibition space more playful and immersive with the help of the projected images and sounds. Two of the art pieces were suspended from the ceiling hanging at eye level and the rest, including the media pieces, were placed on small tables (see Figure 7.4 for an illustration of the setup).

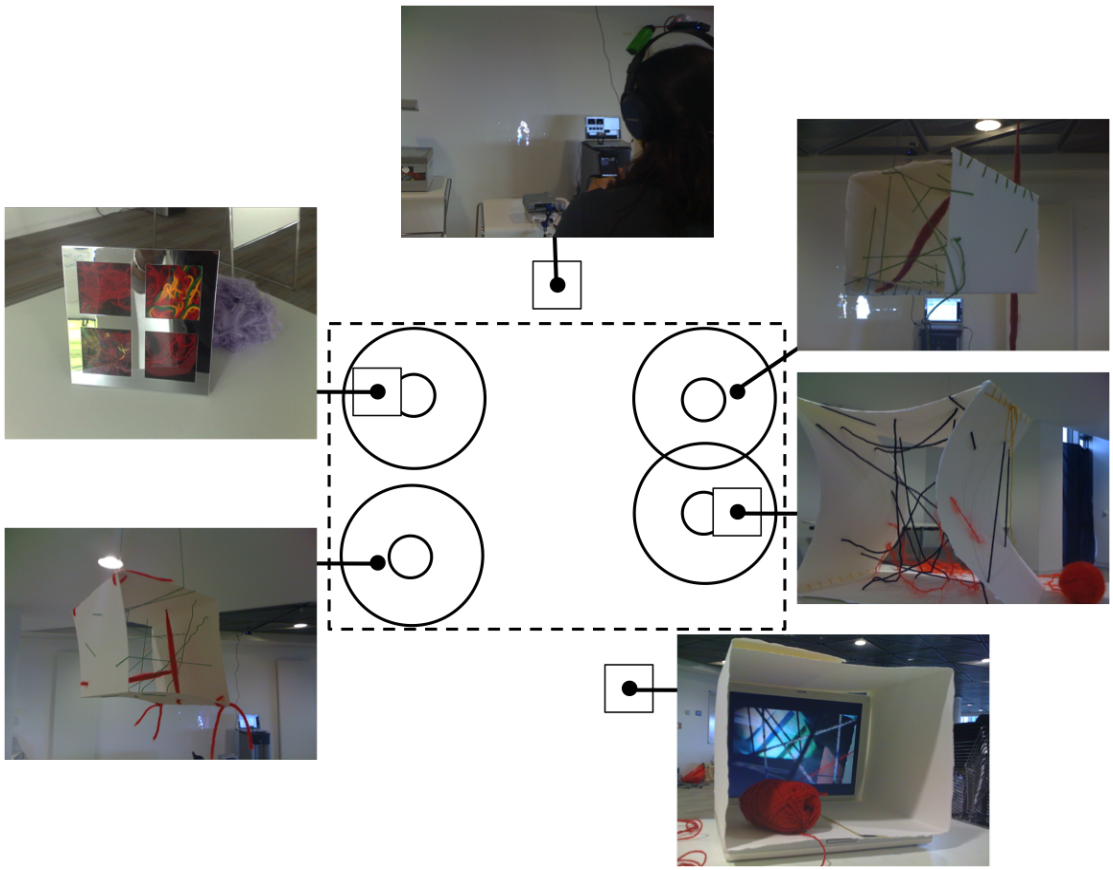


Figure 7.4: Illustration of the exhibition area layout and the top-level sonification layer showing the location of the proximity and activation zones surrounding each art piece. The dashed-line area identifies the audio-augmented exhibition space measuring 3m(x) x 3.85m(y). The small squares with a dot at its centre identify the art pieces placed on tables and a dot alone the ones that hung from the ceiling.

User Location Tracking

In this study, the indoor tracking system consisted of a PlayStation® Eye camera modified to work as an IR camera and an IR tag used to track the user location in the exhibition space. The IR camera was attached to the ceiling using velcro and connected via USB to a MacBook computer running Community Core Vision - CCV (`http://ccv.nuigroup.com`) and Processing (`http://processing.org`) open source software. CCV takes a video input stream and outputs tracking data as coordinates and it is frequently used in building multi-touch applications (Correia *et al.*, 2010; Fu *et al.*, 2010; Roth *et al.*, 2010; Leftheriotis and Chorianopoulos, 2011; Zhang *et al.*, 2012). In this study CCV tracked the position of the IR tag mounted on the partici-

pant's headphones. Then, the Processing application used the TUIO (Table-Top User Interfaces Objects) API to decode the TUIO messages sent out from CCV and output coordinate information at 2Hz to a multicast network socket. The coordinate information could then be accessed by the mobile phone running the audio-augmented environment. The Processing application also plotted user location in real-time on the computer screen so the experimenter could confirm the tracking was active.

This indoor tracking system was first calibrated before measuring the tracking error (i.e., the average Euclidean distance deviation in centimetres (cm) between the actual position of a number of target points and the estimated user position as detected by the IR camera at the same target points). Three volunteers of different heights (1.58cm, 1.73cm and 1.92cm) took part in the initial IR camera calibration. The IR camera was first calibrated with the middle height (1.73cm) using a total of nine reference points across the 3m x 3.85m exhibition space and these camera settings were used for all three participants. In this way, the tracking error for shorter or taller heights could be calculated. Then, a total of 19 fixed target points were identified across the tracked exhibition space. An application was devised to present the fixed target points one at a time on the screen of the mobile device so the participant could then walk to the location indicated. Once the location was reached, pressing a button on the device logged the user location. Results showed that the tracking error was 55.69cm, 28.45cm and 76.33cm for the 1.58cm, 1.73cm and 1.92cm high participants respectively. Given that the tracking error varied considerably depending on height and it was lower for the middle height used to calibrate the IR camera in the first place, it was decided that the IR camera would be calibrated individually for each of the participants taking part in the evaluation study.

Multi-level Auditory Display Design and Stimuli

There were two levels in the multi-level auditory display: 1) A top-level sonification layer and 2) a secondary interactive layer.

As in the case study reported in Chapter 6, Earcons were used to advertise the content of each exhibit. In that earlier study, choosing sounds that fitted the environment and had an ambient quality contributed to a sense of immersion (See *Audio spatialisation: Spatial3D* in Section 6.2.3). For this study the following sounds were chosen:

- Top-level sonification layer: Chattering voices were used to advertise content about each art piece. This sound was chosen because it fitted the gallery environment by representing an item of public interest which would encourage discussion.

- Secondary interactive layer: For each art piece, different sounds were used to identify comments left by the artist and those left by non-expert reviewers. In order to provide a uniform listening environment in this layer, sounds representing the elements (i.e., water, fire and wind) were chosen. The sound of "water waves" was chosen for the artist's comments to represent a deeper understanding of the work, an "open crackling fire" sound representing warmth and excitement was chosen for positive non-expert reviews, and a "stormy wind" sound was used to represent the negative (cold) non-expert reviews.

The top-level sonification layer attracted visitors towards the artwork and advertised the existence of information at that location. As in the case study reported in Chapter 6 (see Sections 6.2.1 and 6.2.2), a circular proximity zone (radius 1.25m) advertised content and a smaller circular activation zone (radius 0.75m) enabled user access it. The chattering voices sound used in this layer was mono, 16-bit and sampled at 16kHz (see Appendix A 4). The chattering voices were presented within the proximity zone surrounding each art piece using an exocentric design (sound positions were updated in real-time according to the user orientation and the loudness of the sound increased as the distance to the art piece decreased (for more details see 6.2.2)) to provide the user with orientation and distance information, while the activation zone was user-activated. The proximity zones overlapped for two of the art pieces while the other two were isolated (see Figure 7.4).

The secondary interactive layer was user-activated and only accessible when in the activation zone of the top-level sonification layer. This display contained an audio menu with information about the art piece. It consisted of a variable number of audio menu items from a minimum of one to a maximum of three. Each audio menu item was identified with a different sound, namely, "water waves", "open crackling fire" or "stormy wind" and each of these menu items included information that was less than 25 seconds long. User interaction with the audio menu items varied for the different experimental conditions, as will be described in the next section. Both the audio menu items and their related information were mono, 16-bit and sampled at 16kHz (see Appendix A 4).

Both the chattering sound in the top-level sonification layer and the audio menu items in the secondary interactive layer were adjusted to conversational volume (approx. 60-70dB).

7.2.2 Design of the Experiment

Participants

Thirty-two participants (21 males, 11 females, aged 18 to 39 years) were recruited, all were studying or working at the University. They all reported normal hearing, were right-handed and paid £6 for participation, which lasted just over an hour. 12.5% (n=4) of the participants reported that they rarely went to museums or art galleries, 12.5% (n=4) reported they went once a year at most, 53.1% (n=17) two to three times a year, 18.8% (n=6) no more than once a month and 3.1% (n=1) at least once a week. Only 15.6% (n=5) of the participants had never used an interactive museum system in the past.

Participants were split equally into two groups: sequential and simultaneous presentation, in a between-subjects design. In the sequential audio group the audio menu items in the interactive auditory display were presented sequentially one at a time, whereas in the simultaneous presentation group audio items were presented simultaneously all at the same time.

Conditions

Each group (*sequential* and *simultaneous presentation* group) was tested in three different conditions in which the secondary interactive layer varied in complexity:

1. *Baseline*: Each audio menu item was *always* played sequentially at each push of the navigation switch either right or left for both presentation groups. There was no spatialisation of the audio items so they seemed to originate from within the users head. The aim was to recreate a traditional audio guide style interaction in which users triggered the audio content by the press of a button in a sequential order. See Figure 7.5a.
2. *Egocentric*: Each audio menu item was presented in a radial menu (virtually located around the users head to the right, left or in front of the user's nose) and played one at a time when selected by pushing the navigation switch for the sequential presentation group (see Figure 7.5b). When an audio menu item was selected in the simultaneous presentation group, the volume increased for the selected item to bring it into focus and decreased for the rest. Selection was performed pushing the navigation switch either right or left and the audio menu items were located at 0°, -90° and +90° azimuth (see Figure 7.5c).
3. *Exocentric*: Each audio menu item was situated in the exhibition space exocentrically in front of the art piece orientated towards the centre of the exhibition space and at a

minimum 45° separation of each other (see Figure 7.6). For both the sequential and simultaneous presentation groups the audio menu items were perceived as if they were fixed to a location. Selection of the audio menu items for the sequential presentation group was performed pushing the navigation switch either right or left (see Figure 7.5d), whereas in the simultaneous presentation group selection was performed by walking around an art piece and then standing at the location were the audio menu item was situated (see Figure 7.5e). A loudness cue identified the activation area where audio menu items could be selected. The Proximity zone around each audio menu item was 3m to ensure all items would overlap and play simultaneously and the Activation zone was 1m. Here, a consistent design across the exocentric top-level sonification layer and an exocentric secondary interactive layer was tested.

The order of the conditions was randomised per participant to control for ordering effects in both the sequential and simultaneous presentation groups (see Table 7.1 for a summary of experimental conditions). Participants were tested in the mobile environment provided by the conceptual art exhibition space.

Procedure

The experiment included a calibration procedure and a training session before the test conditions. First, the indoor tracking system was calibrated for the height of each participant. This followed a training session using the starting test condition to familiarise the participant with the multi-level auditory displays around one of the art pieces in the exhibition space. For each test condition, participants were asked to explore the exhibition space and find as much information as possible about the art pieces by interacting with the different auditory displays. The experimental instructions and brief introduction to the exhibition can be found in Appendix F.1 and Appendix F.2 respectively. The auditory display description per test condition can be found in Appendix F.3. As participants walked closer to the audio-augmented art pieces, the proximity zone in the top-level sonification layer was triggered and the sound of chattering voices was

	Condition		
Presentation group	Baseline	Egocentric	Exocentric
Sequential	BaseSeq	EgoSeq	ExoSeq
Simultaneous	BaseSim	EgoSim	ExoSim

Table 7.1: Summary of Experimental conditions.

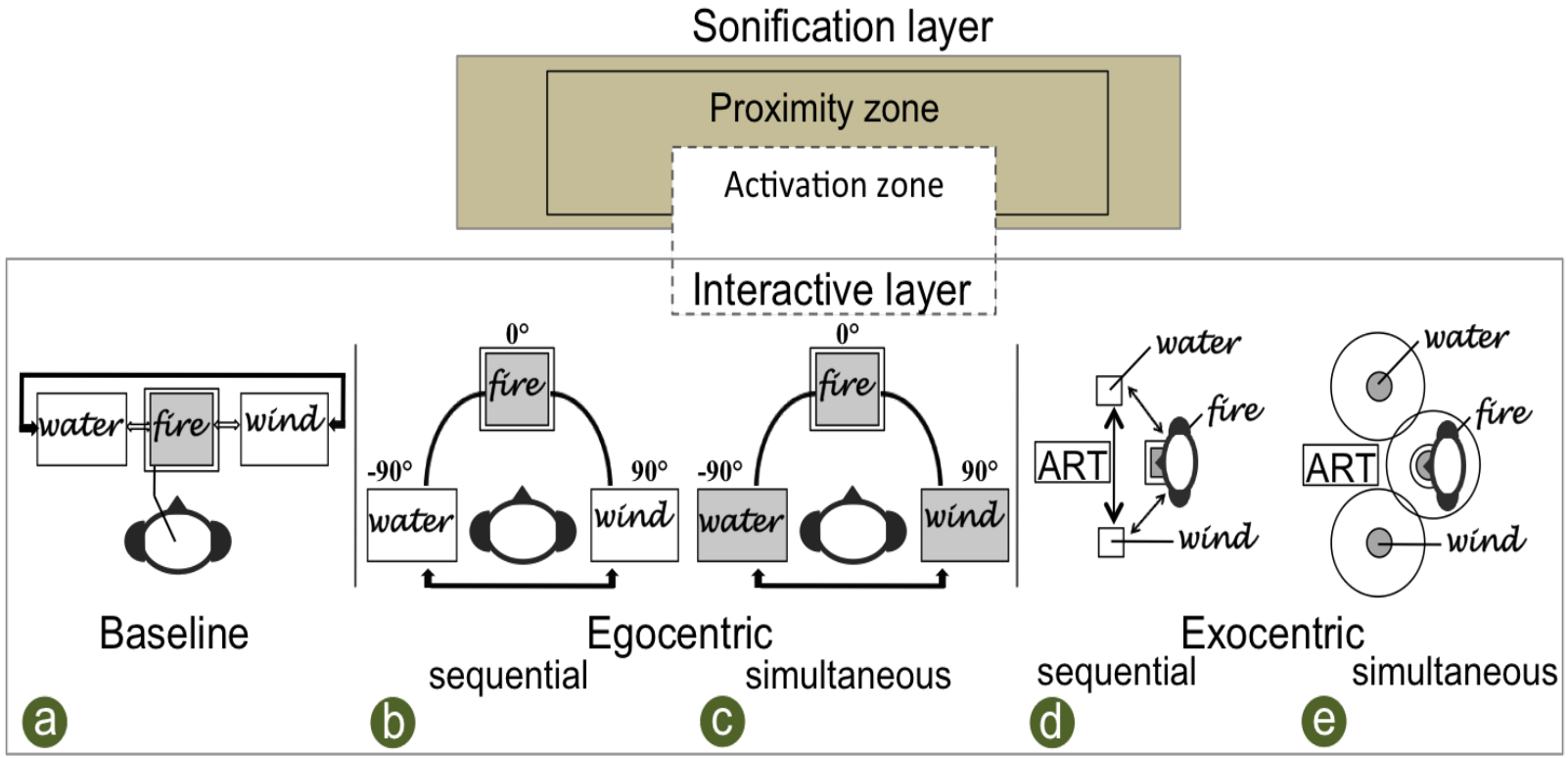


Figure 7.5: Schematic illustration of the multi-level auditory displays surrounding each art piece for the different experimental conditions. Each multi-level auditory display consisted of a top-level sonification layer and a secondary interactive layer. In the sonification layer, there was a proximity zone and an activation zone. The interactive layer could only be activated when the user was situated in the activation zone. The interface design tested in the interactive layer varied in complexity for the different experimental conditions (The greyed-out areas indicate the number of sound sources playing at one time): **a)**Baseline: non-spatialised audio menu items were played sequentially by pushing the navigation switch right or left for both the sequential and simultaneous presentation groups. **b)** Egocentric *sequential*: each audio menu was played sequentially from a location around the user's head at each navigation button push. **c)** Egocentric *simultaneous*: all audio menu items were played simultaneously and selecting one item using the navigation switch would increase the volume of the selected item and decrease the volume of the non-selected items. **d)** Exocentric *sequential*: audio menu items were situated in the exhibition space and perceived as if they were fixed to a location in the physical space. Selection of an audio item was performed by pushing the navigation switch independently of were the user was situated around the art piece. **e** Exocentric *simultaneous* all audio menu items were played simultaneously from a fixed location around the art piece. Circular proximity zones around each art piece enabled items to play simultaneously. To select and audio item, users walked around the art piece till the audio item was perceived as louder than the rest. This indicated the audio item was selected and could be activated.

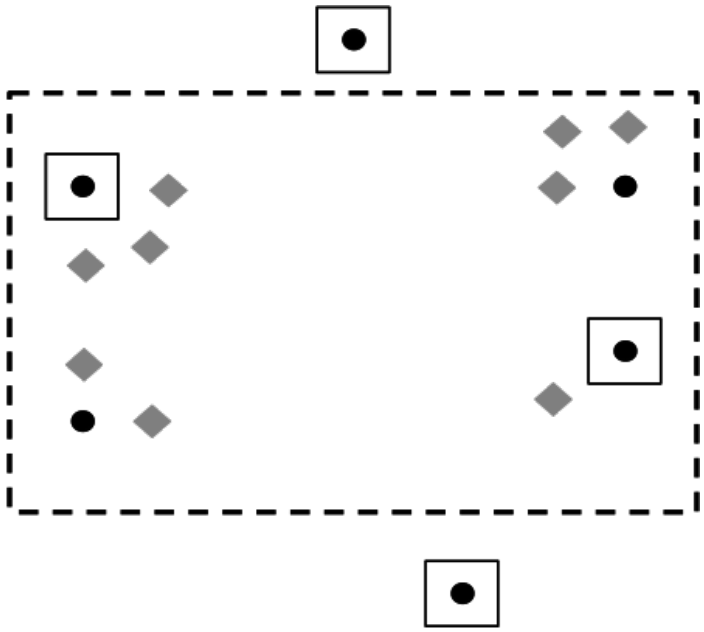


Figure 7.6: Illustration of the exhibition space identifying the location of the audio menu items for each of the art pieces in the Exocentric condition. The dashed-line area identifies the audio-augmented exhibition space measuring 3m(x) x 3.85m(y). The small squares with a dot at its centre identify the art pieces placed on tables and a dot alone the ones that hung from the ceiling. Grey diamonds identify the location of the audio menu items for each of the art pieces situated in the exhibition space.

played to indicate the presence of information at that location. As participants approached the art piece more closely, they were able to reach the activation zone in which they were able to activate the secondary interactive layer. Participants knew they had reached the interactive area when the chattering voices were louder and heard in both ears. To activate the interactive layer, the participant pressed down (long press, i.e. more than 2 secs) the navigation switch on the SHAKE sensor pack. Once activated, a number of available menu items could be browsed by pushing the navigation switch to the right or left. To select one of the menu items the navigation switch was pressed down (short press, i.e. less than 2 secs). Once the menu item was selected, its content was made available to the user. When the participant finished listening to the information available in the menu items, pressing down (long press) the navigation switch exited the interactive layer and the sound of the chattering voices in the top-level sonification auditory display was played again. The participant could then walk away and find the next audio-augmented artwork. In order to keep each experimental session within one hour duration, participants were

given a maximum of 10 minutes of exploration time for each test condition. There was no minimum time and participants could choose to stop whenever they wanted. All the participants had time to explore the art pieces in the allocated time. After each test condition, participants were asked to complete a NASA-TLX subjective workload assessment (see Appendix D) and a satisfaction questionnaire (see Appendix F.4) and also provide some informal feedback on their experience interacting with the system being tested in that condition. Once all three test conditions were completed, participants were instructed to provide feedback on how the different auditory interfaces tested in the exhibition space compared to each other. Finally, participants were invited to add an entry to a visitors' book especially created for the exhibition in which they could write any thoughts the information contained in the art pieces had provoked in them, if any.

Experiment hypotheses and metrics

In this experiment, two hypotheses were formulated:

i A consistent design (the same *exocentric* auditory display design in both the top-level sonification layer and the secondary interactive layer) in the multi-level audio display would follow the design principle of consistency (as discussed in Section 7.1) and reduce subjective workload and increase user satisfaction.

ii The use of 3D audio techniques in the secondary interactive layer of the multi-level auditory display will encourage an exploratory behaviour, which will result in significantly more time taken interacting with the system without a significant drop in user satisfaction or a significant increase in perceived workload (see Section 6.3 for a discussion on the assessment of exploratory behaviour).

In this evaluation user satisfaction and workload metrics (user experience) together with performance indicators were combined to assess the effectiveness of the interactive displays. The independent variable (IV) was the type of condition (the *Baseline* condition, the *Egocentric* condition and the *Exocentric* condition) per sequential and simultaneous presentation group, and the dependent variables (DVs) were a combination of subjective (level of user satisfaction and perceived subjective workload) and objective measures (time taken while interacting with the secondary interactive layer). In addition to participants' comments and opinions, user location coordinates and head orientation data were also collected for an in-depth analysis of participant behaviour.

The satisfaction questionnaire used in this experiment was a modified version (see Appendix F.4) of the one used in Wakkary and Hatala (2007) to evaluate the overall reaction to the system, the user interface, learning how to use the system, perceptions of the system's performance, the experience of the content, and degree of navigation and control. The questionnaire used in this study was modified to reflect the differences in the design of the system, in particular the user interface (questions on the SHAKE sensor pack and open-back headphones instead of the original interaction cube and wireless headset) and the content management (questions on the audio menus instead of the original audio preface). In addition, a question on the level of immersion experienced by the user was added to assess the overall reaction to the system. The inclusion of this question was motivated by the user feedback reported in the previous chapter, in which participants remarked on the level of immersion experienced when interacting with a fully spatialised system (see *Audio spatialisation: Spatial3D* in Section 6.2.3).

7.2.3 Results

User Experience

User Satisfaction The user satisfaction questionnaire included 62 questions (see Appendix F.4) and each question was rated on a continuum from "low" (1) to "high" (5) satisfaction. Questions were grouped into eight different subscales: *overall reaction to the system*, *user input interface*, *comfort level of headphones or headset*, *learning how to use the system*, *perceptions of the system's performance*, *quality of the content*, *audio experience*, and *degree of navigation and control*. Satisfaction mean scores were calculated from the participants' responses to the multiple questions contained in each subscale. These means scores could then be analysed using parametric statistics (Boone Jr and Boone, 2012). A statistical analysis is presented for the *overall reaction to the system* subscale alone, as no significant differences between the experimental conditions for the different presentation groups were found for the other seven subscales.

A two-way mixed-design ANOVA on the *overall reaction to the system* subscale mean scores with condition type as a within-subjects factor and presentation group as a between-subjects factor showed a significant interaction between condition type and presentation group ($F(2,60)=9.134$, $p<0.001$). No significant main effect was found for presentation group or condition type. *Post hoc* paired samples t-tests with Bonferroni correction for condition type showed that the satisfactory reaction was significantly higher for the Baseline condition (mean=4.10, SD=.49) ($t(15)=3.014$, $p<0.030$) and Egocentric condition (mean=3.99, SD=.56) ($t(15)=4.011$, $p<0.005$) than for the Exocentric condition (mean=3.56, SD=.73) in the simultaneous presentation group. No significant differences were found between the conditions for the sequential presentation

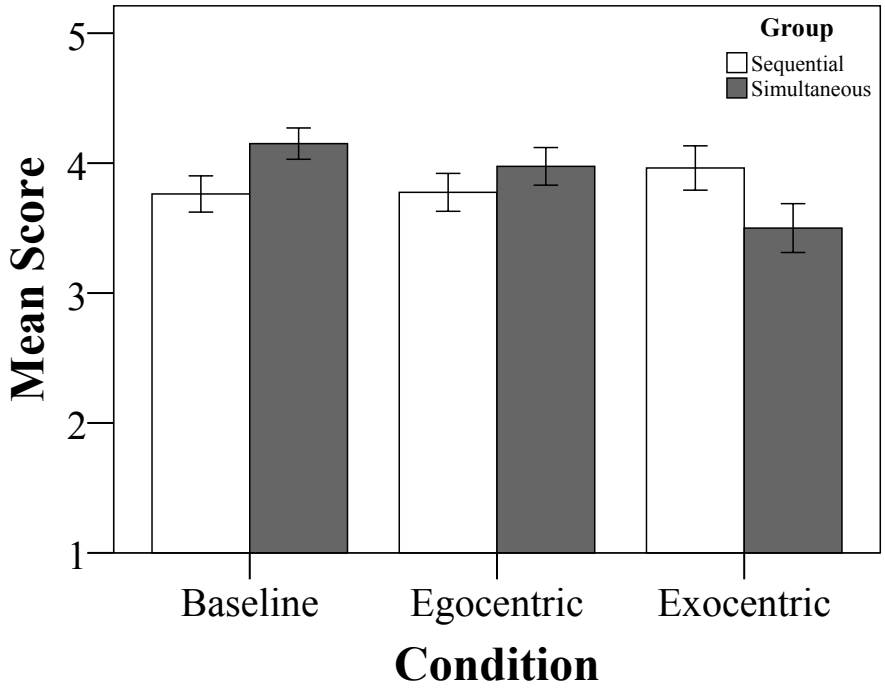


Figure 7.7: Mean scores for the 'overall reaction to the system' factor per condition and presentation group. Error bars show Standard Error of Mean ± 1.0.

group.

This result shows that users were less satisfied with the simultaneous exocentric design than the other designs. See Figure 7.7. Thus, the hypothesis that a consistent design across layers in a multi-level auditory display would increase user satisfaction was rejected, according to these results. However, the second hypothesis was partially confirmed, as the other spatial conditions did not show a significant drop in user satisfaction.

Overall Workload Raw overall workload means were calculated from the NASA-TLX questionnaire completed after each condition (see Figure 7.8). A two-way mixed-design ANOVA on overall workload with condition type as a within-subjects factor and presentation group as a between-subjects factor showed a significant main effect for condition type ($F(2,60)=4.606$, $p<0.015$). There was also an interaction between condition type and presentation group ($F(2,60)=4.672$, $p<0.015$). No significant difference was found between presentation groups. *Post hoc* Paired samples t-tests with Bonferroni correction per presentation group showed that overall workload was significantly higher for the Exocentric condition ($t(15)=-3.480$, $p<0.01$) (mean=

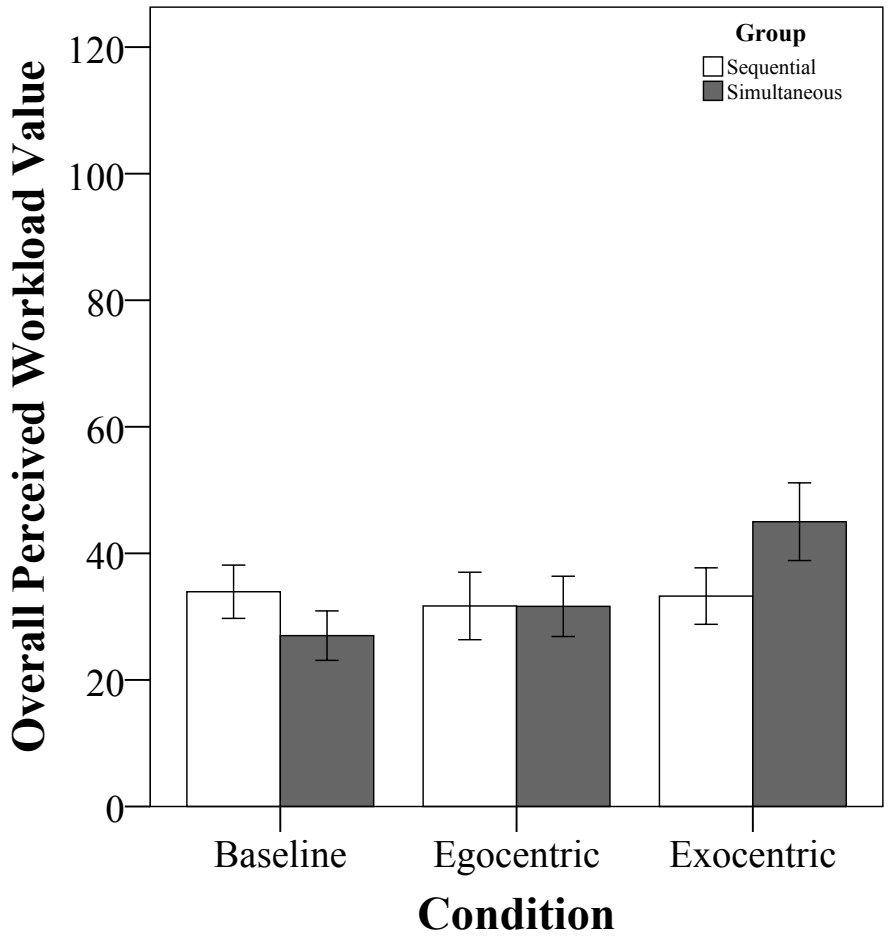


Figure 7.8: Overall perceived workload per condition and presentation group. Error bars show Standard Error of Mean ± 1.0.

45.00, SD=24.55) than the baseline (mean= 27.00, SD=15.62) and the Egocentric condition ($t(15)=-3.406$, $p<0.015$) (mean=31.63, SD=19.05) for the simultaneous presentation group. No significant differences were found between the conditions for the sequential presentation group.

These results show that workload was higher for the simultaneous exocentric design than for the other designs and supports the rejection of the hypothesis that a consistent design across layers in a multi-level auditory display would reduce subjective workload. However, the second hypothesis was partially confirmed, as the other spatial conditions did not show a significant increase in perceived workload.

User Performance

Time taken: interactive layer The total time participants spent interacting with the secondary interactive layer was also computed. See Figure 7.9. A two-way mixed-design ANOVA on time taken with condition type as a within-subjects factor and presentation group as a between-subjects factor showed a significant main effect for condition type ($F(2,60)=5.971$, $p<0.005$). No significant main effect was found for presentation group or interactions with condition type. *Post hoc* Pairwise Comparisons with Bonferroni correction for condition type showed participants spent significantly less time interacting with the auditory display in the Baseline conditions (mean= 249 secs, SD= 50) when compared to the spatial conditions (Egocentric: mean= 307 secs, SD= 109, $p<0.025$; Exocentric: mean= 322 secs, SD= 102, $p<0.005$).

These results show that the multi-level auditory displays designed with a spatialised secondary interactive layer encouraged users to spend longer interacting with the artwork and together with results from workload and user satisfaction confirms the second hypothesis formulated in this study.

User Feedback

Based on the user feedback collected after each condition was completed, twenty-nine participants out of the thirty-two that took part in this study found the experience enjoyable or interesting and they agreed that the provision of audio comments about the art pieces enhanced their experience. In addition, three participants reported that this experience had made them more likely to use an audio guide next time they visited a museum/gallery. In conditions where spatialised audio was used to present information sequentially, three participants described the interfaces as “thought provoking”. The occasional spatial audio latency problem affected the user experience in the spatial conditions but overall all participants enjoyed the idea of being able to walk around the space freely. Informal user feedback is presented for each of the three multi-level auditory display conditions.

Baseline In general, the interaction with this auditory display was described as “easy”, “enjoyable” and “most of all playful and entertaining”, with “informative” audio content. Two participants felt “more in control” when using this auditory display as the audio content was triggered by a simple press of a button in a sequential order. However, other participants reported that, although this display was “faster to use”, it was simply “less immersive” and “less fun” than the spatialised ones. One participant remarked: “I felt the experience was slightly less immersive and interacting with the menus less enjoyable. I felt a little less certain that I had

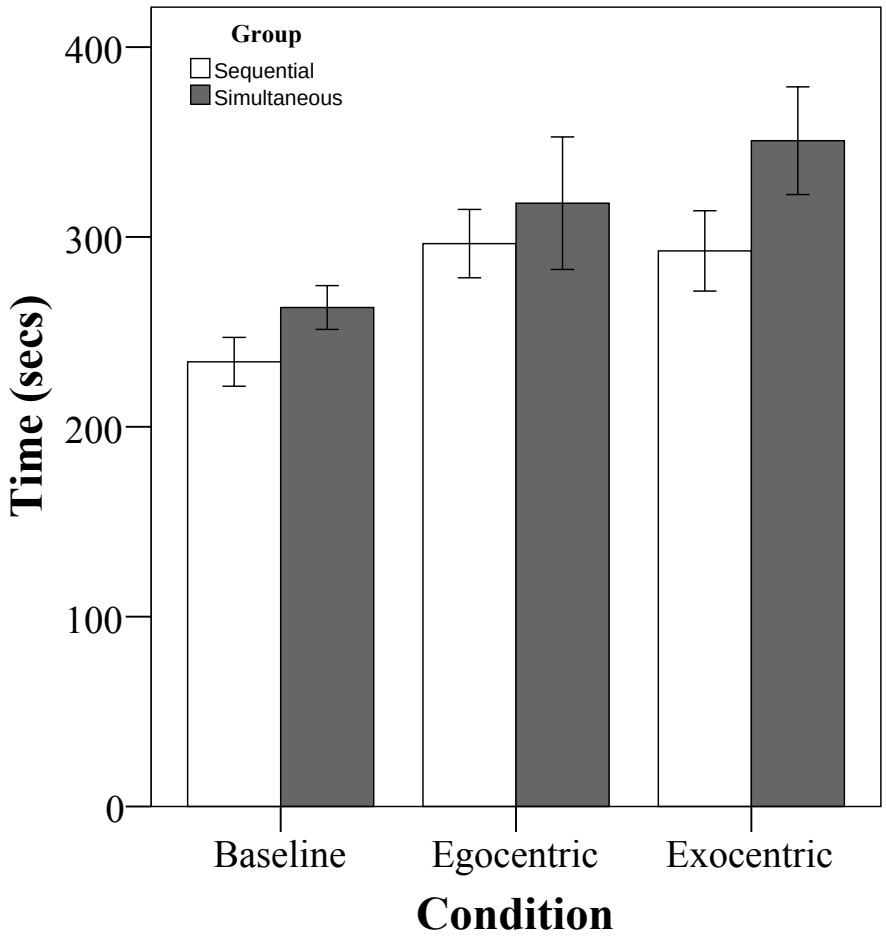


Figure 7.9: Mean time taken interacting with the secondary interactive layer per condition and presentation group. Error bars show Standard Error of Mean ± 1.0.

heard all the comments when they were not categorised hierarchically. However, I would say I put less mental effort into using the system, but it was closer to a simple accompaniment tape, which was much less interesting".

Egocentric The participants' experience of the Egocentric sequential condition, in which the audio menu items were spatialised around the users head and played one at a time when selected, was described as "novel", "fun", "enjoyable", "easy" and "informative". Although one participant reported that the spatialisation of the audio menu items did not affect the experience when using this auditory display, overall, the use of spatialisation had a positive impact in the user experience and interface usability. One participant felt that "having audio menu items separated

spatially made them easier to differentiate" and another suggested that "it gave a real life feeling, as if someone was indeed talking to me".

Having all the menu items play simultaneously in the Egocentric simultaneous condition was found to "enhance the experience" and to be "less mechanic and artificial". One participant reported: "I liked the fact that the sounds played simultaneously, that I had much more control over which of them I played. It was also much easier to confirm that I had listened to all that were available, and it was nice to be able to control the movement of the sound around your head." On the other hand, three participants remarked that playing the menu items simultaneously made it a little bit more difficult to remember to which option they had already listened to and consequently made the interaction more confusing. Two participants suggested that more training could offer a solution to enjoying the system more.

Exocentric Participants in the Exocentric sequential condition, in which audio menu items were perceived as if they were fixed to a location in the exhibition space and selected sequentially by the press of a button, described their experience of the auditory display as "enjoyable", "stimulating", "fun" and "informative" with one participant finding the concept "innovative". Having the auditory sources fixed to a physical location was reported to have "made the experience even more immersive" as "having the audio cues distributed around the artifact encouraged you to examine it from all sides". This interface was found to be "easy to use" and "quick to learn". One participant highlighted how "it was unique in the way the menu was in your head and you had to use your hearing to navigate through the menu".

Audio menu items in the Exocentric simultaneous condition were not only presented simultaneously but participants were also required to walk around the art pieces in order to locate and select a menu item. Some participants felt distracted by the need to move around more to locate the menu items and perhaps for that reason they felt unsure of whether they had found all the items around the art piece. One participant remarked: "it required a lot more physical activity going back and forth, focusing more on the commentary and it made me focus less on the artwork". Although there were participants that enjoyed having to move more and felt that it added to the playfulness, entertainment and "our awareness of space", the small size of the exhibition that resulted in menu items being closer together could still have a negative effect on the user experience. One participant reported: "Having the audio cues distributed around the artifact encouraged you to examine it from all sides, but finding the correct space to play certain cues was occasionally tricky". However, two participants remarked on the potential of this interface. One participant suggested that "if art pieces were much larger walking around to get menu items would make more sense" and another reckoned that "making you *search* for the audio information instead of it being available straight to you, I think this would be applied more to the public

exhibitions in numerous very new experiences [*sic*]".

User Behaviour

The logged data (including user location coordinates and head orientation data) showed a much simpler and shorter pattern of exploration for participants in the baseline and Egocentric conditions when compared to the Exocentric condition. Figure 7.10a shows an example of one of the participants in the simultaneous Baseline condition walking in a straighter trajectory between the art pieces and then staying mainly stationary once the secondary interactive layer was activated. Figure 7.10b shows the same participant taking more time to explore around the art pieces once the secondary interactive layer was activated in the Egocentric condition. However, Figure 7.10c shows that the same participant spent more time exploring both the top-level sonification and secondary interactive layers when in the Exocentric condition than when in the baseline or the Egocentric conditions. In contrast, a participant in the exocentric condition from the sequential group, Figure 7.10d, showed an exploratory behaviour while in the secondary interactive layer that resembled more the one found in the Egocentric condition (see Figure 7.10b) rather than the one found in the Exocentric simultaneous condition (see Figure 7.10c). This user behaviour suggests that spatial audio information encouraged participants to spend more time exploring the exhibition space. In addition, participants in the Exocentric condition also exhibited a *searching behaviour* (this searching behaviour is similar to that previously noted in Section 6.2.3 of the sound garden case study) altering their head position in order to determine the direction of the sound sources (see spline clusters in Figure 7.10c and d). Participants found this searching process most enjoyable when workload was low, as the user satisfaction for the Exocentric sequential condition showed; but less enjoyable under higher workload, as in the Exocentric simultaneous condition in which a greater amount of time was spent moving around as part of this searching process.

7.2.4 Discussion

In this study, a number of different multi-level auditory display designs that varied in complexity were evaluated as part of a non-guided mobile audio-augmented reality environment. Previous work on eyes-free interaction design has focused on evaluations of different mobile spatial audio designs in semi-controlled task-based assessments (Marentakis and Brewster, 2006; Brewster *et al.*, 2003), whereas work on mobile audio-augmented reality has mainly focused on the design of a unique auditory display to deliver location-based information as part of the main system implementation (Wakkary and Hatala, 2007; Eckel, 2001; Heller and Borchers, 2011). This

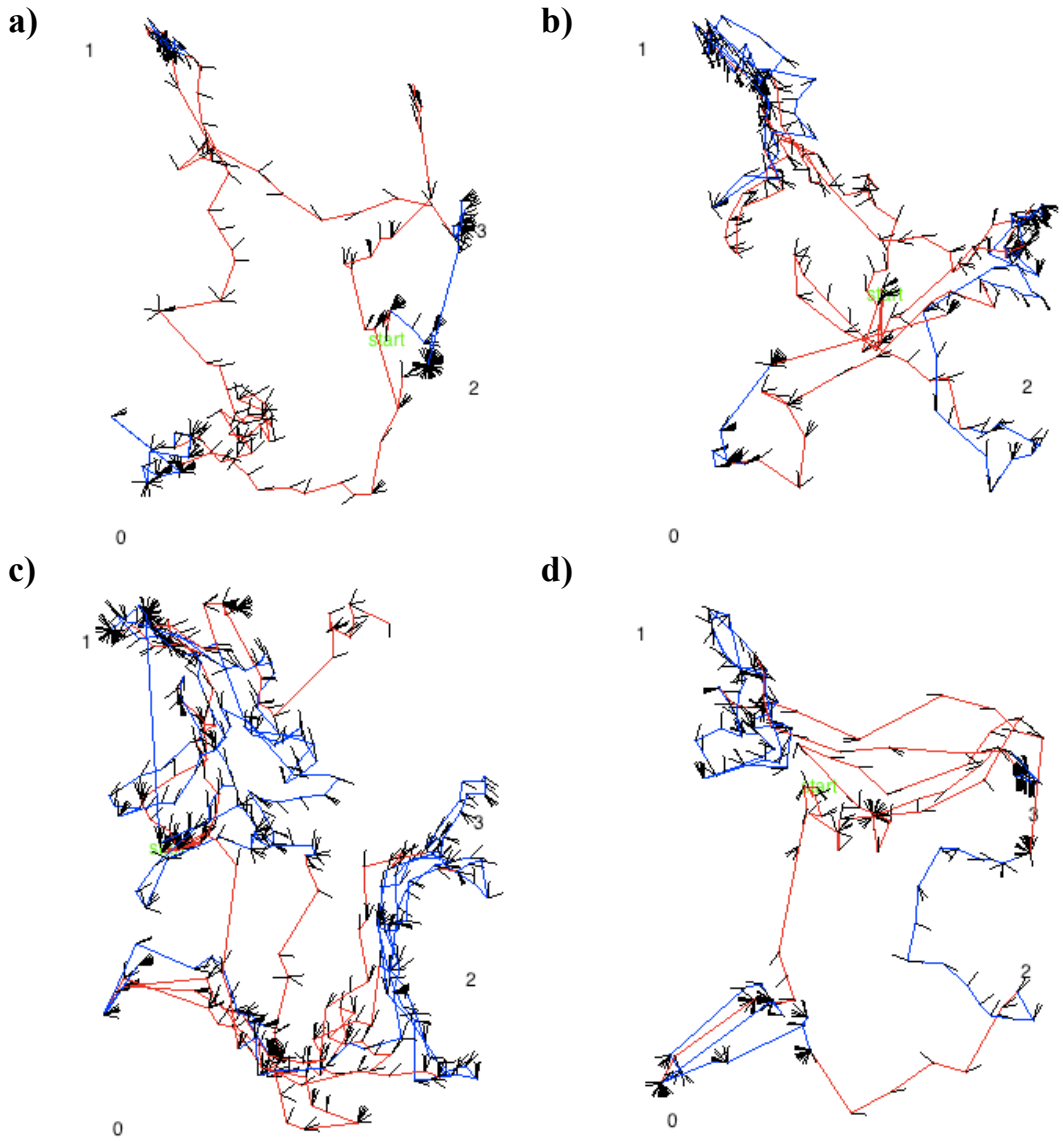


Figure 7.10: Route taken around the art pieces (0-3) by one participant from the *simultaneous* presentation group when in the a) Baseline b) Egocentric and c) Exocentric conditions, and the *sequential* presentation group when in the d) Exocentric condition. Solid red and blue lines illustrate the direction of exploration in the sonification layer and the interactive auditory display respectively. Short splines illustrate the participant's head direction approx. every half a second (mean=0.52 secs, SD=0.11).

study combined both a systematic assessment of mobile spatial auditory displays and location-based interactions within a mobile audio-augmented reality system to provide usability and user

behaviour information.

Results showed that consistency across the multi-level auditory display *did not* improve usability. The Exocentric simultaneous auditory display used an exocentric design that was consistent across levels. However, users felt under higher workload and were less satisfied with this auditory display. In this study, it was the appropriateness of the interface to the task, rather than the consistency of the auditory display design which was the key requirement. The top-level sonification layer was used for searching for audio-augmented locations but using the same exocentric design for browsing content in the secondary interactive layer was not acceptable for users.

The other three spatialised secondary interactive layers in the Exocentric sequential, Egocentric simultaneous, Egocentric sequential auditory displays were found to encourage users to spend longer interacting with the artwork without a drop in user satisfaction or increase in workload reflecting greater exploration. Changing from an exocentric to an egocentric perspective did not appear to confuse the users who were able to move smoothly between the different layers without an increase in workload, or drop in user satisfaction.

As hypothesised, informal feedback included in the *User feedback* section (see Section 7.2.3) suggests that the secondary spatialised interactive layer allowed for a more exploratory behaviour. Although, overall, the interface tested in the Baseline condition was reported as easy and faster to use, it was also found to be less immersive and less fun than the spatialised interfaces. Users liked the control over the interaction with the location-based information provided by the egocentric design and found that the exocentric design made the experience even more immersive. Performance results supported user feedback and helped characterise an exploratory behaviour as one where interaction times will increase without an increase in workload and a decrease of user satisfaction. In this way, the Baseline and the Exocentric display with simultaneous presentation did not encourage an exploratory behaviour. However, the other three spatialised auditory displays, including the Egocentric simultaneous display, did encourage an exploratory behaviour without a significant increase in workload. As one participant wrote in the visitors' book, the simultaneous presentation of "the menu items gave the whole experience a nice ambience and encouraged me to spend more time exploring the pieces."

Overall, the mobile audio-augmented reality environment implemented for this study provided a successful user experience. As one participant wrote in the visitors' book: "I enjoyed the idea of being able to move around a space and have the commentary adapt to me rather than the other way round. An altogether pleasant experience." In addition, this system was able to engage participants who mostly identified themselves as 'non-arty' by making the exhibit 'thought provoking'. As another participant wrote in the visitors' book: "the audio comments helped provoke thoughts and appreciate the exhibition in a way I usually wouldn't. As a person

who is not very 'arty' I spent more time looking at the pieces than I usually would".

Despite the successful implementation of this mobile audio-augmented reality prototype system, user location tracking still poses challenges. In this study the system had to be calibrated for each user due to their height having an effect on the positional accuracy of the system. A location tracking system that would allow for the removal of this anatomical dependency would further improve this system for deployment in the wild.

7.3 Conclusions

This chapter was devoted to studying the efficiency and usability of complex spatial auditory displays designed to enable user interactions with concentrated areas of information in a location-based system. The results presented in this chapter complement those discussed in Chapter 7 and together contribute to answering the second research question posed by this thesis: "How can 3D audio techniques be used to disambiguate multiple auditory sources in order to access location-based information in a mobile eyes-free interface?". An initial case study presented in the previous chapter showed that an interface combining 3D audio techniques, including variations in the amplitude and direction of the sources, and Earcons was the most effective auditory display when multiple audio landmarks overlapped in an exploratory mobile audio-augmented reality environment. It also put forward a framework for the analysis and description of user behaviour in such exploratory environments. This chapter sought to take these initial findings and methods further by investigating more complex interaction techniques.

The experiment presented in this chapter compared the users' experience and performance when interacting with a number of multi-level spatial auditory displays in an exploratory mobile audio-augmented reality environment. Multi-level displays enable the presentation of simultaneous auditory streams and allow the structuring of information in concentrated areas in a location-based system. Both egocentric and exocentric designs were combined in the multi-level auditory display to test whether a consistent design across levels would be preferred over a mixed-design and whether these 3D audio techniques would encourage an exploratory behaviour. The results showed that using a consistent exocentric design in the multi-level auditory display was not preferred. Also, by including a formal assessment of perceived workload and user satisfaction as part of the evaluation of user experience, it was possible to determine that a consistent exocentric design also failed to encourage an exploratory behaviour. However, the combination of a top-level exocentric configuration and an egocentric secondary configuration did encourage exploratory behaviour without overloading the user, even when auditory sources were presented simultaneously.

These results suggest that spatial audio encourages both an immersive experience and an exploratory behaviour but it is important to avoid overloading the user. Results also show that users can switch between egocentric and exocentric display types readily, so using the same configuration is less important than using an appropriate configuration for the task at hand. Informal feedback suggests that an interface allowing for simultaneous presentation can also be more immersive but such an interface should be very carefully designed as simultaneous presentation can increase workload.

These findings should allow designers to make more informed decisions when designing eyes-free auditory interfaces for mobile audio-augmented reality environments.

Chapter 8

Conclusions

8.1 Introduction

This thesis investigated to what extent 3D audio can be effectively incorporated into mobile auditory interfaces to offer users eyes-free interaction for both multitasking and accessing location-based information. A key contribution of this thesis has been a systematic evaluation of spatial audio techniques to determine their usability in an interactive mobile environment where we need to consider how to manage multiple auditory streams without overloading the user. This evaluation work has shown that spatial audio techniques, such as spatial minimisation and those used to design spatialised multi-level auditory displays, offer an effective means of presenting and interacting with multiple auditory streams simultaneously in an eyes-free mobile interface. However, the design of such interfaces is affected by attention demands, localisation error and subject preference.

This chapter provides a summary of the work reported in this thesis and relates the findings to the two research questions identified in Chapter 1:

RQ 1 To what extent can 3D audio techniques aid the user to maintain coherent attention on multiple auditory streams in a mobile eyes-free interface?

RQ 2 How can 3D audio techniques be used to disambiguate multiple auditory sources in order to access location-based information in a mobile eyes-free interface?

In addition, a set of guidelines are outlined for the design of eyes-free auditory interfaces for mobile applications. Finally, limitations of the research work carried out in this thesis and directions for future work are discussed and some general conclusions drawn.

8.2 Summary of the Thesis

This thesis began by introducing the motivation for this work, namely that spatial auditory interfaces can offer a solution to the problem of effective eyes-free mobile interactions, especially when such interfaces have to support multiple auditory streams. To interact effectively with an audio-only interface, users need to be able to direct their attention selectively to each individual auditory stream. By using spatial audio, the discrimination between the different streams can be achieved by placing each auditory source at a different location around the user's head, thus mirroring how humans perceive sounds in real life. Chapter 2 provided a background on sound and more specifically on spatial sound (or 3D audio). The chapter started with a brief account of the characteristics of sound that was followed by an overview of how the human auditory system is able to perceive and localise real sound sources and a review of current audio techniques that are able to model human sound localisation in order to create virtual audio environments. The chapter concluded that simulating spatial audio is still not as accurate as in the physical world, however, the use of spatial audio based on HRTF filters allow for an effective positioning of audio around the user and provide a solid basis for designing eyes-free spatial auditory interfaces. Chapter 3 started with a brief definition of a spatial auditory display and continued with a review of existing research on the design of spatial auditory displays, with special attention to those designed for eyes-free interaction. This review showed that the design of such auditory displays has varied enormously but little systematic user evaluation has been carried out to determine how these differences in design may affect their usability and the user experience, especially on mobile interfaces. Thus, in this chapter, a number of design considerations for both the presentation and the spatial arrangement of auditory sources in an auditory display were identified that would be then systematically evaluated in the experimental chapters of this thesis.

But, before mobile spatial auditory displays could be designed, a calibration of the 3D audio controls on the specific mobile platform to be used for the thesis was required. This is due to the significant differences in the implementation of spatial audio amongst different mobile platforms. Chapter 4 reported an evaluation of the positional 3D audio controls supported by the mobile device of choice for this work, the Nokia N95 8GB. It showed that the 3D audio capabilities on the Nokia device allowed users to identify unique targets at 45 intervals, so these location controls were considered appropriate for developing a 3D auditory interface.

Using the 3D audio controls successfully evaluated in Chapter 4, a number of eyes-free mobile spatial auditory interfaces supporting multiple audio presentation were designed and evaluated in Chapters 5-7. Chapter 5 investigated single level egocentric designs, Chapter 6 single level exocentric designs, while Chapter 7 examined a mixed multi-level spatial design.

Chapter 5 reported the design, implementation and evaluation of a number of different spatial (egocentric) and non-spatial audio techniques for supporting eyes-free mobile multitasking that included spatial minimisation. The efficiency and usability of these techniques was evaluated under varying cognitive load. The results of this evaluation showed an important interaction between cognitive load and the method used to present multiple auditory streams. The spatial minimisation technique offered an effective means of presenting and interacting with multiple auditory streams simultaneously in a selective-attention task (low cognitive load). However, spatialisation techniques were not as effective in a divided-attention task (high cognitive load), in which the interaction benefited significantly from the interruption of one of the stream. This chapter concluded that, given an appropriate task structure that minimises cognitive load, 3D audio techniques offer a means of designing effective auditory interfaces to support eyes-free mobile multitasking.

Chapters 6 and 7 investigated a location-based approach to supporting multiple information streams in a realistic eyes-free mobile environment. Chapter 6 reported an experimental case study which compared exocentric spatial auditory display designs, in which sounds are placed relative to the real world, to non-spatial designs. This study was conducted in an *outdoor* mobile audio-augmented exploratory environment that allowed for the analysis and description of user behaviour in a purely exploratory environment. A quantitative and qualitative analysis of the data gathered from this study showed that 3D audio was an effective technique to disambiguate multiple sound sources in a mobile exploratory environment. In addition, a combined analysis of the quantitative and qualitative data showed that 3D audio provided a more engaging and immersive experience and encouraged an exploratory behaviour. However, the exocentric spatial auditory displays tested in Chapter 6 only allowed for interactions with one location-based information item and interactions with multiple information were not investigated.

Chapter 7 extended the work of the previous chapter by evaluating a number of complex multi-level spatial auditory displays that enabled interaction with multiple location-based information. Multi-level displays enable the presentation of multiple location-based information by structuring the information found in concentrated areas of location-based systems. A top level exocentric structure, like the one successfully tested in Chapter 6, was used to advertise the existence of information at a specific location and a secondary interactive layer was used to interact with the actual information. The work reported in this chapter focused on how to design effective and usable multi-level spatial auditory displays. The experimental study was conducted in an *indoor* mobile audio-augmented exploratory environment that improved user tracking accuracy. A consistent exocentric design across levels was tested but it failed to reduce subjective workload or increase user satisfaction, so this design was widely rejected by users. On the other hand, the

rest of the spatial auditory displays tested encouraged an exploratory behaviour similar to that described in the previous chapter, here further characterised by increased user satisfaction and low perceived workload. In addition, users were able to switch between exocentric and egocentric designs in the multi-level auditory display, which suggests that using the same configuration is less important than using an appropriate configuration for the task at hand.

The spatial minimisation technique presented in Chapter 5 and the multi-level auditory display design introduced in Chapter 7 were both based on visual metaphors. The former on minimisation, where a visual display object is reduced and moved away from the user's focus of attention and the latter inspired by Zoomable User Interfaces (ZUIs), which are visual interfaces used to display large amounts of data in order to improve usability and reduce cognitive load. In this work, the minimisation metaphor was successful when supporting a selective-attention task but not a divided-attention task. Also, the ZUI metaphor presented in Chapter 7 did not result in a usable interface when a consistent design was used across levels.

8.3 Contributions of the Thesis

This thesis has presented the first systematic evaluation of 3D audio techniques used to design mobile auditory interfaces supporting eyes-free user interaction for both multitasking and accessing location-based information. In addition, this thesis has investigated the efficiency and usability of two different 3D audio techniques (i.e. *spatial minimisation* and *spatialised multi-level audio displays*). In particular, the design of the spatialised multi-level auditory displays included a novel combination of egocentric and exocentric techniques within the same audio interface that enabled the user to effectively access and manage location-based information. The novel contributions of this thesis will be outlined in further detail around the two research questions introduced in Chapter 1.

RQ 1: *To what extent can 3D audio techniques aid the user to maintain coherent attention on multiple auditory streams in a mobile eyes-free interface?*

The experimental work presented in this thesis has provided substantial evidence that 3D audio techniques aid users to maintain coherent attention on multiple auditory streams in a multitasking environment. In Chapter 5 two baselines, interruption and simultaneous presentation, were compared to two spatialised conditions. The extent spatialised audio could help in the presentation of multiple auditory streams was shown to depend on the extent the user was placed under cognitive load. In a divided-attention task, when users were placed under high cognitive load, users preferred streams to be interrupted than to be played simultaneously whatever simultaneous presentation method was applied. However, in a selective-attention task, when the user

was under less cognitive load, the minimisation technique was significantly preferred than in the divided-attention task, while preference for the other simultaneous conditions remained the same.

In Chapter 6, two spatialised approaches were used to present simultaneous overlapping Earcons. Although no direct comparison of *user attention* was possible in this experiment, a greater sense of immersion was reported in the fully spatialised condition. In Chapter 7, a greater sense of immersion was also reported by users in spatialised simultaneous conditions providing cognitive load was not high. However, as in Chapter 5, when users experienced higher cognitive load due to the searching behaviour required in the simultaneous exocentric condition, the spatialised design was rejected.

RQ 2: *How can 3D audio techniques be used to disambiguate multiple auditory sources in order to access location-based information in a mobile eyes-free interface?*

Chapter 6 and 7 concentrated on techniques to disambiguate multiple auditory sources in order to access location-based information. In the sound garden experiment presented in Chapter 6, four different auditory display designs were tested with varying use of Earcons to advertise location-based content and different degrees of spatialisation. In this experiment, simultaneous presentation was effectively used to add to the user experience. Users were able to track the virtual sound sources in order to find advertised content in the audio-augmented spaces, and found simultaneous presentation more immersive despite an increase in cognitive load. The fully spatialised exocentric design was found to be most effective for encouraging exploratory behaviour in the users.

The spatialised exocentric design in Chapter 6 was then extended in Chapter 7 with a secondary display that allowed users to interact with multiple items of audio content at each audio-augmented location. The top-level exocentric display introduced in Chapter 6 was shown to be effective for advertising content in a much smaller indoor environment. In addition, a set of different secondary interactive displays were evaluated. Using the same top-level exocentric simultaneous design for the secondary display was not effective due to the extra workload placed on users by the required searching behaviour. The rest of the secondary display designs were effective in that they allowed users to access the location-based information, however the spatialised designs also increased the sense of user immersion and encouraged exploratory behaviour. Ultimately, no negative results were found from users switching between a top-level exocentric display to a secondary egocentric interactive display.

8.4 Guidelines

The systematic evaluation of the 3D audio techniques developed in this thesis lead to the creation of a set of guidelines that are presented in this section.

8.4.1 Guidelines for the Design of Mobile Spatial Auditory Interfaces Supporting Multitasking

The experimental work presented in Chapter 5 lead to the following guidelines:

- When in a selective-attention task, spatial audio offers an efficient means of presenting concurrent auditory streams.
- When in a selective-attention task, spatial minimisation is able to maintain task efficiency when interacting with concurrent auditory streams.
- When in a divided-attention task, avoid the simultaneous presentation of auditory streams and always interrupt the auditory streams the user is not attending to.
- When in a divided-attention task, if concurrent presentation is required expect a drop in user performance and a significant increase in workload.

8.4.2 Guidelines for the Design of Mobile Spatial Auditory Interfaces Supporting Multiple Location-Based Information in Audio-Augmented Reality Environments

The experimental work presented in Chapters 6 and 7 lead to the following guidelines:

- Earcons are an effective technique for presenting multiple location-based information in an eyes-free spatial auditory display.
- When using a multi-level auditory display, using the same consistent configuration across levels is less important than using an appropriate configuration for the task at hand.
- Spatial audio in an exocentric auditory display can be used to encourage exploratory behaviour.
- Monitoring users head movement, as well as their position, helps understand users' exploratory behaviour.
- An exocentric auditory display operated by physical displacement increases user workload. Take care when using such a display in an interactive multi-level design.

8.5 Limitations and Future Work

8.5.1 Limitations

There are a number of limitations to the work presented in this thesis, which are related to technical constraints and the scope of the current research.

Throughout the experimental work presented in this thesis, all the auditory sources were limited to the horizontal plane around the user's head. Considering elevation in the design of egocentric displays was beyond the scope of this thesis. Furthermore, in the design of the exocentric displays it was not required, as all the audio-augmentation was limited to the X-Y plane. Including elevation in an spatial auditory display could offer greater flexibility and more complex design possibilities, however further baseline studies would be required to test its accuracy as part of an audio interface.

Another limitation in this thesis is related to the use of non-individualised HRTFs to position virtual sound sources around the user. Individualised HRTFs provide better localisation results as they are custom generated for each individual user but they are difficult to employ as the setup and equipment to acquire them is complex and very expensive. Non-individualised HRTFs provide worse localisation results but can be used by a much bigger number of users, which is the main reason why HRTF-based mobile phones use non-individualised HRTFs.

Connected to this decrease in localisation accuracy are the finding in Chapter 4 on earedness. In this chapter data were presented that suggested that right-eared users tend to perceive sources as being more central than users whose right ear is not dominant. In addition, our single left-handed user appeared to perform differently from the rest but without more data it is not possible to say if this was caused by left-handedness alone. In the light of these results it was decided that, for the remaining experiments presented in this thesis, participants would be screened for hand dominance and only right-handed users would be allowed to take part in the experiments.

Other technical limitations resulted in users having to wear a separate sensor device to make it possible to track head position and orientation. Usually, sensors for tracking orientation, such as a digital compass, come already integrated into mobile phone devices but this does not help greatly with tracking head movement. A more integrated approach in which the sensor device would be embedded in the headphones instead could be a solution. Also, user location tracking technology for both indoors and outdoors systems still poses challenges. In-built GPS receivers commonly found in mobile phones vary in their quality and accuracy and work poorly indoors. Audio-augmentation is sensitive to location inaccuracy. In the work presented in this thesis, only one type of GPS receiver was used. In addition, GPS error can also be affected by changes in the

number of satellites available in line of site. In the sound garden experiment reported in Chapter 6, a relatively open vista minimised these effects, but in more built up areas this could become a more serious problem.

Indoor user tracking is a less mature technology than outdoor GPS tracking and the system presented in Chapter 7 had a number of important limitations. It needed to be calibrated for each subject because of height variation and a separate IR emitter and head orientation device was required. The fact that only one participant could be tracked at any time, and the multiple devices required made the system inappropriate for a less controlled environment. Furthermore, a single camera limited the available tracking space. In order to track subjects in a broader space the use of multiple cameras would be required.

8.5.2 Future Work

Individual user differences

For auditory displays to become more mainstream interfaces in mobile devices, the issues of user differences in 3D audio localisation ability need to be addressed. Firstly, anatomical differences mean that systems using non-individualised HRTFs perform differently for different users. There is scope for systems to adapt to a user or be calibrated more easily by a user. For example, a set of headphones could adapt to the user's head width, or a system could analyse a photographic image of a user's pinnae, altering the HRTFs accordingly (see work by University of Sidney and University of York (2013) on mapping HRTFs to Ear Morphology). These approaches would not address non-anatomical individual variation in 3D audio perception, such as that caused by ear dominance. In this case, either some type of smart user adaptation would be required or some sort of initial calibration. Some systems already offer users a choice between several HRTFs with a simple calibration exercise to help choose the most appropriate (Papa Sangre, 2012).

However, the extent such a calibration phase could improve an auditory display is unclear. Should interfaces be designed to be resilient to individual variation or should they be designed to take user variation into account? Future work is required to deal with these issues, especially if auditory displays become more complex and more widespread.

Animated auditory streams

3D audio techniques do not just allow for a sound source to be positioned in a 3D space, they also allow for the sound source to be moved in this space. Moving sound sources, like animation in visual displays, could be used to bring auditory displays alive. Such movement could be used to add richness to both egocentric and exocentric displays. Further baseline work is required to investigate how such animation should be incorporated into spatial audio interface design.

Robust user tracking

In an exocentric display, user position together with real-time updating of the sound sources relative to the user orientation make users perceive the sound sources as being fixed to the physical space. Critical to the accuracy, immersiveness and believability of such interfaces is the precision and responsiveness of user location and head orientation tracking.

Recent work in computer vision (Eichner *et al.*, 2012) looking at human pose estimation could be applied to this problem. Potentially, such a system would be able to detect both user location and user head orientation for more than one user in a space equipped with multiple cameras. Such a system would also offer the advantage that no additional sensors would be required other than a user's mobile phone. These advantages would allow for deployment in a semi-public space, such as an art gallery or supermarket.

8.6 Final Remarks

In 2011 more smart phones than PCs were shipped (Engadget, 2012). Ubiquitous computing is a reality that modern interface design has to contend with. The experimental work presented in this thesis has demonstrated that spatial auditory interfaces offer a practical solution to a number of eye-free mobile interface design challenges. In addition, this work has shown that complex 3D auditory interfaces contribute to an immersive and playful user experience. As auditory interfaces become more widely incorporated into devices and applications, the baseline work and design guidelines presented here could be used to make such systems fun, effective and compelling for users.

Appendix A

Experimental Stimuli: DVD

The audio files used in the experimental chapters of this thesis are included in separate electronic folders. In the printed version of this thesis, these folders are included in a DVD. See below for a list of folders.

- *A 1. 3D Audio Controls Evaluation Stimuli (Chapter 4)*
- *A 2. Spatial Auditory Interfaces for Eyes-Free Multitasking Stimuli (Chapter 5)*
- *A 3. Sound Garden Stimuli (Chapter 6)*
- *A 4. Audio-Augmented Conceptual Art Exhibition Stimuli (Chapter 7)*

Appendix B

Experimental Groundwork Study - Instructions for Participants

Evaluation of spatial audio localization

Introduction

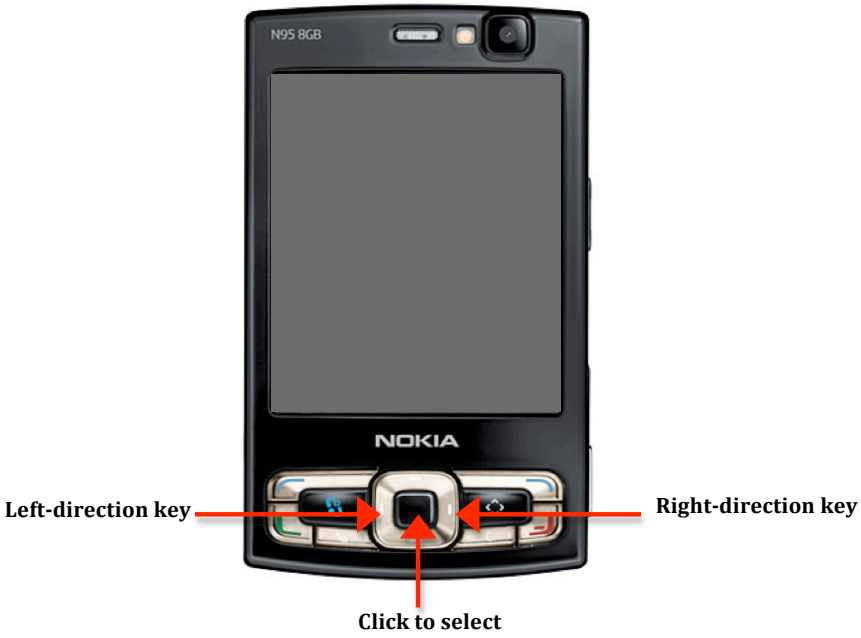
A spatial sound is a synthesized sound that is experienced as a virtual source at a desired location in space. The aim of this study is to test the user's capability to localize a spatial sound when no visual feedback is present. We will test this in two different conditions using two different types of sounds: pink noise and speech. You will be presented with two identical sounds, one that will be static and another that you will be able to move. They will be repeated one after the other. For each trial, you will be asked to adjust the direction of the sound you can move till it matches that of the static sound. You will adjust the sound direction by pressing on the **left** and **right keys** on the direction pad on the phone. These adjustments will be carried out as quickly and accurately as possible. Once the adjustment is completed, press on the **central navigation button** to make a selection.

Important

It is extremely important that you alert the experimenter if you feel any discomfort during the experiment. If you feel uncomfortable, please tell the experimenter immediately.

The Task

The experiment will be run on a Nokia N95 8GB phone shown below



You will hold the device in the upright position in your hand. No visual feedback will be displayed on the screen of the device. You will then control the direction of the sound. You can move it by pressing the right direction key or the left direction key. When you are happy that the two sounds match the same location, press central button to confirm your selection. Continue until no more trials are left. This will be notified to you by returning to the main menu. The experiment consists of two blocks of 15 trials.

It is important that you are comfortable during the length of the experiment. A rest will be provided between the blocks, but if you need any assistance, please alert the experimenter.

Appendix C

Eyes-free Multitasking Study - Experimental Instructions, Recall Questions and Preference forms

C.1 Divided-Attention group

C.1.1 Instructions for Participants

Foreground/Background audio interactions

Introduction

The aim of this study is to examine how easy it is to deal with the issue of presenting concurrent audio streams. You will be asked to perform a number of tasks while listening to a podcast:

1. Find out what is the *next audio track title*.
2. Check *appointments for Tuesday*.
3. Find out the *current time*.

Important

It is extremely important that you alert the experimenter if you feel any discomfort during the experiment. If you feel uncomfortable, please tell the experimenter immediately.

The Task

The experiment will be run on a Nokia N95 8GB phone shown below

Central navigation button: click to proceed to menu task

Menu up-direction key

Menu down-direction key

The experiment consists of four blocks with 3 tasks per block starting with two training sessions. In the first training session you will get familiar with the audio menu and will be given the time to practice the tasks till you are comfortable browsing the menu for information. The second training session will present you with a podcast and will allow you to practice interacting with the audio menu while listening to the podcast.

During each of the four blocks you will hold the device in your hand in the orientation shown above. No visual feedback will be displayed on the screen of the device. When the experiment starts a podcast will start playing. While listening to the podcast a BEEP sound will prompt you to start the tasks. To start the tasks press the **central navigation button** to listen to the audio menu. Perform one task after the other from this point in your own time. To navigate the audio menu, use: Up/Down/Right/Left-direction keys.

Once each block is completed, you will be asked to:

- Fill in a listening comprehension exercise on the information delivered in each task and on specific details of the podcast.
- Complete a NASA-TLX form.

It is important that you are comfortable during the length of the experiment. A rest will be provided between the blocks, but if you need any assistance, please alert the experimenter.

C.1.2 Recall Questions

Training. *Listening Comprehension*

1. Which city is the story about?

2. Due to which endangered species construction has been banned on the beaches?

3. In what year was construction of further buildings banned for fear of damaging archeological evidence?

4. What colour is the Ely hotel painted in?

5. What is the name of the moderate Islamic party?

6. What was the current time?

Condition 1. *Listening Comprehension*

1. Which country encouraged people to vote in the European elections showing a film with a woman screaming and a man with a blood-stained ax?

2. According to the correspondent, what do nightmares wear?

3. Which party do Nicola Benoit and Manu Garcia support?

4. What is the name of Nicola Benoit and Manu Garcia's company?

5. Which American short distance runner were the strikers compared to?

6. What was the appointment for Tuesday?

Condition 2. *Listening Comprehension*

1. In which city was the correspondent living in 1996?

2. Waiting up for who was the same as being allowed to stay up late to watch Eurovision when the correspondent was a little boy?

3. What was the name of the presenter for the 'Song of Europe' contest?

4. What was the name of Russia's first ever Eurovision song?

5. In what year did four countries tie for the first place in the Eurovision contest?

6. What was the current time?

Condition 3. *Listening Comprehension*

1. What is the surname of the person addressing the first afternoon rally the correspondent attends?

2. What did the correspondent stick in his ear?

3. Which gate number does the correspondent use to get into the rally?

4. What is the surname of the former Tamil movie superstar?

5. What is the name of the correspondent's cameraman?

6. What was the next track song title?

Condition 4. *Listening Comprehension*

1. What is the name of Texan George Bush senior's envoy?

2. The summer of what year was a decisive moment for the American Middle East diplomacy?

3. On what was the message printed, "We welcome the summit and the prospects for peace"?

4. What tragedy happens on the very same day the summit ends?

5. What is the name of the hugely influential pro-Israel lobby?

6. What was the appointment for Tuesday?

C.1.3 Order of Preference Form

Name	Task	Date

Out of the four different ways you used to interact with the audio menus while listening to the podcast, which one did you like the most?

PLEASE, enter number in the box in order of preference.

Condition 1. ☐

Both podcast and audio menu were playing simultaneously and heard in both ears.

Condition 2. ☐

Podcast was interrupted when interacting with the audio menu.

Condition 3. ☐

Podcast was only heard in the right ear for the whole duration of the condition.

Condition 4. ☐

Podcast was moved to the right ear only when interacting with the audio menu.

C.2 Selective-Attention group

C.2.1 Instructions for Participants

Foreground/Background audio interactions

Introduction

The aim of this study is to examine how easy it is to deal with the issue of presenting concurrent audio streams. You will be asked to perform a number of tasks while listening to a piece of classical music:

1. Find out what is the *next audio track title*.
2. Check *appointments for Tuesday*.
3. Find out the *current time*.

Important

It is extremely important that you alert the experimenter if you feel any discomfort during the experiment. If you feel uncomfortable, please tell the experimenter immediately.

The Task

The experiment will be run on a Nokia N95 8GB phone shown below

Central navigation button: click to proceed to menu task

Menu up-direction key

Menu down-direction key

The experiment consists of five blocks with 3 tasks per block starting with two training sessions. In the first training session you will get familiar with the audio menu and will be given the time to practice the tasks till you are comfortable browsing the menu for information. The second training session will present you with a piece of classical music and will allow you to practice interacting with the audio menu while listening to the music.

During each of the five blocks, you will hold the device in your hand in the orientation shown above. No visual feedback will be displayed on the screen of the device. When the experiment starts a piece of classical music will start playing. While listening to the music a BEEP sound will prompt you to start the tasks. To start the tasks press the **central navigation button** to listen to the audio menu. Perform one task after the other from this point in your own time. To navigate the audio menu, use: Up/Down/Right/Left-direction keys.

Once each block is completed, you will be asked to:

- Fill in a listening comprehension exercise on the information delivered in each task.
- Complete a NASA-TLX form.

It is important that you are comfortable during the length of the experiment. A rest will be provided between the blocks, but if you need any assistance, please alert the experimenter.

C.2.2 Recall Questions

Training. *Listening Comprehension*

1. What was the current time?

Condition 1. *Listening Comprehension*

1. What was the appointment for Tuesday?

Condition 2. *Listening Comprehension*

1. What was the current time?

Condition 3. *Listening Comprehension*

1. What was the next track song title?

Condition 4. *Listening Comprehension*

1. What was the appointment for Tuesday?

C.2.3 Order of Preference Form

Name	Task	Date

Out of the four different ways you used to interact with the audio menus while listening to the podcast, which one did you like the most?

PLEASE, enter number in the box in order of preference.

Condition 1. ☐

Both music and audio menu were playing simultaneously and heard in both ears.

Condition 2. ☐

Music was interrupted when interacting with the audio menu.

Condition 3. ☐

Music was only heard in the right ear for the whole duration of the condition.

Condition 4. ☐

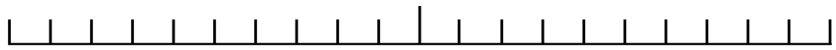
Music was moved to the right ear only when interacting with the audio menu.

Appendix D

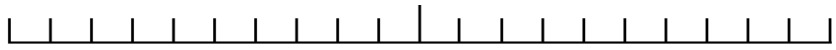
NASA-TLX Questionnaire Form

Name	Task	Date

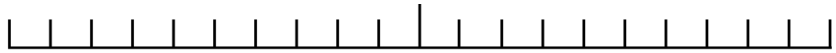
Mental Demand — How much mental and auditory activity was required?



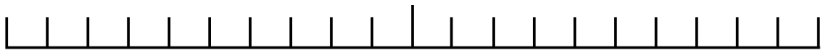
Physical Demand — How much physical activity was required?



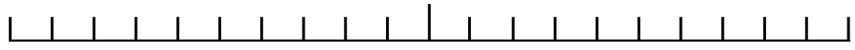
Temporal Demand — How hurried or rushed did you feel while performing the tasks?



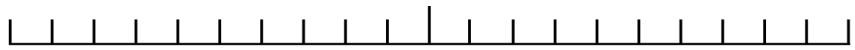
Performance — How successful do you think you were at accomplishing the task set by the experimenter?



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



Frustration — How much frustration did you experience? (e.g. were you relaxed, content, stressed, irritated, discouraged)



Appendix E

Outdoor Mobile Audio-Augmented Reality Study - Experimental Instructions, Think Aloud Sheet and User Questionnaire

E.1 Instructions for Participants

Exploring spatial audio in the sound garden

Introduction

The aim of this study is to determine how easy it is to localize sound sources anchored to a number of different locations in a sound garden, i.e. Jardim Municipal, Funchal. You will be asked to walk around the park at your own leisure (the experimenter will accompany you at all times) following the different paths already existing in this garden holding a mobile phone in your hand.

Important

It is extremely important that you alert the experimenter if you feel any discomfort during the experiment. If you feel uncomfortable, please tell the experimenter immediately.

The Task

The experiment will be run on a Nokia N95 8GB phone shown below

Central navigation button: press to play message.

You will hold the device in this orientation in your hand. After starting the sound garden application and listening to the welcoming message by the Café do Teatro, the experimenter will walk with you to the Jardim Municipal and follow you while you walk along the different stone paths in the garden in your own time. You will have 30 minutes to find a number of audio messages planted in this garden. Do not rush though and do not walk too fast as the GPS will not be able to update your position if you walk at a fast pace. In some conditions the messages will be triggered as you walk near them and in others you will hear an animal sound, as you get closer. In the last case, you will need to press the central navigation button on the mobile phone provided to hear the message planted at that particular position. You will know whether you have arrived at the location where the sound has been planted because the sounds will be heard in both your ears.

Please give feedback to the experimenter of your experience WHILE you are walking around the garden. Any impressions you have or anything you would like to comment on as you walk and immerse yourself in this experience. The experimenter will take note of everything you say.

Once the 30 minutes are over, the experiment will be over and the experimenter will go through a few questions with you and will also ask you to provide some informal feedback about your experience in the sound garden.

It is important that you are comfortable during the entire length of this study. A rest will be provided if requested, but if you need any assistance, please alert the experimenter.

E.2 Sample Think Aloud Note Taking Sheet

Sample Think Aloud Note Taking Sheet

Task	Time	Time taken	Observations	Criteria

E.3 User Questionnaire

Participant's name: _____

Date: _____ Time: _____

Sex: Male ____ Female ____

Expertise with GPS-based applications: YES _____ NO _____

Questionnaire:

1. Overall, how much did you enjoy this sound garden experience?

   Not much 1 2 3 4 5 Very much

   Feedback:

   _____
   _____
   _____

2. How easy was it to navigate the garden to find the sounds/audio messages?

   Very easy 1 2 3 4 5 Very difficult

   Feedback:

   _____
   _____
   _____

3. What did you think of the overall quality of the sounds?

   Very bad 1 2 3 4 5 Very good

   Feedback:

   _____
   _____
   _____

4. To what extent the use of Speech Synthesis to reproduce the audio messages negatively affected your experience of the sound garden?

   Very negative 1 2 3 4 5 Very positive

Feedback:

_____
_____
_____

5. IF the sounds were spatialised, how difficult was it to identify the position of the sound?

   Very difficult 1 2 3 4 5 Very Easy -NA

   Feedback:

   _____
   _____
   _____

6. IF the sounds were spatialised, was the distance effect from sound appropriate?

   Not appropriate 1 2 3 4 5 appropriate -NA

   Feedback:

   _____
   _____
   _____

7. IF the sounds were spatialised, was the spatial presentation of sound helpful?

   Not helpful 1 2 3 4 5 Very helpful -NA

   Feedback:

   _____
   _____
   _____

8. How useful were the animal noises in helping find the information in the sound garden?

   Not useful 1 2 3 4 5 Very useful -NA

   Feedback:

   _____
   _____
   _____

9. If you could choose the information delivered to you when targeting a sound, what kind of audio message would you like to hear in a context like this, i.e. sound garden? Would you say the ones used in this study were appropriate?

   _____
   _____
   _____
   _____
   _____
   _____

10. If there was something you could change in the sound experience you just had, what would that be?

    _____
    _____
    _____
    _____
    _____
    _____

11. Overall, what would you highlight from this sound garden experience?

    _____
    _____
    _____
    _____
    _____
    _____

Appendix F

Indoor Mobile Audio-Augmented Reality Study - Experimental Materials

F.1 Instructions for Participants

Designing mobile spatial audio interfaces for exploration indoors

Introduction

The aim of this study is to assess the effectiveness and usability of spatial audio cues when supporting user interactions with location-based audio targets indoors while mobile. You will be asked to explore a room in which various artworks are being exhibited and, for each artwork, interact with the audio information provided.

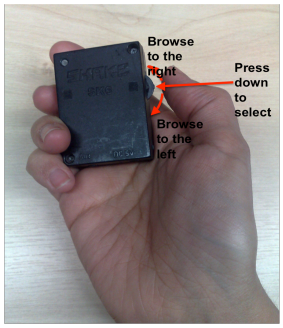
Important

It is extremely important that you alert the experimenter if you feel any discomfort during the experiment. If you feel uncomfortable, please tell the experimenter immediately.

The task

The experiment will be run on a Nokia N95 8GB phone and you will be asked to wear a pair of headphones. A sensor device to track your head movements and an infrared tag to track your location will be mounted on the headphones. The experimenter will provide you with the mobile device and a SHAKE sensor pack (see an illustration below). Place the phone around your neck using the lanyard provided and carry the SHAKE sensor pack in your hand. Please, once the experiment starts, keep the mobile device out of sight as no visual feedback will be provided. At the end of each block, the experimenter will ask you to hand it over.

The experiment consists of three blocks starting with a short calibration of the space and an initial training session per block. In the training session you will get familiar with exploring the exhibition room and will be given the time to practice with audio information related to one of the artworks. During each of the three blocks you will be asked to explore the exhibition room and the artwork within it. As you get closer to the artwork, audio will be triggered indicating the presence of information about the artwork at that location. As you approach the artwork more closely, you will reach the area in which you will be able to interact with the audio information. You will know you have reached the interactive area when the audio is louder and is heard in both your ears. At this point, you can start interacting with the audio using the navigation switch on the SHAKE sensor pack (see illustration below). Push the wheel to the right or left to indicate in which direction you want to browse the available audio items, or press down to select. A short press will *select* an audio item and a long press will **de-select** and exit the audio item. You will have a total of 10 minutes to explore the room during each block.



- **short press** SELECTS
- **long press** DE-SELECTS

At the end of each block, you will be asked to complete a NASA-TLX form and a satisfaction questionnaire. In addition, once all three blocks are completed, the experiment will be over and the experimenter will ask you to add an entry to the exhibition's visitors' book. Please, provide your informal comments on your response to the art displayed in the exhibition room (min. 20 words).

It is important that you are comfortable during the length of the experiment. A rest will be provided between the blocks, but if you need any assistance, please alert the experimenter.

F.2 Introduction to the Conceptual Art Exhibition

Weaving the City started some years ago, as a series of ephemeral interventions in urban space, which aimed to *physically connect* locations through woollen threads. The project, deals with the relational space between human bodies and urban surfaces, and thus, highlights, the undefined character of the threshold *between* private space, the SKIN, and public realm, the CITY. The project, investigates the *textural qualities* of surfaces and connections, the *perceptual experience* of being: in the street, indoors, in an audiovisual installation, or in an online space. The pieces are temporal, just like sound, and although *repeated*, they entail variation.

The locations of weaves and installations are documented and published online, becoming digital tags and references to the past. The temporal weaves *are* available on a more abstract and intangible weave: the NET. The most recent pieces are paper-boxes pierced and woven with wool and threads, onto which I project audiovisuals of old weaves and installations that have *long ago* disappeared. These paper boxes are portable, miniaturised installation rooms, *ephemeral and fragile*, designed to be placed outdoors, and to degrade over time as a consequence of weather conditions, and the *natural process of* existence and disappearance, which draws on Heidegger's concept of the *circularity of being*.

F.3 Audio Interface Descriptions

This section contains the descriptions provided to participants of the audio interfaces tested in Chapter 7. These descriptions are presented here per experimental condition.

Baseline

As you walk around the exhibition space, when you get closer to an artwork that has been audio-augmented, you will listen to some chattering voices. This sound indicates that you can access extra audio information about the artwork. In order to play this extra information, walk to the place where the chattering voices are at their loudest (loud in both your ears) and push (short-press) the SHAKE sensor pack navigation switch. This will play the audio menu sounds available for that artwork. The audio menu sounds are: *water waves* identifying information provided by the artist, *open crackling fire* identifying positive comments left by previous visitors and *stormy wind* identifying negative comments left by previous visitors. **Notice that not all artworks have been audio-augmented and individual artworks may have different numbers of audio menu sounds**.

The presentation of the audio menu sounds in this condition will be sequential (one at a time). In order to browse the audio menu items push the SHAKE sensor pack navigation switch right or left. If you reach the end of the audio menu items they will loop around. To listen to the artwork information, short-press when listening to the targeted audio menu item and the information will start playing. Once the information has finished playing, the system automatically goes back to the audio menu items. You can stop the information from playing at any time with a long-press. To exit the audio menu items long-press again. You will then be back to listening to the chattering voices.

Egocentric Sequential

As you walk around the exhibition space, when you get closer to an artwork that has been audio-augmented, you will listen to some chattering voices. This sound indicates that you can access extra audio information about the artwork. In order to play this extra information, walk to the place where the chattering voices are at their loudest (loud in both your ears) and push (short-press) the SHAKE sensor pack navigation switch. This will play the audio menu sounds available for that artwork. The audio menu sounds are: *water waves* identifying information provided by the artist, *open crackling fire* identifying positive comments left by previous visitors and *stormy wind* identifying negative comments left by previous visitors. **Notice that not all artworks have been audio-augmented and individual artworks may have different numbers of audio menu sounds**.

The presentation of the audio menu sounds in this condition will be sequential (one at a time) and spatialised around your head, i.e. perceived as originating from the right, left or in front of you. In order to browse the audio menu items push the SHAKE sensor pack navigation switch right or left. If you reach the end of the audio menu items they will loop around. To listen to the artwork information, short-press when listening to the targeted audio menu item and the information will start playing. Once the information has finished playing, the system automatically goes back to the audio menu items. You can stop the information from playing at any time with a long-press. To exit the audio menu items long-press again. You will then be back to listening to the chattering voices.

Exocentric Sequential

As you walk around the exhibition space, when you get closer to an artwork that has been audio-augmented, you will listen to some chattering voices. This sound indicates that you can access extra audio information about the artwork. In order to play this extra information, walk to the place where the chattering voices are at their loudest (loud in both your ears) and push

(short-press) the SHAKE sensor pack navigation switch. This will play the audio menu sounds available for that artwork. The audio menu sounds are: *water waves* identifying information provided by the artist, *open crackling fire* identifying positive comments left by previous visitors and *stormy wind* identifying negative comments left by previous visitors. **Notice that not all artworks have been audio-augmented and individual artworks may have different numbers of audio menu sounds**.

The presentation of the audio menu sounds in this condition will be sequential (one at a time) and audio menu items will be perceived as if they were fixed to a location. In order to browse the audio menu items push the SHAKE sensor pack navigation switch right or left. If you reach the end of the audio menu items they will loop around. To listen to the artwork information short-press when listening to the targeted audio menu item and the information will start playing. Once the information has finished playing, the system automatically goes back to the audio menu items. You can stop the information from playing at any time with a long-press. To exit the audio menu items long-press again. You will then be back to listening to the chattering voices.

Egocentric Simultaneous

As you walk around the exhibition space, when you get closer to an artwork that has been audio-augmented, you will listen to some chattering voices. This sound indicates that you can access extra audio information about the artwork. In order to play this extra information, walk to the place where the chattering voices are at their loudest (loud in both your ears) and push (short-press) the SHAKE sensor pack navigation switch. This will play the audio menu sounds available for that artwork. The audio menu sounds are: *water waves* identifying information provided by the artist, *open crackling fire* identifying positive comments left by previous visitors and *stormy wind* identifying negative comments left by previous visitors. **Notice that not all artworks have been audio-augmented and individual artworks may have different numbers of audio menu sounds**.

The presentation of the audio menu sounds in this condition will be simultaneous (all audio menu sounds will be presented at the same time) spatialised around your head, i.e. perceived as originating from the right, left or in front of you. In order to browse the audio menu items push the SHAKE sensor pack navigation switch right or left. When an audio menu item is selected the volume will increase. If you reach the end of the audio menu items they will loop around. To listen to the artwork information short-press when listening to the targeted audio menu item and the information will start playing. Once the information has finished playing, the system automatically goes back to the audio menu items. You can stop the information from playing

at any time with a long-press. To exit the audio menu items long-press again. You will then be back to listening to the chattering voices.

Exocentric Simultaneous

As you walk around the exhibition space, when you get closer to an artwork that has been audio-augmented, you will listen to some chattering voices. This sound indicates that you can access extra audio information about the artwork. In order to play this extra information, walk to the place where the chattering voices are at their loudest (loud in both your ears) and push (short-press) the SHAKE sensor pack navigation switch. This will play the audio menu sounds available for that artwork. The audio menu sounds are: *water waves* identifying information provided by the artist, *open crackling fire* identifying positive comments left by previous visitors and *stormy wind* identifying negative comments left by previous visitors. **Notice that not all artworks have been audio-augmented and individual artworks may have different numbers of audio menu sounds**.

The presentation of the audio menu sounds in this condition will be simultaneous (all audio menu sounds will be presented at the same time) and audio menu items will be perceived as if they were fixed to a location. In order to browse the audio menu items you will have to walk around the artwork and stand at the location where you think the audio menu item is situated. When you are standing at the right location the volume will increase. To listen to the artwork information short-press when listening to the targeted audio menu item and the information will start playing. Once the information has finished playing, the system automatically goes back to the audio menu items. You can stop the information from playing at any time with a long-press. To exit the audio menu items long-press again. You will then be back to listening to the chattering voices.

F.4 User Satisfaction Questionnaire

Audio-augmented Exhibition: User *satisfaction* questionnaire

Age: _____ **Gender** : ___ male ___ female

Part 1: System Experience

1.1 On average how much time do you spend at Museums/Art Galleries?

___ at least once a week
___ no more then once a month
___ Two to three times a year
___ Once a year at most
___ I rarely find myself at museums/art galleries
___ Other

Part 2: Past Experience

2.1 How many interactive museum systems have you used in the past.

___ None
___ 1
___ 2
___ 3 to 4
___ 5 to 6
___ more than 6

2.2 Of the following interactive museum systems which ones are you familiar

___ Audio tape systems
___ PDA style systems
___ Docent tours
___ Tour guide robotics
___ Interactive Kiosks
___ Film and video
___ Seated / ride based systems
___ Kinesthetic sensor systems

Part 3: Overall User Reactions

Please circle the numbers which most appropriately reflect your impressions about using this system. Not applicable = NA

3.1 Over all reactions to the system:	Terrible					wonderful	
	1	2	3	4	5	NA	
3.2	Frustrating					satisfying	
	1	2	3	4	5	NA	
3.3	Dull					Stimulating	
	1	2	3	4	5	NA	
3.4	Difficult					Easy	
	1	2	3	4	5	NA	
3.5	Rigid					Flexible	
	1	2	3	4	5	NA	
3.6	Non-Immersive					Immersive	
	1	2	3	4	5	NA	

Part 4: System interface

Interaction device (SHAKE sensor pack)

4.1 The interaction device was	Heavy					Light	
	1	2	3	4	5	NA	
4.2 Holding the interaction device was	Difficult					Easy	
	1	2	3	4	5	NA	
4.3 Manipulating the interaction device was	Difficult					Easy	
	1	2	3	4	5	NA	
4.4 Within the exhibition environment, the interaction device seemed	Inappropriate					Appropriate	
	1	2	3	4	5	NA	
4.5 Understanding how to use the interaction device was	Difficult					Easy	
	1	2	3	4	5	NA	
4.6 In general the interaction device was	Annoying					Enjoyable	
	1	2	3	4	5	NA	

Headphones

4.7 The headphones were	Comfortable					uncomfortable
	1	2	3	4	5	NA
4.7 headphones transmission was clear	Always					Never
	1	2	3	4	5	NA

Part 5: Learning

5.1 Learning to operate the system was	Difficult					Easy
	1	2	3	4	5	NA
5.2 Getting started with the system was	Difficult					Easy
	1	2	3	4	5	NA
5.3 Exploration of features seemed	Risky					safe
	1	2	3	4	5	NA
5.4 Discovering new features seemed	Difficult					Easy
	1	2	3	4	5	NA
5.5 Remembering rules about interacting with the system was	Difficult					Easy
	1	2	3	4	5	NA
5.6 The number of steps needed to complete a task was	Too many					just right
	1	2	3	4	5	NA
5.7 Steps to complete a task followed a logical order	Never					always
	1	2	3	4	5	NA
5.8 Feedback on the completion of steps was	Unclear					clear
	1	2	3	4	5	NA

Part 6: System Capabilities

6.1 System speed was	Too slow				fast enough	
	1	2	3	4	5	NA
6.2 Response time for most operations	Too slow				fast enough	
	1	2	3	4	5	NA
6.3 The system is reliable	never				always	
	1	2	3	4	5	NA
6.4 Operations were	undependable				dependable	
	1	2	3	4	5	NA
6.5 System failures occur	Frequently				seldom	
	1	2	3	4	5	NA
6.6 System warns you about potential problems	Never				always	
	1	2	3	4	5	NA
6.7 Ease of operation depends on your level of experience	Never				always	
	1	2	3	4	5	NA
6.8 You can accomplish tasks knowing only a few commands.	With difficulty				easily	
	1	2	3	4	5	NA
6.9 You can use system features	With difficulty				easily	
	1	2	3	4	5	NA

Part 7: Content management

Audio Content

7.1 The audio content seemed	Loud				quiet	
	1	2	3	4	5	NA
7.2	Uninformative				informative	
	1	2	3	4	5	NA
7.3	Inaccurate				accurate	
	1	2	3	4	5	NA
7.4	Incomprehensive				comprehensive	
	1	2	3	4	5	NA

7.5	Generalized customized for me 1 2 3 4 5 NA
7.6	Confusing clear 1 2 3 4 5 NA
7.7	Irrelevant relevant 1 2 3 4 5 NA
7.8	Rigid playful 1 2 3 4 5 NA
7.9	Un-entertaining entertaining 1 2 3 4 5 NA
7.11 In relation to the visual elements in the exhibit the audio content was	Distractive synergistic 1 2 3 4 5 NA
	Inappropriate appropriate 1 2 3 4 5 NA

Audio menu sounds

7.14 Audio menu sounds were	Non-seductive seductive 1 2 3 4 5 NA
7.15	Rigid playful 1 2 3 4 5 NA
7.16	Irritating enjoyable 1 2 3 4 5 NA
7.17	Predictable surprising 1 2 3 4 5 NA

Audio experience

7.18 In general the audio experience was	Monotonous entertaining 1 2 3 4 5 NA
7.19	Predictable surprising 1 2 3 4 5 NA
7.20	Clunky flowing 1 2 3 4 5 NA

7.21	Frustrating satisfying 1 2 3 4 5 NA
7.22	Confusing clear 1 2 3 4 5 NA
7.23	Rigid playful 1 2 3 4 5 NA
7.24	Inappropriate appropriate 1 2 3 4 5 NA
7.25	Mechanical human like 1 2 3 4 5 NA
7.26 I felt that my time using this system was.	Wasteful valuable 1 2 3 4 5 NA

Part 8: Navigation and interaction

8.1 I was able to navigate the presented information in a ***meaningful*** way	Never always 1 2 3 4 5 NA
8.1 I was able to navigate the presented information in a ***efficient*** way	Never always 1 2 3 4 5 NA
8.2 The organization of information was	Confusing clear 1 2 3 4 5 NA
8.3 Audio sounds playing simultaneously was	Annoying enjoyable 1 2 3 4 5 NA
8.4 I found I wanted to replay my choice options	Never Always 1 2 3 4 5 NA
8.5 I found myself feeling lost in the system.	Always Never 1 2 3 4 5 NA
8.6 I found myself uncertain as to the state of system	Always Never 1 2 3 4 5 NA

Bibliography

Algazi, A. R., R. O. Duda, D. M. Thompson, and C. Avendano (2001) The CIPIC hrtf database. In *Proceedings of 2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 99–102.

Audacity (2010) `http://audacity.sourceforge.net/`.

Aylett, M. P. and C. J. Pidcock (2007) The cerevoice characterful speech synthesiser sdk. In *AISB*, pp. 174–178. Newcastle, UK.

Batteau, D. W. (1967) The role of the pinna in human sound localization. In *Proceedings of the Royal Society*, vol. 168, pp. 158–180.

Bederson, B. B. (1995) Audio augmented reality: A prototype automated tour guide. In *Proceedings of CHI'95*, vol. 2, pp. 210–211. ACM Press.

Bederson, B. B. (2011) The promise of zoomable user interfaces. *Behaviour & Information Technology*, **30**, 6, 853–866.

Begault, D. R. (1994) *3D Sound for Virtual Reality and Multimedia*. Boston, MA. USA: Academic Press.

Begault, D. R. and E. M. Wenzel (1993) Headphone localization of speech. *Human Factors*, **35**, 361–376.

Begault, D. R., E. M. Wenzel, and M. R. Anderson (2001) Direct comparison of the impact of head tracking, reverberation and individualized head-related transfer functions on the spatial perception of a virtual speech source. *Journal of the Audio Engineering Society*, **49**, 10, 904–916.

Berkhout, A. J., D. de Vries, and P. Vogel (1993) Acoustic control by wavefield synthesis. *Journal of the Acoustical Society of America*, **93**, 5, 2754–2778.

Best, V., A. Ihlefeld, and B. G. Shinn-Cunningham (2005) The effect of auditory spatial layout in a divided attention task. In *Proceedings of ICAD 2005*, pp. 17–22. Limerick, Ireland.

Blattner, M. M., D. A. Sumikawa, and R. M. Greenberg (1989) Earcons and icons: Their structure and common design principles. *Human-Computer Interaction*, **4**, 1, 11–44.

Blauert, J. (1997) *Spatial Hearing: The psychophysics of human sound localization.* Cambridge (MA), USA: MIT Press.

Bly, S. A. (1985) Communicating with sound. In *Proceedings of CHI '85*. San Francisco, California, USA: ACM Press.

Boone Jr, H. N. and D. A. Boone (2012) Analyzing likert data. *Journal of Extension*, **50**, 2.

Bregman, A. S. (1990) *Auditory Scene Analysis: The Perceptual Organization of sound*. Cambridge, MA, USA: MIT Press.

Brewster, S. A. (2002) Non-speech auditory output. In J. Jacko and A. Sears (eds.), *Human-Computer Interaction Handbook*, pp. 220–239. Mahwah, NJ.: Lawrence Erlbaum Associates.

Brewster, S. A. and M. G. Crease (1999) Correcting menu usability problems with sound. *Behaviour and Information Technology*, **18**, 3, 165–177.

Brewster, S. A., J. Lumsden, M. Bell, M. Hall, and S. Tasker (2003) Multimodal 'eyes-free' interaction techniques for wearable devices. In *Proceedings of CHI 2003*, pp. 463–480. Fort Lauderdale, Florida, USA: ACM Press.

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

Bronkhorst, A. W. (2000) The cocktail party phenomenon: a review of research on speech intelligibility in multiple-talker conditions. *Acustica*, **86**, 117–128.

Brungart, D. S., M. A. Ericson, and B. D. Simpson (2002) Design considerations for improving the effec- tiveness of multitalker speech displays. In *Proceedings of ICAD 2002*, vol. 1, p. 424430. Kyoto, Japan.

Brungart, D. S., B. D. Simpson, M. A. Ericson, and K. R. Scott (2001) Informational and energetic masking effects in the perception of multiple simultaneous talkers. *Journal of Acoustical Society of America*, **110**, 5, 2527–2538.

Buxton, B. (1995) Integrating the periphery and context: A new model of telematics. In *Proceedings of Graphics Interface '95*, pp. 239–246. Quebec, Canada.

Carlile, S. (1996) *Virtual Auditory Space: Generation and Application*. Austin, TX. USA: R. G. Landes Company.

Carlile, S., P. Leong, S. Hyams, and D. Pralong (1997) The nature and distribution of errors in the localization of sounds in humans. *Hearing Research*, **114**, 179–196.

Cater, K., R. Hull, K. O'Hara, T. Melamed, and B. Clayton (2007) The potential of spatialised audio for location based services on mobile devices: Mediascapes. In *SAMD: Workshop on Spatialised Audio for Mobile Devices, MobileHCI 2007*. Singapore: ACM Press.

Cherry, E. C. (1953) Some experiments on the recognition of speech, with one and with two ears. *Journal of Acoustical Society of America*, **25**, 5, 975–979.

Cohen, J. (1994) Monitoring background activities. In G. Kramer (ed.), *Auditory Display: Sonification, Audification and Auditory interfaces*, pp. 499–522. Westview Press.

Cohen, M. (1993) Throwing, pitching and catching sound: audio windowing models and modes. *International Journal of Man-Machine Studies*, **39**, 2, 269–304.

Cohen, M., S. Aoki, and N. Koizumi (1993) Augmented audio reality: Telepresence/vr hybrid acoustic environments. In *Proceedings of RO-MAN: 2nd IEEE International Workshop on Robot and Human Communication*, pp. 361–364. Tokyo, Japan.

Cohen, M. and L. F. Ludwig (1991) Multidimensional audio window management. *International Journal of Man-Machine Studies*, **34**, 3, 319–336.

Coleman, P. (1963) An analysis of cues to auditory depth perception in free space. *Psychological Bulletin*, **60**, 302–315.

Correia, N., T. Mota, R. Nóbrega, L. Silva, and A. Almeida (2010) A multi-touch tabletop for robust multimedia interaction in museums. In *ACM International Conference on Interactive Tabletops and Surfaces*, ITS '10, pp. 117–120. Saarbrücken, Germany: ACM Press.

Crispien, K., K. Fellbaum, A. Savidis, and C. Stephanidis (1996) A 3d-auditory environment for hierarchical navigation in non-visual interaction. In *Proceedings of the 3rd International Conference on Audio Display (ICAD)*, pp. 18–21. Palo Alto, California, USA.

D'Appolito, J. (1998) *Testing Loudspeakers*. Audio Amateur Press.

Deutsch, D. (1999) Grouping mechanisms in music. In D. Deutsch (ed.), *The Psychology of Music*, p. 299348. San Diego, CA. USA: Academic Press, 2nd edn.

Devore, S. and B. G. Shinn-Cunningham (2003) Perceptual consequences of including reverberation in spatial auditory displays.

Dicke, C., K. Wolf, and Y. Tal (2010) Foogue: eyes-free interaction for smartphones. In *Proceedings of MobileHCI'10*, pp. 455–458. Lisbon, Portugal: ACM Press.

Draycott, S. G. and P. Kline (1996) Validation of the agard stres battery of performance tests. *Human Factors*, **38**, 2, 347–361.

Eckel, G. (2001) Immersive audio-augmented environments: The listen project. In *IV'01: Proceedings of the Fifth International Conference on Information Visualisation*, pp. 571–573. London, England, UK: IEEE Computer Society Press.

Eichner, M., M. J. Marn-Jimnez, A. Zisserman, and V. Ferrari (2012) 2d articulated human pose estimation and retrieval in (almost) unconstrained still images. *International Journal of Computer Vision*, **99**, 2, 190–214.

Engadget (2012) `http://www.engadget.com/2012/02/03/canalys-more-smartphones-than-pcs-shipped-in-2011/`.

Etter, R. and M. Specht (2005) Melodious walkabout: Implicit navigation with contextualized personal audio contents. In *In Adj. Proc. Pervasive Computing*, p. 43. Technology.

Friedman, M. (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, **32**, 200, 675–701.

Fry, D. (1979) *The Physics of Speech*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Fu, C., W. Goh, and J. A. Ng (2010) Multi-touch techniques for exploring large-scale 3d astrophysical simulations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pp. 2213–2222. Atlanta, Georgia, USA: ACM Press.

Gardner, W. G. (1999) 3d audio and acoustic environment modeling. *Technical report*, Wave Arts Inc., Arlington, MA. USA.

Gaver, W. (1997) Auditory interface. In t. helander and p. prabhu (eds.), *Handbook of Human-Computer Interaction*, pp. 1003–1041. Amsterdam, The Netherlands: Elsevier Science, 2nd edn.

Gaver, W. W. (1989) The sonicfinder: An interface that uses auditory icons. *Human-Computer Interaction*, **4**, 1, 67–94.

Goldstein, E. B. (2009) *Sensation and Perception*. Wadsworth Cengage Learning, 8th edn.

Goßmann, J. and M. Specht (2002) Location models for augmented environments. *Personal and Ubiquitous Computing*, **6**, 5-6, 334–340.

Grossman, T. and R. Blakrishnan (2005) The bubble cursor: enhancing target acquisition by dynamic resizing of the cursor's activation area. In *Proceedings of CHI 2005*, pp. 281–290. Portland, Oregon, USA: ACM Press.

Handedness Earedness Questionnaires (2009) `http://www.jackielam.net/handedness/`.

Handel, S. (1989) *Listening: An introduction to the perception of auditory events*. MIT Press, Cambridge, MA.

Hardyck, C. and L. F. Petrinovich (1977) Left-handedness. *Psychological Bulletin*, **84**, 385–404.

Hart, S. G. and L. E. Staveland (1988) Development of nasa tlx (task load index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati (eds.), *Human Mental Workload*, pp. 139–183. Amsterdam: North Holland Press.

hartmann, W. M. (1995) The physical description of signals. In B. C. J. Moore (ed.), *Hearing, Handbook of Perception and Cognition*, p. 140. San Diego, CA. USA: Academic Press Inc.

Hartmann, W. M., P. X. Zhang, and J. F. Culling (2001) Earedness: Left-eared and right-eared listeners. *Acoustical Society of America Journal*, **115**, 5, 2534–2535.

Hassenzahl, M. (2005) The thing and i: Understanding the relationship between user and product. In M. Blythe, K. Overbeeke, A. Monk, and P. Wright (eds.), *Funology*, vol. 3 of *Human-Computer Interaction Series*, pp. 31–42. Springer Netherlands.

Hawley, M. L., R. Y. Litovsky, and J. Culling (2000) The "cocktail party problem" with four types of maskers: Speech, time-reversed speech, speech-shaped noise, or modulated speech-shaped noise.

Helander, M., T. Landauer, and P. Prabhu (1997) *Handbook of Human-Computer Interaction*. North Holland.

Heller, F. and J. Borchers (2011) Corona: Audio augmented reality in historic sites. In *Proceedings of MobileHCI 2011*, pp. 51–54. Stockholm, Sweden: ACM Press.

Heller, F., T. Knott, M. Weiss, and J. Borchers (2009) Multi-user interaction in virtual audio spaces. In *Extended Abstracts of CHI 2009*, pp. 4489–4494. Boston, Massachusetts, USA: ACM Press.

Herron, J. (1980) *Neuropsychology of Left-handedness.* New York, USA: Academic Press.

Holland, S., D. R. Morse, and H. Gedenryd (2002) Audiogps: spatial audio in a minimal attention interface. *Personal and Ubiquitous Computing*, **6**, 4, 253–259.

Ihlefeld, A. and B. G. Shinn-Cunningham (2008) Spatial release from energetic and informational masking in a divided speech identification task. *Journal of Acoustical Society of America*, **123**, 4380–4392.

International Conference on Auditory Display (ICAD) (2012) `http://www.icad.org/`.

JAKE Sensor Pack (2010) `http://code.google.com/p/jake-drivers/`.

Jones, M., S. Jones, G. Bradley, N. Warren, D. Bainbridge, and G. Holmes (2008) Ontrack: Dynamically adapting music playback to support navigation. *Personal and Ubiquitous Computing*, **12**, 7, 513–525.

JSR-234 Advanced Multimedia Supplements API (AMMS) (2009) `http://theoreticlabs.com/dev/api/jsr-234/javax/microedition/amms/package-summary.html`.

Kobayashi, M. and C. Schmandt (1997) Dynamic soundscape:mapping time to space for audio browsing. In *Proceedings of CHI 1997*, pp. 194–201. Atlanta, Georgia, USA.

Kruskal, W. H. and A. Wallis (1952) Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, **47**, 583–621.

Kryter, K. D. (1972) Speech communication. In H. P. V. Cott and R. G. Kinkade (eds.), *Human Engineering Guide to Equipment Design*. McGraw–Hill.

Lam, H. and T. Munzner (2010) *A guide to visual multi-level interface design from synthesis of empirical study evidence*. San Rafael, Calif. (1537 Fourth Street, San Rafael, CA 94901 USA): Morgan & Claypool.

Leftheriotis, I. and K. Chorianopoulos (2011) User experience quality in multi-touch tasks. In *Proceedings of the 3rd ACM SIGCHI symposium on Engineering interactive computing systems*, EICS '11, pp. 277–282. Pisa, Italy: ACM Press.

Lemordant, J. and A. Guerraz (2007) Mobile immersive music. In *Proceedings of the 2007 International Computer Music Conference, ICMC 2007*, pp. 21–24. ICMA, San Francisco, USA.

Litovsky, R. Y. (2008) *Binaural Hearing*. Cochlear.

Litovsky, R. Y., H. S. Colburn, W. A. Yost, and S. Guzman (1999) The precedence effect. *Journal of Acoustical Society of America*, **106**, 1633–1654.

Ludwig, L., N. Pincever, and M. Cohen (1990) Extending the notion of a window system to audio. *IEEE Computer*, **23**, 8, 66–72.

Lund, A. M. (1997) Expert ratings of usability maxims. *Ergonomics in Design*, **5**, 3, 15–20.

Lyons, K., M. Gandy, and T. Starner (2000) Guided by voices: An audio augmented reality system. In *Proceedings of the International Conference on Auditory Display – ICAD 2000*, pp. 57–62. Atlanta, GA, USA.

Magnusson, C., B. Breidegard, and K. Rassmus-Gröhn (2009) Soundcrumbs – hansel and gretel in the 21st century. In S. LNCS (ed.), *HAID 2009*. Dresden, Germany.

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

Malham, D. and A. Myatt (1995) 3-d sound spatialization using ambisonic techniques. *Computer Music Journal*, **19**, 4, 58–70.

Mann, H. B. and D. R. Whitney (1947) On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, **18**, 1, 50–60.

Marentakis, G. and S. A. Brewster (2005) A comparison of feedback cues for enhancing pointing efficiency in interaction with spatial audio displays. In *Proceedings of MobileHCI 2005*, pp. 55–62. Salzburg, Austria: ACM Press.

Marentakis, G. N. and S. A. Brewster (2006) Effects of feedback, mobility and index of difficulty on deictic spatial audio target acquisition in the horizontal plane. In *Proceedings of CHI 2006*, pp. 359–368. Montreal, Quebec, Canada: ACM Press.

Mariette, N. (2007) From backpack to handheld: The recent trajectory of personal location aware spatial audio. In *Proceedings of DAC 2007: 7th digital arts and culture conference*, pp. 233–240. Perth, Australia.

Mariette, N. (2010) Navigation performance effects of render method and head-turn latency in mobile audio augmented reality. In *CMMR/ICAD 2009*, vol. LNCS 5954, pp. 239–265. Copenhagen, Denmark.

Martin, G. (2006) *Introduction to Sound Recording*. www.tonmeister.ca.

McCarthy, J. and P. Wright (2004) *Technology as Experience*. MIT Press.

McGookin, D., S. Brewster, and P. Priego (2009) Audio bubbles. employing non-speech audio to support tourist wayfinding. In M. E. Altinsoy, U. Jekosch, and S. Brewster (eds.), *HAID 2009*, vol. LNCS 5763. Dresden, Germany.

McGookin, D. K. (2004) *Understanding and improving the identification of concurrently presented earcons.* Phd thesis, School of Computing Science, Glasgow, UK.

McGookin, D. K. and S. A. Brewster (2004) Space the final frontearcon: The identication of concurrently presented earcons in a synthetic spatialised auditory environment. In *Proceedings of ICAD 2004*. Sydney, Australia.

Microsoft (1995) *The Windows Interface Guidelines for Software Design*. Microsoft Press.

Middlebrooks, J. (1997) Spectral shape cues for sound localization. In R. H. Gilkey and T. R. Anderson (eds.), *Binaural and spatial hearing in real and virtual environments*, pp. 77–98. Mahwah, NJ. USA: Lawrence Erlbaum Associates.

Middlebrooks, J. and D. Green (1991) Sound localization by human listeners. *Annual Psychology Review*, **42**, 135–159.

Mills, A. W. (1958) On the minimal audible angle. *Journal of the Acoustical Society of America*, **30**, 237–246.

Mills, A. W. (1972) Auditory localization. In J. V. Tobias (ed.), *Foundations of modern auditory theory*, vol. 2, pp. 301–348. New York, USA: Academic Press.

Mobile Trail Explorer (2010) `http://code.google.com/p/mobile-trail-explorer/`.

Moore, B. C. J. (2004) *An introduction to the psychology of hearing*. London, UK: Academic Press, 5th edn.

Morrison, A. J., P. Mitchell, and M. Brereton (2007) The lens of ludic engagement: evaluating participation in interactive art installations. In *Proceedings of the 15th international conference on Multimedia, ser. MULTIMEDIA 07*, pp. 509–512. Augsburg, Germany.

Musicant, D. and A. Butler (1985) Influence of monaural spectral cues on binaural localization. *Journal of the Acoustical Society of America*, **77**, 1, 202–208.

Mynatt, E., M. Back, R. Want, M. Baer, and J. B. Ellis (1998) Designing audio aura. In *Proceedings of CHI 1998*, pp. 566–573. Los Angeles, CA, USA: ACM Press.

Nielsen, J. (1994) *Heuristic evaluation*. New York, NY, USA: John Wiley & Sons.

Nokia N95 8GB (2009) http://www.nseries.com/index.html#l=products,n95_8gb.

Noonan, M. and S. Axelrod (1981) Earedness (ear choice in monaural tasks): Its measurement and relationship to other lateral preferences. *Journal of Auditory Research*, **21**, 263–277.

Oldfield, S. R. and S. P. A. Parker (1984) Acuity of sound localisation: A topography of auditory space i. *Perception*, **13**, 581–600.

Oppenheim, A. V. and R. W. Schafer (1989) *Discrete Time Signal Processing*. Englewood Clis, NJ. USA: Prentice Hall.

Papa Sangre (2012) http://www.papasangre.com/.

Porac, C. and S. Coren (1981) *Lateral preferences and human behavior*. New York, USA: Springer-Verlag Berlin and Heidelberg GmbH & Co. K.

Pressnitzer, D., A. de Cheveigné, S. McAdams, and L. Collet (2006) *Auditory Signal Processing: Physiology, Psychoacoustics, and Models*. Springer.

Pulkki, V. and T. Hirvonen (2005) Localization of virtual sources in multichannel audio reproduction. *IEEE Transactions on Speech and Audio Processing*, **13**, 1, 105–119.

Qstarz GPS Receiver (2010) http://www.qstarz.com/Products/GPS%20Products/BT-Q1000X-F.htm.

Raman, T. V. (1997) *Auditory Interfaces: Towards the speaking computer*. Kluwer Academic Publishers.

Rauscher, F. H., G. L. Shaw, and C. N. Ky (1993) Music and spatial task performance. *Nature*, **365**, 6447, 611–611.

Rayleigh, L. (1907) On our perception of sound direction. *Philos Mag*, **13**, 214–232.

Reid, J., E. Geelhoed, R. Hull, K. Carter, and B. Clayton (2005) Parallel worlds: Immersion in location-based experiences. In *Proceedings of CHI 2005*, vol. 2, pp. 1733–1736. Portland, Oregon, USA: ACM Press.

Roth, V., P. Schmidt, and B. Güldenring (2010) The IR ring: authenticating users' touches on a multi-touch display. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, UIST '10, pp. 259–262. New York, New York, USA: ACM Press.

Rozier, J., K. Karahalios, and D. J. (2000) Hear & there: An augmented reality system of linked audio. In *Proceedings of the International Conference on Auditory Display – ICAD 2000*. Atlanta, Georgia, USA.

Sawhney, N. and C. Schmandt (2000) Nomadic radio: Speech and audio interaction for contextual messaging in nomadic environments. *ACM Transactions on Computer-Human Interaction*, **7**, 3, 353–383.

Schmandt, C. (1998) Audio hallway: a virtual acoustic environment for browsing. In *Proceedings of 11th annual ACM symposium on User interface software and technology*, pp. 163–170. Francisco, California, USA.

Schmandt, C. and C. Mullins (1995) Audiostreamer: exploiting simultaneity for listening. In *Proceedings of CHI 1995*, pp. 218–219. Denver, Colorado, USA: ACM Press.

SHAKE SK6 sensor pack (2010) `http://code.google.com/p/shake-drivers/`.

Shepard, M. (2006) Tactical sound garden toolkit. In *3rd International Workshop on Mobile Music Technology*. Brighton, UK.

Shinn-Cunningham, B. G. (2002) Speech intelligibility, spatial unmasking, and realism in reverberant spatial auditory displays.

Shinn-Cunningham, B. G. and A. Ihlefeld (2004) Selective and divided attention: extracting information from simultaneous sound sources. In *Proceedings of ICAD 04*. Sydney, Australia.

Shneiderman, B. (1998) *Designing the User Interface*. Reading, Massachusetts, USA: Addison-Wesley.

Stahl, C. (2007) The roaring navigator: A group guide for the zoo with a shared auditory landmark display. In *Proceedings of MobileHCI 2007*, pp. 282–386. Singapore: ACM Press.

Stevens, S. S. (1955) The measurement of loudness. *Journal of the Acoustical Society of America*, **27**, 815–829.

Stifelman, L. J. (1994) The cocktail party effect in auditory interfaces: a study of simultaneous presentation. *Tech. rep.*, MIT Media Laboratory Technical Report.

Strachan, S., P. Eslambolchilar, and R. Murray-Smith (2005) Gpstunes: Controlling navigation via audio feedback. In *Proceedings of MobileHCI 2005*, pp. 275–278. ACM Press.

Terrenghi, L. and A. Zimmermann (2004) Tailored audio augmented environments for museums. In *IUI 04: Proceedings of the 9th international conference on Intelligent user interfaces*, pp. 334–336. Funchal, Madeira, Portugal: ACM Press.

Thurlow, W. R. and P. S. Runge (1967) Effect of induced head movements on the localization of direction of sounds. *Journal of the Acoustical Society of America*, **42**, 2, 480–488.

University of Sidney and University of York (2013) `http://sydney.edu.au/engineering/electrical/carlab/hrtfmorph.htm`.

Vazquez-Alvarez, Y. and S. Brewster (2009a) Audio minimization: Applying 3D audio techniques to multi-stream audio interfaces. In M. E. Altinsoy, U. Jekosch, and S. Brewster (eds.), *HAID 2009*, vol. LNCS 5763. Dresden, Germany: Springer.

Vazquez-Alvarez, Y. and S. Brewster (2009b) Investigating background & foreground interactions using spatial audio cues. In *CHI '09 Extended Abstracts on Human Factors in Computing Systems*, pp. 3823–3828. Boston, MA, USA: ACM Press.

Vazquez-Alvarez, Y. and S. A. Brewster (2010) Designing spatial audio interfaces to support multiple audio streams. In *Proceedings of MobileHCI 2010*, pp. 253–256. Lisbon, Portugal: ACM Press.

Vazquez-Alvarez, Y. and S. A. Brewster (2011) Eyes-free multitasking: The effect of cognitive load on mobile spatial audio interfaces. In *Proceedings of CHI 2011*, pp. 2173–2176. Vancouver, Canada: ACM Press.

Vazquez-Alvarez, Y., I. Oakley, and S. A. Brewster (2012) Auditory display design for exploration in mobile audio-augmented reality. *Personal and Ubiquitous Computing*, **16**, 8, 987–999.

Wakkary, R. and M. Hatala (2007) Situated play in a tangible interface and adaptive audio museum guide. *Journal of Personal and Ubiquitous Computing*, **11**, 3, 171–191.

Walker, A. and S. A. Brewster (2000) Spatial audio in small display screen devices. *Personal Technologies*, **4**, 2, 144–154.

Walker, A., S. A. Brewster, D. McGookin, and A. Ng (2001) Diary in the sky: A spatial audio display for a mobile calendar. In *Proceedings of BCS IHM-HCI 2001*, pp. 531–540. Lille, France.

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

Watson, R. and O. Downey (2009) *The Little Red Book of Acoustics: A Practical Guide*. Blue Tree Acoustics.

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

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

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

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

Yost, W. A., J. Dye, R. H., and S. Sheft (1996) A simulated "cocktail party" with up to three sound sources. *Perception and Psychophysics*, **58**, 1026–1036.

Zahorik, P. (2002) Auditory display of sound source distance. In *Proceedings of ICAD 2002*. Kyoto, Japan.

Zatorre, R. J. (2003) Sound analysis in auditory cortex. *Trends in Neurosciences*, **26**, 5, 229–230.

Zhang, H., X. Yang, B. Ens, H. Liang, P. Boulanger, and P. Irani (2012) See me, see you: a lightweight method for discriminating user touches on tabletop displays. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, CHI '12, pp. 2327–2336. Austin, Texas, USA: ACM Press.

Zhao, S., P. Dragicevic, M. Chignell, R. Balakrishnan, and P. Baudisch (2007) Earpod: eyes-free menu selection using touch input and reactive audio feedback. In *Proceedings of CHI 2007*, pp. 1395–1404. San Jose, California, USA: ACM Press.