QUEEN MARY UNIVERSITY OF LONDON

PHD THESIS

Binaural virtual auditory display for music discovery and recommendation

Author:
Rishi Chirantan Shukla

Submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

in the

Centre for Digital Music
School of Electronic Engineering and Computer Science

Queen Mary
University of London
Science and Engineering

March 29, 2023

Statement of originality

I, Rishi Chirantan Shukla, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

Signature:

Rishi Shukla

Date: March 29, 2023

Details of collaboration

Four collaborations were undertaken in the course of this thesis. The input of collaborators is specified in the body of the document wherever this occurred and is also summarised below.

Chapter 3 – ***Study 1*** was completed whilst acting as a visiting researcher at BBC R&D. The conceptualisation and implementation of the two systems used in the study was completed entirely personally. Audio content curation and preparation was completed in collaboration with Alex Robertson, Music Editor. The mixed methods research design benefitted from the equal inputs of Holly Challenger, Senior User Experience Researcher, and Joanna Rusznica, Senior User Experience Designer. All resulting data analysis and writing-up was completed personally in its entirety, other than interpretation of mental model illustrations, which was pursued in a 30 minute joint analysis session with the two user experience professionals.

Chapter 3 – ***Study 2a*** and **Chapter 4 –** ***Study 2b*** were based on an experiment designed and conducted by Dr Rebecca Stewart, prior to the start of the doctoral research presented in this thesis. Data collection was undertaken by Dr Stewart as part of postdoctoral research whilst on secondment to New York University in 2011. Neither the study design, methodology, or the data from that research had been previously reported or interrogated in any form. Full documentation and analysis of that investigation was conducted entirely personally.

Chapter 5 – ***Binaural virtual auditory display system and measurement*** was pursued entirely independently, other than software implementation of the reference first order Ambisonics system used for comparison purposes, which was implemented through supervision of a Masters research student, Teodor Radu.

Chapter 6 – ***Prototype binaural playlist editor*** was pursued entirely independently, other than software implementation of the the 3D synthetic reverberation component of the final system, which was adapted as described from work completed via supervision of a Masters research student, Christopher Yeoward.

List of publications

The following publications have resulted from work pursued in relation to this thesis:

Peer-reviewed conference papers

R. Shukla, R. Stewart, A. Roginska and M. Sandler (August 2018). "User selection of optimal HRTF sets via holistic comparative evaluation". In *AES International Conference on Audio for Virtual and Augmented Reality*. Redmond, WA, USA, pp. 1-10.

R. Shukla, T. Radu, R. Stewart and M. Sandler (March 2019). "Real-time binaural rendering with virtual vector base amplitude panning". In *AES International Conference on Immersive and Interactive Audio*. York, UK, pp. 1-10.

R. Shukla, R. Stewart and M. Sandler (June 2021). "User HRTF Selection for 3D Auditory Mixed Reality". In *18th Sound and Music Computing Conference*. Online Virtual Conference, pp. 84-91.

Industry research report

R. Shukla (2019). *Voice-led interactive exploration of audio.* Tech. rep. London, UK: BBC R&D.

Peer-reviewed journal paper

C. Yeoward, R. Shukla, R. Stewart, M Sandler and J. Reiss (2021). "Real-time binaural room modelling for augmented reality applications". In: *Journal of the Audio Engineering Society* 69.11, pp. 818–833.

QUEEN MARY UNIVERSITY OF LONDON

Abstract

Faculty of Science and Engineering
School of Electronic Engineering and Computer Science

Doctor of Philosophy

Binaural virtual auditory display for music discovery and recommendation

by Rishi Chirantan Shukla

Emerging patterns in audio consumption present renewed opportunity for searching or navigating music via spatial audio interfaces. This thesis examines the potential benefits and considerations for using binaural audio as the sole or principal output interface in a music browsing system. Three areas of enquiry are addressed.

Specific advantages and constraints in spatial display of music tracks are explored in preliminary work. A voice-led binaural music discovery prototype is shown to offer a contrasting interactive experience compared to a mono smartspeaker. Results suggest that touch or gestural interaction may be more conducive input modes in the former case. The limit of three binaurally spatialised streams is identified from separate data as a usability threshold for simultaneous presentation of tracks, with no evident advantages derived from visual prompts to aid source discrimination or localisation.

The challenge of implementing personalised binaural rendering for end-users of a mobile system is addressed in detail. A custom framework for assessing head-related transfer function (HRTF) selection is applied to data from an approach using 2D rendering on a personal computer. That HRTF selection method is developed to encompass 3D rendering on a mobile device. Evaluation against the same criteria shows encouraging results in reliability, validity, usability and efficiency.

Computational analysis of a novel approach for low-cost, real-time, head-tracked binaural rendering demonstrates measurable advantages compared to first order virtual Ambisonics. Further perceptual evaluation establishes working parameters for interactive auditory display use cases.

In summation, the renderer and identified tolerances are deployed with a method for synthesised, parametric 3D reverberation (developed through related research) in a final prototype for mobile immersive playlist editing. Task-oriented comparison with a graphical interface reveals high levels of usability and engagement, plus some evidence of enhanced flow state when using the eyes-free binaural system.

Acknowledgements

Sincere and especial thanks go to the co-supervisors of this thesis, Mark Sandler and Rebecca Stewart. Their guidance and support have been instrumental not just in the success of its outcomes, but in making the whole experience everything it should be – rewarding, ambitious, joyful and independent. At all turns, Mark's insight and interventions kept standards high and objectives focussed. Becky's consistent generosity and immediate understanding of any problem's intricacies were invaluable. Mark and Becky always made allowances for my commitments outside the research and the research was all the better for that understanding. Thank you both.

This research has crossed through a number of teams within different organisations. It was a pleasure to be part of the Centre for Digital Music throughout. Other members owed particular thanks for their direct contributions towards my thinking include Alo Allik, Jeff King Miller, Andrew McPherson, Giulio Moro, Johan Pauwels, Josh Reiss and Tony Stockman. Similar input was offered very generously by Lorenzo Picinali and Isaac Engel Alonso-Martinez of the Audio Experience Design team, Imperial College London. A special mention is required for Chris Yeoward and Teodor Radu, whose interest in and extension of my research made it stronger and more fruitful. I also feel fortunate to have spent three months at BBC R&D, where amongst many others Nicky Birch, Alex Roberston, Holly Challenger and Joanna Rusznica were particularly close collaborators, whose leaps of faith and help allowed me to pursue work that was truly exciting and memorable.

The practical support of the Media and Arts Technology team and community meant that this thesis never felt like a solitary endeavour. Thank you Jonathan, Nick, Geetha and Karen for always finding the way to make whatever needed to happen happen. Without the good company and cheer of MAT peers the challenge would have been far more difficult to surmount. Particular appreciation is due to James for frequent reflection and distraction, to Angela for journeys in spatial audio, and to Vanessa, Sophie, An, Mei and Nida for continual collective encouragement.

I have also benefitted from the financial support of the Media Arts Technology Centre for Doctoral Training, sponsored by the Engineering and Physical Sciences Research Council / Arts and Humanities Research Council under grant EP/L01632X/1. They funded my time, research costs, published outputs and presentations. It would not have been possible for me to achieve this work without the backing of these institutions. I will be forever grateful for having had this opportunity.

Most importantly, thank you Lucy for your half of the effort. Life will never be this hard again. I promise.

Contents

List of Figures

List of Tables

List of Abbreviations

3DoF	**T**hree **D**egrees **o**f **F**reedom
AAE	**A**uditory **A**rchive **E**xplorer
AAR	**A**udito **A**ugmented **R**eality
ASA	**A**uditory **S**tream **A**nalysis
API	**A**pplication **P**rogramming **I**nterface
BBC R&D	**B**ritish **B**roadcasting **C**orporation **R**esearch and **D**evelopment
BinVAD	**Bin**aural **V**irtual **A**uditory **D**isplay
FOA	**F**irst **O**rder **A**mbisonics
GUI	**G**raphical **U**ser **I**nterface
HCI	**H**uman **C**omputer **I**nteraction
HOA	**H**igher **O**rder **A**mbisonics
HRIR	**H**ead **R**elated **I**mpulse **R**esponse
HRTF	**H**ead **R**elated **T**ransfer **F**unction
ILD	**I**nteraural **L**evel **D**ifference
IPD	**I**nteraural **P**hase **D**ifference
ITD	**I**nteraural **T**ime **D**ifference
LUFS	**L**oudness **U**nits **R**eferenced to **F**ull **S**cale
RT	**R**esponse **T**ime
SDN	**S**cattering **D**elay **N**etwork
VAD	**V**irtual **A**uditory **D**isplay
VBAP	**V**ector **B**ase **A**mplitude **P**anning

To Dylan and Franklyn, for everything this process created and took away. Please remember the times it gave us and forgive the ones we lost.

Chapter 1

Introduction

Wish I had a dollar
For every time they say:
"Don't you miss the feeling music gave you
Back in the day?"

'Musicology'
Prince

Mobile computing, smart technology and on-demand streaming have revolutionised dominant practices in music consumption globally. Binaural synthesis techniques and even spatial auditory music browsing have been the subject of research prior to these recent innovations in digital distribution of audio. This thesis addresses the potential design opportunities and considerations that spatial auditory display might now bring, in the light of these radically transformed modes for music listening. It explores the role that interactive binaural audio (i.e. 3D sound over headphones) could play within increasingly internalised, translocated and patchwork music consumption trends, to facilitate more inquisitive user engagement with third party curated content.

1.1 Context and motivation

This research was motivated by recent patterns in digital audio consumption through smart technology. An overview of these trends is provided to give context for the investigation. Statistics on global music revenue strands and consumption patterns are referenced from data collection co-ordinated by the International Federation of the Phonographic Industry.

1.1.1 Developments in digital audio distribution

2016 marked a dual watershed in global music distribution patterns. For the first time, worldwide revenue from digital formats constituted the majority of all income from recorded music. Perhaps more significant is that this was achieved with a 60.4% increase that put streaming services into the digital ascendancy against a concomitant 20.5% decline in downloads (IFPI, 2017). By 2017, streaming generated the largest proportion of revenue at 38%, overtaking the 30% combined share of all physical formats (IFPI, 2018b). The following year, subscription streaming (as opposed to free streaming funded by advertising) became the biggest single source of income for the phonographic industry, constituting 37% of all revenue. At the time, that portion of the market incorporated 255 million members of paid services accessing over 50 million licensed tracks (IFPI, 2019a,b). Most recent data now sees subscription streaming providing 47.3% of the industry's income and ad-supported services contributing 17.7%, so approximately two thirds of all revenue (IFPI, 2022). In summary, digital delivery has become very rapidly and firmly established as the dominant mode of audio distribution.

Recent industry surveys suggest similar signs of traction for audio streaming services in consumer listening patterns. In 2019, 64% of respondents from 19 top music consuming countries had used an audio streaming service in the past month. Of all time spent listening to on-demand music in these territories, 37% was through a paid subscription to an audio provider and 15% via a free audio service, with the remainder occurring on video platforms (IFPI, 2019b). By 2021, 78% of respondents in the same territories were using audio streaming services on at least a monthly basis (IFPI, 2021). These emergent trends present an intriguing challenge for recording companies and listeners alike: how can virtually limitless content be accessed and explored effectively and enjoyably by music fans?

1.1.2 Developments in consumer smart technology

At the same time, the growth of mobile devices as a means of playing music has complemented listeners' transition towards on-demand audio streaming. In 2018, 75% of consumers reported using their smartphone to listen to music (IFPI, 2018a). Most recent statistics from 2019 indicated that 31% of all listening time occurred on mobile devices (IFPI, 2019b). Alongside the context of mobile device adoption, therefore, audio streaming does not just represent the possibility of delivering virtually any music any time, but also anywhere.

Yet smartphones are now a relatively established technology, with the iPhone celebrating its fifteenth anniversary in 2022 and the Android operating system only a year behind. Much of the leading-edge in consumer mobile device development is focussed on wearable technology. Smartwatches released in the last five years, such as the Apple Watch and Samsung Gear series, present a mechanism for remote control of audio playback from smartphones via Bluetooth. On-board WiFi functionality or – in cases such as the Apple Watch 3 and above, or the Samsung Galaxy Watch series – cellular networking even provide the capability for direct connectivity to streaming services.[1] Associated

[1] www.apple.com/uk/watch/compare/; www.samsung.com/uk/wearables/smart-watch/ [accessed 19/06/2022].

developments are seen in the emerging market for "hearables" or smart headphones and earbuds, where products such as Apple AirPods and Bose Frames devices enable speech or gestural interaction to control paired phones or watches.[2] These options for interacting with on-demand audio services via smartphones, watches or headphones present scenarios where there is in each case (respectively) restricted, little or no visual real estate for providing a graphical user interface (GUI). This presents an additional layer of complexity to the problem of interaction design for libraries with near limitless music tracks.

1.1.3 Developments in consumer extended reality technology

A further significant development in home media technology started to gain traction in 2016. A new wave of virtual reality (VR) and augmented reality (AR) headsets were released to consumer markets, delivering improved immersive experiences and gaming to the home. The most prominent devices included the HTC Vive, Sony PlayStation VR and Facebook Oculus Rift (all tethered VR units released in 2016), the Samsung Gear VR (an untethered VR solution released in late 2015 that utilised the screen of compatible smartphones for stereoscopic display) and the Microsoft Hololens (an untethered AR headset also released in 2016) (Morris, 2015).

This generation of devices was significant for the music and audio technology community, because they supported dynamic 3D audio to heighten the realism of virtual or augmented interaction (Kelly, 2016). These products therefore firmly established interactive, 3D spatialised audio over headphones (i.e. simulated surround sound that is responsive to real-time changes in listener orientation) as a relevant technology for home consumers. At the time of finalising this thesis, uptake and engagement with VR and AR units remained relatively low and uncertain, with one industry source estimating the active user base of VR and AR headsets to be less than 90 million people worldwide at the end of 2022.[3] In contrast, the release of Apple Music's Spatial Audio streaming format was only unveiled as recently as 2021, at which point the same industry source estimated the service to have 78 million subscribers.[4] This augmented audio content allows the production of 3D surround sound mixes. The Dobly Atmos standard is deployed in combination with capable headphones to enable immersive and interactive placement of the listener within a virtual sound scene.[5] Within 18 months, Apple announced that 80% of subscribers (so, upwards of 62 million listeners) were accessing Spatial Audio tracks through compatible devices and that monthly listens to immersive content had increased

[2] www.apple.com/uk/airpods/compare/; www.bose.co.uk/en_gb/products/frames.html [accessed 19/06/2022].

[3] www.statista.com/chart/28467/ [accessed 29/01/2023].

[4] www.statista.com/statistics/604959/ [accessed 29/01/2023].

[5] www.apple.com/uk/newsroom/2021/05/apple-music-announces-spatial-audio-and-lossless-audio/; www.dolby.com/technologies/dolby-atmos/ [accessed 29/01/2023].

by 1000% since its release.[6] At this juncture, therefore, deploying spatial audio to expand the horizons of recorded music consumption can be viewed as a distinct and potentially more readily adopted area of technological development than VR and AR.

1.2 Research questions

Binaural audio describes the process of delivering 3D surround sound over headphones. The commercial and consumer trends outlined above provide a clear motivation for revisiting binaural virtual auditory environments as a mechanism for music content exploration. Advances in binaural synthesis and previous research in the use of spatial audio to navigate music collections are addressed in detail in the following chapter. However, at this point it is necessary to briefly note that binaural signal processing depends fundamentally on the use of head related transfer function (HRTF) measurements. Each individual has their own unique HRTF set, which it is not possible to measure or easily replicate under normal circumstances. The choice of HRTF set used in a binaural system will therefore determine the quality of the resulting spatial effect differently, depending on the given listener.

This doctoral research focusses on the specific use case of using a binaural spatial auditory interface to access volumes of music content in flexible and fluid ways. It is assumed that such a system would deploy spatially arranged combinations of speech, synthetic sound and, of course, representative excerpts from music tracks to enable mechanisms for presenting and navigating through content. The research questions that determine the scope of this investigation are:

1. *What interaction design approaches might be employed to enable exploration of large collections of music using only 3D binaural sound as the display interface?* This question is addressed in exploratory research first in Chapter 3, then revisited summatively in Chapter 6.

2. *How can personalisation through HRTF selection be achieved to benefit interaction with music content in a 3D auditory environment using a mobile device?* This question is addressed in detail throughout Chapter 4.

3. *How can a 3D virtual auditory display system for presenting music content be implemented on a portable computing platform and within significant processing constraints?* This question is addressed in detail throughout Chapter 5, then further explored in Chapter 6.

[6] //www.apple.com/newsroom/2023/01/apple-celebrates-a-groundbreaking-year-in-entertainment/ [accessed 29/01/2023].

1.3 Thesis outline and main contributions

The remainder of the thesis's structure and its main contributions are summarised below:

Chapter 2 – *Background* outlines fundamental concepts and surveys related research that underpin the basis of this thesis.

Chapter 3 – *Binaural auditory display affordances for music exploration* investigates potential benefits of both sequential and concurrent spatial arrangement of music content binaurally, via two separate studies.

- Presents a novel design for voice-only exploration of a defined catalog of unfamiliar music, which was evaluated as a notably different interactive experience when pursued using a binaural implementation.
- Provides evidence showing that three binaurally spatialised concurrent music tracks is the maximum that can be used effectively in searching or browsing. Further demonstrates that there is no apparent advantage in accuracy or speed gained from displaying accompanying visual representations of sound source locations.

Chapter 4 – *HRTF selection for interactive auditory display of music content* assesses a mechanism for personalising 2D rendering on desktop computers, before proposing and evaluating an extended approach for 3D customisation on mobile devices.

- Establishes a set of four criteria (*reliability*, *validity*, *usability* and *efficiency*) for assessing end user HRTF selection procedures.
- Presents a novel design for HRTF selection in 3D on mobile devices, which shows demonstrable success when evaluated against the above criteria.

Chapter 5 – *Mobile 3D virtual auditory display design and evaluation* investigates the viability of an alternative method for low-cost, head-tracked, 3D binaural rendering that is more suited to interactive auditory display, compared to the dominant approach used currently for immersive media applications.

- Measures the perceptual accuracy of the proposed 'virtual vector base amplitude panning' method and demonstrates substantially improved representation of lateral sound localisation cues compared to a virtual first order Ambisonics baseline equivalent.

Chapter 6 – *Prototype binaural playlist editor* outlines and evaluates a design for customising the results returned from a search or recommender system using only touch surface gestural interaction and binaural virtual auditory display, which was rendered by the system specified in Chapter 5.

- Demonstrates that the proposed implementation was intuitively adopted and how the marginally higher error rate compared to an equivalent graphical interface can be addressed through design refinements.
- Illustrates evidence indicating that the binaural virtual auditory display system encouraged an enhanced flow state, compared to the equivalent graphical interface.

Chapter 7 – *Conclusion* reflects on the research questions from section 1.2 in light of findings from individual chapters and considers what pathways of future research and development these might present.

Timeline	Research activity	Chapters
autumn 2017	2D HRTF selection / concurrent music stream data analysis	4 & 3
spring 2018	Embedded binaural renderer development and analysis	5
autumn 2018	3D HRTF selection / binaural renderer subjective validation	4 & 5
summer 2019	Auditory Archive Explorer user study (at BBC R&D)	3
spring 2020	Binaural playlist editor user study	6

TABLE 1.1: Timeline indicating the sequence of research activity mapped to thesis chapters

Finally, it is necessary to highlight that the narrative sequence of the thesis does not directly reflect the chronology of the research undertaken in its preparation. Table 1.1 indicates, for subsequent reference, the principle phases of the investigation, their approximate starting points and how each of those strands of activity maps to the chapters outlined above and which now follow.

Chapter 2

Background

Ever feel that somehow, somewhere,
You've lost your way?
And if you don't get a help quick you won't make it
Through the day?
Could you call on Lady Day?
Could you call on John Coltrane?

'Lady Day and John Coltrane'
Gil Scott-Heron

To start the investigation this chapter surveys three strands relevant to research on auditory display of music content. It begins by outlining core principles of binaural perception and synthesis techniques that underpin all subsequent chapters in this inquiry. This is followed by a discussion of previous work related auditory music browsing systems, including early designs and their limitations, potential affordances of both sequential and concurrent sound source display, and relevant developments in mobile computing. Finally, the specific requirements and implementation techniques for effective binaural rendering on a mobile computing platform are discussed with regard to work carried out in later chapters.

2.1 Fundamentals of binaural perception and synthesis

Binaural perception and synthesis principles concern physical or virtual placement of sounds around a listener and their capacity to locate the source positions. Literature related to these domains uses the spherical co-ordinates system to specify spatial locations and this thesis is no exception. From hereon, the following convention is adopted (other than in Chapter 4, where there is detailed discussion of a third party system that uses an alternate convention for specifying azimuth):

- *Azimuth* defines horizontal position with respect to the lateral plane bisecting a listener's ears. It is expressed in angular degrees from 0° (immediately in front of the subject, coinciding with the median plane) through to +/-180°. Positive increments

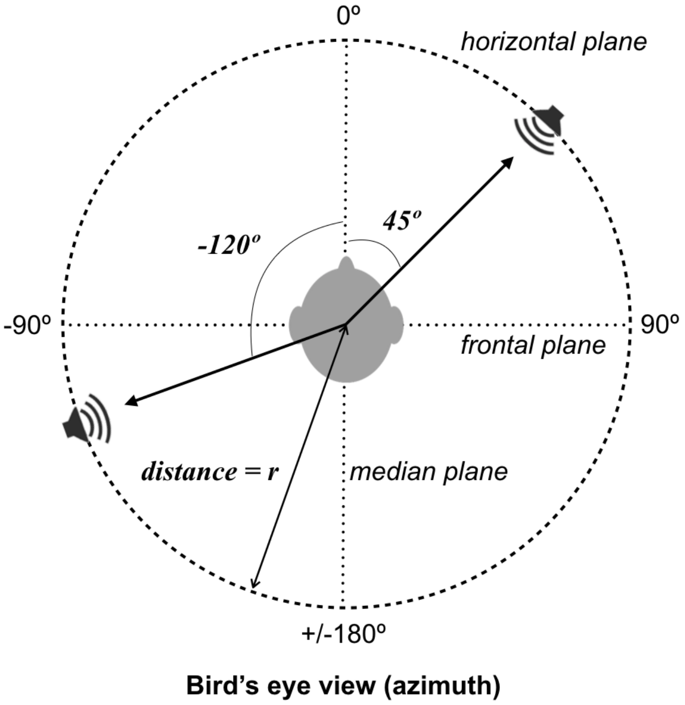


FIGURE 2.1: Illustration of co-ordinate system used to identify azimuth angle, represented from an overhead perspective.

represent positions to the right of the subject and negative increments indicate positions to their left, as illustrated in Figure 2.1.

- *Elevation* defines vertical position with respect to the median plane bisecting a listener's nose. It is expressed in angular degrees from 0° (at ear level) through to +/-90°. Positive increments represent positions above the subject's ear level and negative increments indicate positions below their ear level, as illustrated in Figure 2.2.

- *Distance* defines the separation between the listener and the sound source, as an absolute radial measurement in meters, as illustrated in Figures 2.1 and 2.2.

2.1.1 Interaural differences

The first major theoretical conceptualisation of spatial hearing was advanced by Rayleigh (1907). *Duplex theory* identified interaural time difference (ITD) and interaural level difference (ILD) as the two fundamental perceptual cues for determining horizontal sound source location in humans. These two principles continue to provide the basis for current models of lateral binaural localisation and are illustrated in Figure 2.3.

At lower frequencies, a fractional difference in the timing of sound pressure changes at either ear is interpreted to help establish the acoustic point of origin. The human hearing system allows binaural discrepancies in the phase of sinusoidal sound components

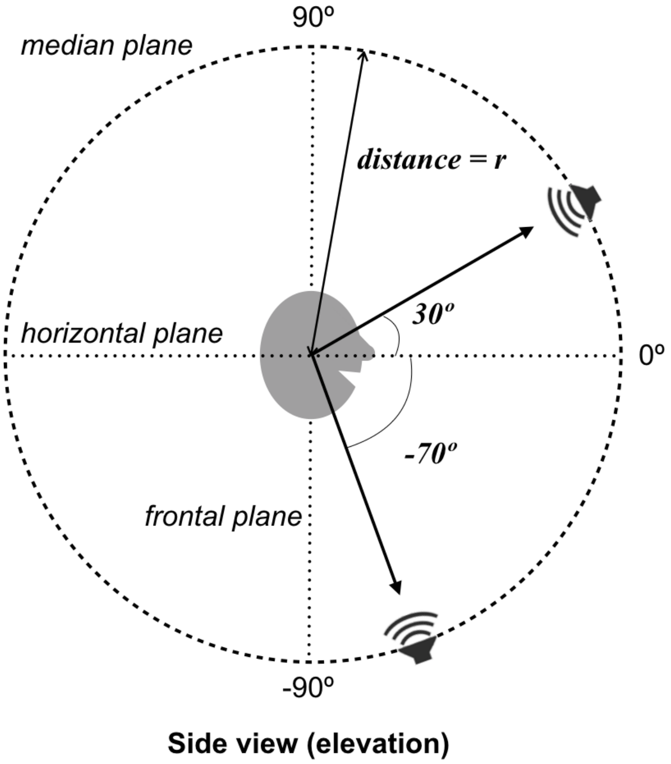


FIGURE 2.2: Illustration of co-ordinate system used to identify elevation angle, represented from a side-on perspective.

– i.e. interaural phase difference (IPD) – to be perceived and cross-correlated below approximately 1 kHz (Wenzel, Begault and Godfroy-Cooper, 2018). The precise frequency at which IPD perception starts to become confounded is dependent on individual head size and shape. However, 650 μsec (i.e. 0.00065 sec) is the generally observed maximum ITD value for an average head – i.e. in cases where the sound source is positioned on the horizontal plane (0° elevation) at +/-90° azimuth (Begault, 1994). The frequency range to which ITD as a spatial cue remains perceptually pertinent is more complex and depends on factors such as the overall spectral content and amplitude envelopes occurring in the sound source. However, there is general acceptance that the influence of ITD begins to dissipate in favour of ILD somewhere within the range of 1-1.5 kHz (Begault, 1994; Blauert, 1997; Wenzel, Begault and Godfroy-Cooper, 2018). At higher frequencies, 'shadowing' from the listener's own head reduces the intensity of a horizontally displaced sound at the contralateral ear. Under approximately 1 kHz intensity differences start to become negligible. Frequencies within this range successfully diffract around the head, since its obstructive size is smaller than the wavelength of the sound itself (Begault, 1994; Wenzel, Begault and Godfroy-Cooper, 2018).

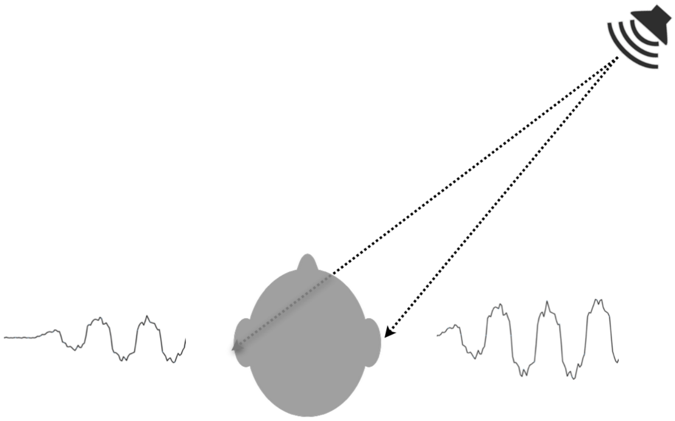


FIGURE 2.3: Illustration of ITD and ILD as identified in *duplex theory*. Sound pressure changes at the contralateral ear arrive later and with reduced intensity.

2.1.2 Monaural and binaural spectral cues

ITD and ILD alone do not determine the full range of cues through which humans discern sound direction. Perception of vertical displacement is determined primarily by shaping in the frequency spectrum of a source and is detected monaurally. The highly complex and individualised shape of the pinnae (outer ears) create occlusion, resonances and reflections with delay times ranging between 0 to 300 μsec (Begault, 1994). Together, these micro-acoustic effects result in a spectral response that is specific to a relative location (Blauert, 1997). Variations are generally most evident in frequencies above approximately 2 kHz (Begault, 1994). Particularly important for vertical perception is a prominent notch that shifts between a range of around 5 to 10 kHz, depending on the listener and source position (Wenzel, Begault and Godfroy-Cooper, 2018).

Additionally, pairs of source location that share very similar ITD and ILD cues occur along the entirety of the median plane and in positions along the surface of what is termed a *cone of confusion*, both shown in Figure 2.4. In these instances, opposing angles of origin mirrored across either the lateral plane (i.e. ’up/down confusion’), or frontal plane (i.e. ’front/back confusion’) are disambiguated due to the asymmetric shape and associated filtering effects of the pinnae (Blauert, 1997). To a lesser extent, reflection and diffraction off the head itself, shoulder reflections and occlusion by the torso all contribute (to a respectively diminishing extent) towards spectral cues that aid hemispherical discrimination (Begault, 1994; Wenzel, Begault and Godfroy-Cooper, 2018).

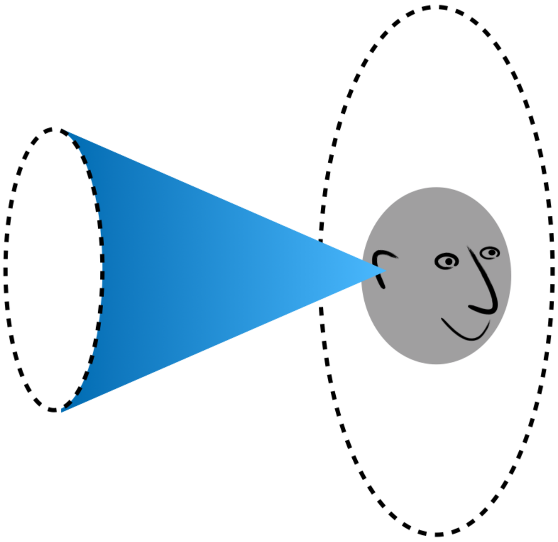


FIGURE 2.4: Illustration of the median plane and an example *cone of confusion*. Opposing angles mirrored across the frontal or horizontal planes are difficult to discriminate along the median plane, or on the surface of any imaginary orthogonal cone extending from the ear canal.

Finally, simple application of ITD and ILD over headphones is known to result in perception of the source as occurring ’inside-the-head’ by the listener. The same morphological factors that contribute towards frontal and lateral plane disambiguation count critically towards achieving a true sense of ‘externalisation’ (Blauert, 1997), though the effect of head movement and room reverberation are also significant in this regard (Roginska, 2018).

2.1.3 Binaural synthesis with head-related transfer functions

The simplest way of creating binaural audio is using a dummy head microphone to record sound generated in a physical setting. These devices mount a pair of microphone capsules within simulated ear canals of an artificial human head, as in Figure 2.5 a). Dummy head microphones incorporate designs that just model cranial morphology, or those that extend to the shoulders and even torso. The main limitation of dummy head recording is that a static binaural scene is captured from a single listening perspective. Binaural recording can be applied dynamically either by using a dummy head that is movable during recording, or instead by placing small binaural microphone capsules at the entrance of a human subject’s ear canals. In these cases, the dummy head or person can take a specific path or journey through the auditory scene, as it unfolds and is captured, to document different locations and orientations in the recording. However, a significant restriction persists even in this approach because the binaural audio playback is embedded to a fixed path (Roginska, 2018).

[1] commons.wikimedia.org/wiki/File:Georg_Neumann_Ku_100_Dummy_Head.jpg; commons.wikimedia.org/wiki/File:Binaural_Audio_Recording_Instrument.png [accessed 29/1/2023].

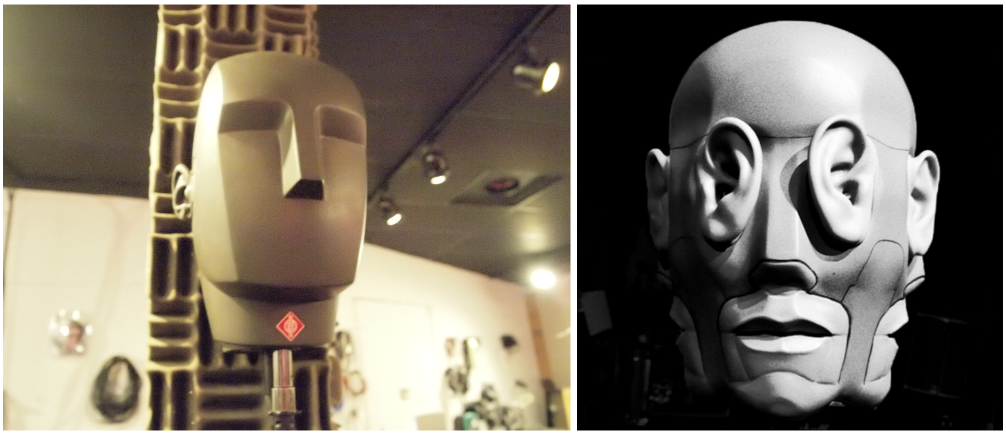


FIGURE 2.5: Example dummy head microphones. a) The Neumann KU100 (left) is characteristic of most designs in that it captures a binaural signal from a single perspective. b) A custom multi-orientation design developed by Chris Milk (right) for immersive media capture. *Sources:* EJ Posselius (CC-BY-SA-2.0), Chris Milk (CC-BY-SA-4.0)[1]

To simulate changes of direction (i.e. to make recordings that can be responsive to a listener's horizontal head movement), some designs of binaural microphone include capsule pairs positioned at multiple orientations, such as in Figure 2.5 b). In this case, a motion-tracking sensor mounted to the listener's headphones is used to track head position and interpolate the playback signal in real time between the required capture orientations (Roginska, 2018). Ultimately, however, all binaural recording methods are limited in that they depend on physical sound events occurring on a fixed timeline, within a given acoustic environment. Whilst this can provide very convincing and appropriate outcomes for linear, immersive media applications, it is not sufficiently flexible for design of interactive systems, which are by definition unfixed in sonic makeup, time and space.

In contrast, binaural signal processing enables 3D arrangement of individual sound components – either recorded or electronically generated – on demand, over headphones. Virtual sound scenes can be synthesised by filtering source signals using head-related transfer function (HRTF) sets. HRTF sets are typically derived from acoustic capture of head-related impulse responses (HRIRs). This process involves systematic measurement of directional test signals via microphones placed within the ears of a human listener or dummy head. Two approaches to microphone placement are used for human HRTF capture. In the first case, small and specially designed binaural microphones with compressible material surrounding either capsule are lodged at the entrances of the subject's ear canals. This technique is known as the *blocked meatus* approach and benefits from relatively consistent and repeatable positioning of the microphones between each measurement step. The alternative is use of small probe microphones that project into each ear canal and sit as close to the eardrum as safely possible (i.e. within 1-2 millimetres). This has the advantage of also capturing the resonant affect of the ear canal itself, which

– although directionally agnostic – is also highly individual and thought to colour sound perception between 3 and 18 kHz. However, due to the largely occluded position of the microphone capsules, this approach comes at the expense of reduced signal-to-noise ratio and inferior frequency response. With the tiny capsules obscured, it is also more challenging to position the probes accurately and consistently (Begault, 1994; Roginska, 2018).

Once the chosen microphone technique is deployed carefully on the subject, a broadband test signal is emitted via a loudspeaker at a fixed distance, within a controlled acoustic environment. The sound source is played back and digitally recorded at all of the desired spatial increments, each of which constitutes the HRIR for the given position. The accuracy of the resulting HRTF set (i.e. the correspondence between the measured and actual HRIRs) will be determined by a number of factors. Firstly, any acoustic reflections from walls or equipment that arrive at the binaural microphones within the timeframe of an HRIR will distort that measurement. Such surfaces need to be treated to eliminate, as far as possible, reflections that would coincide with the binaural measurement locations and time window. Secondly, the quality of the loudspeaker and binaural microphones used may introduce inaccuracies in the frequency response of the measured HRIRs. These errors can be be compensated for, to some extent, by precise application of post-measurement equalisation. Thirdly, any divergence in the positioning of the binaural microphones upon the subject, or of the subject's location and required orientation in relation to the loudspeaker, would also create variance between a measured and actual HRIR. This third set of challenges has a further bearing on the spatial density of HRIRs that may be realistic to capture within the HRTF set, from a logistical and practical point of view. To maintain the required consistency, an HRTF measurement facility may deploy any one or combination of the following elements (Algazi et al., 2001; Andreopoulou, Begault and Katz, 2015; Armstrong et al., 2018; Warusfel, 2003):

- a loudspeaker that can be moved, mechanically and precisely, to a variety of source locations;
- an array of loudspeakers spaced at desired increments;
- a mechanical turntable-operated seat on which the subject can remain in a set position and be automatically rotated;
- a motion capture tracking system to verify the angular positions of the subject's head and torso relative to the loudspeaker source.

Assuming that the the capture process is sufficiently accurate, convolution of an HRIR pair with a source signal simulates the net filtering that occurs at either ear, for the corresponding location. Intermediary points not captured directly within the HRTF set can be synthesised by interpolating between the nearest HRIRs. By these means, an HRTF set incorporates the data necessary to reconstruct interaural discrepancies and monaural spectral cues that form the highly individual set of psychoacoustic information described in the previous subsection (Begault, 1994; Roginska, 2018).

Binaural synthesis is enhanced by incorporating real-time head-tracking to simulate virtual sound scenes that are fixed to the physical setting, rather than locked to listener orientation. Three degrees of freedom (3DoF) head-tracking (i.e. monitoring of changes in pitch, yaw and roll) is applied in binaural synthesis to counter-rotate the spatial sound scene against changes in the listener's orientation (Begault, 1994; Roginska, 2018). This interactive processing improves both discrimination between sound source positions in front or behind (Iwaya, Suzuki and Kimura, 2003) and sense of externalisation (Brimijoin, Boyd and Akeroyd, 2013). Rendering a static sound scene, enhancing perceived location and improving externalisation are all crucial attributes to creating a coherent 3D virtual auditory experience.

2.2 Related work on spatial auditory display

At the turn of the millennium, a significant body of work emerged within HCI and signal processing communities exploring spatial auditory display of sound information. This section discusses the main outcomes from that phase of research activity relevant to this thesis.

2.2.1 Early (spatial) auditory music browsing systems

Sonic Browser was the first prominent interface devised for navigating a collection of audio using playback of sound files in a spatialised environment (Fernström and McNamara, 2005). It established the principle of direct sonification in sound catalogue interfacing, which advocated that the audio content itself should (Fernström, 2005):

- form what is displayed by the environment;
- adhere to principles of direct manipulation, i.e. be highly responsive to and representative of user actions within the system;
- feature a user-controllable *aura*, to determine the quantity of sources presented simultaneously.

Sonic Browser was initially evaluated by ten musicologists, as a tool for exploring song melodies collected through fieldwork. Participants who pursued objective search tasks with a user-controlled *aura* of 1-16 simultaneous melodies were significantly faster than those restricted to single-source browsing (by 27%, on average) (Fernström and McNamara, 2005). Later subjective evaluation by six users also showed that Sonic Browser was strongly preferred to Windows Explorer for searching sound effect files, particularly so in terms of intuitiveness and potential to facilitate discovery for unlabelled sound files (Fernström and Brazil, 2001).

In the years that followed, multiple researchers developed and evaluated systems that used similar audio processing techniques to enable preview, selection and arrangement of audio content with sound-centred interaction. Two broad developments in patterns of technology use at the time seemed to prompt this line of enquiry: i) the growth of multimedia content production on desktop computing environments; ii) the widespread adoption of portable digital music players. It was evident that users in either scenario could benefit from mechanisms to identify and explore audio content through listening, rather than visually-led interaction. A detailed survey of the motivations, designs, findings and limitations of 22 such investigations is presented by Stewart and Sandler (2012). Six of these platforms demonstrated that, when audio playback is a primary modality for the interface, users are more efficient in completing tasks like browsing for music.

Work in this area depleted 10-15 years after the burst of research activity that followed Sonic Browser's inception in 1998. Ten of the designs from this period adopted a 2D/3D map-like construct to present content, eight obviated a visual display altogether and three combined both approaches (Stewart and Sandler, 2012) . Two reasons explain why research in spatial auditory music browsing dissipated at this point. Firstly, several drivers existed for graphics-less design, including improving accessibility for blind or visually impaired users, catering for contexts that require eyes-free use and expanding usability on small- or no-screen devices. The last of these concerns was of particular value in the early years of portable digital music players, but rapidly declined in perceived importance as touch-screen and voice technology enabled more fluid interaction with smartphones and music players. Secondly, there were also three fundamental limitations in the working assumption that characterised half of these designs (Stewart and Sandler, 2011a,b; 2012):

1. *Conceptual flaws in spatial music arrangement* – Mapping-out a collection of audio assumed that sound files could be intuitively represented in a multidimensional space and, further, that a set of content can be easily and quickly presented in this way. In practice it proved difficult to meaningfully arrange something as complex as music along two or three dimensions, especially if seeking to automate the process algorithmically via self-organising maps.

2. *Expanding music collections* – As digital music devices steadily increased in storage capacity, or – from the early 2010s – were enabled to access content through cloud services, the aim of trying to navigate an entire personal or global collection as a single entity became impractical and illogical. Whilst arranging items according to measures of relative similarity assisted one particular browsing approach, this mode of interaction did not lend itself well to defined retrieval tasks (i.e. known item search), socially driven discovery, or playlist construction.

3. *Technical limitations to binaural rendering* – Only two of the interfaces that relied on relational mapping deployed either 2D or 3D binaural spatial audio. Some of the

binaural engineering considerations will be discussed in more detail in subsequent chapters. It can nevertheless be stated at this point that, even ten years ago, the computational constraints of portable devices and paucity of efficient binaural rendering implementations limited the possibilities and efficacy of spatial auditory display techniques. Implementation of head-tracking and reverberation — both of which improve accuracy of perceived localisation and, therefore, the scope and utility of the auditory display design — were not technically viable.

Despite these shortcomings, early audio-led spatial representations of music collections introduced the notion of presenting multiple simultaneous recordings as a specific mechanism for exploration and discovery. While, as noted by Stewart and Sandler (2012), this has shown potential as a means to decrease the time needed for searching or browsing, there has not yet been dedicated research to measure the optimal number of concurrently presented music pieces and whether this is dependent on other variables. Aside from the potential advantages of deploying simultaneous sound streams, the benefits of spatialising sequentially presented auditory information is also of relevance to this thesis. Known affordances of spatialising both sequential and concurrent audio information are therefore discussed further in sections 2.2.2 and 2.2.3.

2.2.2 Sequential use of spatialised auditory information display

Audio-only interaction on contemporary voice-controlled devices enables highly flexible input interfacing. It has the capability to intelligently interpret and respond to a wide range of natural language commands. One challenge for auditory interface design is how to display system feedback. Conveying returned information in an efficient and understandable form is problematic when only audio can be received and sent by the user. Spatialised sound has been explored previously as a means of addressing this limitation.

General applications

Begault (1994) describes two broad applications of binaural signal processing in interactive system design relevant to this research. First, spatial display of auditory icons (i.e. everyday sounds reflecting user actions or returned information (Gaver, 1986)) can be applied as direct substitutes for graphical user interface (GUI) elements and paradigms in non-visual systems. This approach has been typically applied to the design of auditory interfaces for the visually impaired, for example in place of standard window-based operating system visual layouts and functions. The music navigation and discovery use case addressed in this thesis differs in that: i) it is not focussed solely on users with visual impairments and ii) it is not pursued as a direct substitution for of an existing GUI, but for a dedicated and alternate immersive audio technology paradigm. Nevertheless, the principle of applying spatial sound to add or enhance interaction capability in an audio-only interface is common between the two contexts.

Second, Begault highlights how spatialisation of confirmatory notifications for touch screen systems can be applied in lieu of the haptic feedback provided by traditional hardware controls — for example in touch-screen flight simulation interfaces for training pilots. In such contexts, instant confirmation is required that intended operations had in fact been executed and others not inadvertently initiated. By spatialising a pilot's and co-pilot's sound notifications into separate audio hemispheres, both operators were aware of the other's actions without diverting their visual attention (Begault, 1994). Again, this scenario differs substantially from interactive music exploration, since the former concerns a high stakes and collaborative application with GUI elements. However, use of spatialised audio notifications in place of haptic feedback and to improve contextual orientation is another shared feature of both use cases.

Designing menu navigation

Lorho, Hiipakka and Marila (2002) advocate three- or five-option auditory menus presented sequentially, concurrently or staggered, in a left-to-right linear (stereo panned) or arc (binaural) formation. Options can be represented by speech, auditory icons, earcons – i.e. synthetic, abstract, sonic motifs conveying system messages (Blattner, Sumikawa and Greenberg, 1989) – or music. They describe two design approaches:

- *Item display* – a vertical tree where three or five items from the active hierarchical level are presented.
- *Level display* – a horizontal tree where three or five hierarchical levels that branch from the active item are presented.

Referring to music collection browsing, they note that level display enables through-browsing of parent categories, reducing navigation back and forth the content hierarchy. In a three option level display the active selection is in the central spatial position, whilst the originating category and one sub-item are presented in the left and right locations, respectively. So, when an album is selected, for instance, its artist is identified at the left position and its first track is previewed to the right. Scrolling through the album list can continue beyond the active artist and through the whole collection, since changes to the root category are updated immediately to reorient the user. The researchers go on to describe an approach to qualitative evaluation of this navigation model to ascertain how intuitive it is without visual guidance (Hiipakka and Lorho, 2003). The evaluation required ten users to:

1. complete a five part description of the interface after five minutes of unguided self-orientation;
2. answer specific questions about the structure of the music collection having been briefed fully on the input interface;

3. create a playlist of six songs comprising three styles.

The results were found to be encouraging. Seven to ten participants provided an accurate response in the case of four description elements (part 1). All but one answered every follow-up question correctly (part 2) and each successfully completed a playlist (part 3). It should be noted that the system design used only amplitude panning with stereo widening as its spatialisation technique, rather than binaural rendering.

Potential cognitive gains

Baldis (2001) analysed advantages to spatial separation of teleconference delegates' voices against qualities of memory, focal assurance (i.e. assignment of identity traits – such as opinion and attitude – to individual contributors), comprehension and preference. The study presented participants with six different configurations of four-talker conversations. These were derived from two dependent variables of visual aide presence/absence and spatial configuration (achieved via hidden loudspeakers).

Two particular outcomes from Baldis's research are pertinent to work presented later in this thesis. First, even though discussion was largely sequential, spatialisation still brought benefits to participants' information retention, attention and comprehension. Significant differences were observed between the spatialised and monaural display versions in terms of: quantitatively measured memory accuracy; self-reported recall confidence, focal assurance effort, focal assurance confidence, comprehension; overall preference ranking. Second, no practical benefits to visual information display were found, despite its inclusion being preferred by participants.

Kilgore, Chignell and Smith (2003) conducted an examination of the same effects, but using binaural spatialisation rather than loudspeakers. In that instance, they found no significant improvements in quantitatively assessed memory and self-reported confidence in recall, or in scored metrics of focal assurance. However, similar benefits were found for self-reported difficulty in speaker identification, perceived attention allocation and overall preference. One significant shortcoming was that the research used comparatively low-grade 2D (azimuth and distance) spatial rendering, without head-tracked scene stabilisation that would have placed conferees in static virtual positions.

2.2.3 Concurrent display of non-music content

The physical origin of acoustic events is an important factor in how humans process aural information into meaningful streams. Research has demonstrated that spatial position provides a strong cue for cognitive grouping of sound events. In their major work on auditory scene analysis (ASA) Bregman (1999) asserts that listeners' brains do not focus on the separate sensory impulses arriving at either ear, but instead pay attention to and switch between independent events occurring around them. Although spatial provenance can in some cases be superseded — for example by proximity of fundamental frequencies between sounds — source location has a powerful role in assisting the auditory grouping process when consistent with other cues for stream formation, such as timbral character or event predictability.

Begault (1994) points to contexts in which sound placement in 2D or 3D could augment aural representation of information in human computer interaction (HCI), including: designing auditory interfaces for the visually impaired; presenting real-time notifications for mechanical device operation; organising sound information from computing workstations; enhancing speech intelligibility in teleconferencing. In presenting their positions, both Bregman and Begault refer to the phenomenon termed the 'cocktail party effect' (Cherry, 1953), which describes cognitive processes through which we can, in a room filled with multiple simultaneous voices, isolate and focus attention on a single talker's speech. Location is understood to be a significant contributory factor in supporting the auditory stream segregation process that we rely on in environments with concurrent sound information streams.

McGookin and Brewster (2004) define four categories of information that can be conveyed in auditory display systems: sonification, speech, auditory icons and earcons. According to the definition of Fernström and McNamara (2005), auditory display systems for music represent examples of *direct sonification*. In the absence of prior research on perceptual limits to direct sonification of concurrently presented music, it is worthwhile instead to review similar work focussed on the latter three categories.

Earcons

Blattner, Sumikawa and Greenberg (1989) define earcons as synthetic, abstract, sonic motifs used to convey feedback messaging about functions or states of a computer system. Two studies have directly investigated the limits to concurrent spatialised display of earcons over headphones. Lorho, Marila and Hiipakka (2001) conducted listening tests to compare participants' ability to identify and locate five different musical instrument sounds playing simultaneous, two-second phrases in varying stereo and binaurally rendered configurations. Although subjects' accuracy was significantly better for stereophonic presentation when locating a single sound, there was no difference between stereo and binaural display of multiple sources in terms of error, response time or rating preference. Participants performed significantly better with three rather than five items in terms of accuracy and response time. However, the researchers concluded that up to five items could be used concurrently with an acceptable level of accuracy if the sound source sequence and onset interval was optimal, or even controlled by user interaction.

McGookin and Brewster (2004) highlighted that evaluation of concurrent earcon perception requires assessment of more than just a user's ability to locate a target sound, but also their capacity to elicit meaning from a complex auditory scene. Their analysis found that timbral distinction and staggered onsets were critical to improving earcon recognition accuracy and usability ratings. Though they consciously chose to avoid using spatial presentation, they presented four guidelines for concurrent earcon system design:

1. optimise the number of earcons to as few as possible;
2. use inharmonic intervals (i.e. less related musical pitches) when presenting earcons in the same register;
3. make structured use of contrasting instrument families and timbres to encode meaning in a way that minimises perceptual ambiguity;
4. apply a 300ms interval between each sound source's onset.

The work of Lorho, Marila and Hiipakka establishes a yardstick for likely perimeters of concurrent sound source perception. Earcons also leverage fundamental principles of music theory – such as pitch, timing and timbre – to communicate specific system messages. However, they differ substantially from music pieces in that they are typically monophonic or homophonic (i.e. a melody with rhythmically simple harmonic support), use a single instrumental voice, are transitory and carry a literal meaning. Recorded music is polyphonic, multi-timbral, extended and tends to be semantically abstract. Therefore, whilst McGookin and Brewster's guidelines present useful considerations for supporting perception of concurrent sound streams, points 2 and 3 cannot cannot be applied directly to recorded music, where these features are embedded and inherently complex.

Auditory Icons

In contrast to earcons, auditory icons describe the application of real-world sound events to represent data dimensions or conceptual objects in a computer system (Gaver, 1986). This typically involves mapping caricatures of everyday sounds to user actions or returned information. Prominent everyday examples include a "whooshing" projectile to represent dispatched email, motorised shutter noise to indicate capture of a digital image, and a light plastic tap to confirm a touchscreen keyboard stroke. Following their work on the Sonic Browser, Brazil and Fernström (2006) explored concurrent auditory icon recognition and its dependency on overlap of either the object or action category of sounds chosen. Their initial research examined participants' recognition of three to six randomly selected simultaneous auditory icons, compared with pre-categorised combinations that avoided co-occurrence of any object or action. Average identification success for six sources was significantly higher when there was no coincidence in objects or actions, with 89% of sources correctly described (compared to 50% in the random combination experiment). A subsequent study repeated the investigation but with three, six and nine concurrent auditory icons and again found that eliminating object and action convergence yielded improved recognition (by 47% to 39% for nine sources), though this time not to a statistically significant degree (Brazil, Fernström and Bowers, 2009). For both studies, the researchers acknowledged methodological shortcomings in the selection of auditory icons that could have lessened the strength of their findings. Consequently, they note the importance and need for further investigation into effective classification of sounds' perceptual qualities, so that optimal combinations can be reliably determined and presented.

It is possible that the question of sonic contrast (rather than McGookin and Brewster's guidelines on use of pitch and instrumentation) might be an equally relevant factor in users' ability to effectively distinguish simultaneous music signals, albeit in different terms. Categorising music recordings through computational analysis of their audio signals has been established as an effective means of organising content for some time (Pampalk, Dixon and Widmer, 2004). However, for the application being explored in this thesis, measurement of timbral or rhythmic similarity might be a useful means to inform which combination of tracks to avoid, rather than as a classification and organisation tool for typical recommendation purposes.

Speech

Binaural technology has also been of interest in the field of telecommunications. Audio engineering and HCI researchers have explored how spatialisation of multiple-talker conversations can improve memory, comprehension and usability in teleconferencing (Aguilera, Lopez and Cooperstock, 2016; Baldis, 2001; Kilgore, Chignell and Smith, 2003). Dedicated attention has been given to ascertaining the optimal conditions for presenting simultaneous speech in auditory displays. Brungart and Simpson (2005) established a preferred near-far layout for positioning seven talkers on the horizontal plane, which combined two dichotic "internalised" voices separated using standard amplitude panning, with five binaurally presented talkers distributed on a frontal arc of one metre radius (at -90°, -30°, 0°, 30° and 90°). This was accompanied by a custom method for normalising source levels that ensured each was always as intense in the ipsilateral ear as the other two or three sounds on the same side. (In the definition of this model, the centrally positioned source was allocated to the right side grouping). Tests with the seven concurrent voices found the approach improved correct decoding of a target talker's speech to a mean of 42.3%. This compared to 9.8% for non-spatialised (monophonic) presentation and 35.1% for non-normalised spatial distribution at uniform distance and azimuth increments, both of which represented statistically significant difference.

Like the application of simultaneous earcons, use of concurrent speech generates a complex auditory scene, so it is important to evaluate user comprehension as well as accuracy of talker localisation. Although they used amplitude panning rather than binaural spatial audio, Fazal, Ferguson and Johnston (2018) devised a methodology to assess user understanding of two concurrently presented speech streams taken from English language comprehension tests. They compared three spatial configurations – diotic (both streams in both ears), diotic plus monotic (one stream in both ears, the other in one ear), and dichotic (each stream in a separate ear) – and continuous versus intermittent concurrency. Wholly dichotic presentation with intermittent concurrency provided a level of comprehension that was statistically similar to the baseline measurement of sequential presentation.

2.2.4 Parallel developments in mobile computing

Since the 1990s, sound computing research has explored spatially located interactive audio within real world contexts. The virtual tour guide presented by Bederson (1995) was an early design for personalised real-time audio intervention to augment physical environments. The design possibilities for blending acoustic and digitally situated sound also started to be investigated at a similar point in time. Cohen, Aoki and Koizumi (1993) experimented with closed-ear headphones to evaluate parallel use of binaurally captured and spatially synthesised sound sources in an audio system. Sawhney and Schmandt (2000) also outlined their Nomadic Radio design for a wearable speaker communication and information system.

In addition, the potential for auditory display to permit "eyes-free" interactivity on mobile devices was recognised in HCI research prior to the emergence of smartphones as a mainstream consumer technology. Implementing their design on a personal digital assistant, Brewster et al. (2003) showed that a spatialised, egocentric (i.e. locked relative to head position), four-item audio menu paired with head gesture interaction was quicker, less frustrating to use and lower in its impact on walking speed than alternative configurations. The same study also demonstrated that touch surface interaction guided by earcon feedback significantly improved hand gesture accuracy. The latter approach was researched further by Fernström, Brazil and Bannon (2005), who showed that auditory icons elicit mental models of button-based touch interface layouts as effectively as haptic feedback — a technique they term *pseudohaptics using auditory display*.

Audio-centred multi-tasking became a further focus of HCI research as smartphones gained prevalence. For instance, Vazquez Alvarez and Brewster (2010) found that using spatial presentation to support divided attention – i.e. listening to a podcast whilst simultaneously pursuing an audio menu navigation task – was significantly harder, slower, more error-prone and less preferred than simply interrupting the continuous workstream (podcast). Yet, concurrent spatialised presentation was the preferred format in the case of selective attention. Listening to music (rather than a podcast) whilst pursuing the same audio menu navigation task showed no significant difference to simple interrupted attention baseline measures for either workload or speed (Vazquez-Alvarez and Brewster, 2011). The researchers have conducted subsequent investigations in binaural *audio augmented reality* (AAR), specifically to enhance visitor exploration and experience at artistic installations (Vazquez-Alvarez, Oakley and Brewster, 2012; Vazquez-Alvarez et al., 2016).

AAR is now a commercial concept describing smart headphone technologies with interactive capabilities and no visual display, most notably by audio technology company Bose[2]. 'Hearable' headsets feature integrated orientation and often positional sensing to enable immersive audio spatialisation. Direct or electronically assisted openness to environmental sound is also enabled in some instances. Recent evaluation of these devices has highlighted potential affordances and some initial interaction design recommendations. It has also introduced new working definitions to distinguish the variety of hardware features and modes of experience that coexist in *auditory mixed reality*, a term that 'encapsulates any auditory VR and AR experiences' (i.e. forms of sonic virtual or augmented reality) (McGill et al., 2020). These technologies have recently prompted structured experimentation with creative social gameplay (Nagele et al., 2021).

2.2.5 Summary

Binaural display of auditory information is an established approach for organising sonic interfaces. It has also been explored previously in the context of menu design for music collection navigation. Furthermore, research on teleconferencing with predominantly sequential exchanges suggests there may be inherent benefits to cognitive load from spatialisation of audio information streams. Study 1 in Chapter 3 examines potential advantages in binaural presentation of a music discovery experience comprising sequential arrangement of track excerpts, speech and auditory icons.

Existing work on simultaneous presentation of earcons and auditory icons provides guidelines that can inform how best to combine complex musical signals in a spatial auditory display. Analysis of binaurally presented concurrent speech shows evidence that, under the right conditions, co-presentation of continuous and semantically involved audio content can be digested as effectively as consecutive delivery. It follows that simultaneous binaural monitoring of two or more music recordings is an area that is also worthy of analysis. The first part of Study 2, presented in Chapter 3, provides specific insights into how concurrent music streams may be utilised in recommendation and discovery systems.

Mobile computing has progressed considerably in its capability, ubiquity and form factor since the early investigations in auditory music browsing discussed in section 2.2.1. This technological shift brings a focus on personalised interactive sound design that supplants or supplements environmental acoustics. In other words – and given the move towards on-demand streaming – a contemporary binaural music browsing experience should be adaptable to match both the perceptual profile of the user and the physical situation in which they find their self. The final study, in Chapter 6, demonstrates and evaluates how AAR principles could be applied via a mobile music discovery prototype.

[2] www.bose.co.uk/en_gb/products/frames.html [accessed 17/2/2022].

2.3 Binaural rendering for virtual auditory display

Commercial applications of binaural synthesis rarely follow the pure methodology summarised in section 2.1.3. HRTF convolution is usually combined in parallel with a more generic audio spatialisation approach. This simplifies the computational demands of encoding sound fields involving multiple sources and simulation of acoustic reflections that occur in real world spaces.

2.3.1 Ambisonics and binaural rendering for immersive media

It is notable that, of the six studies surveyed in section 2.2.3 (all but one of which span the period 2001-2009), only two explored binaural spatial audio in their research design, one used amplitude panning and three used no spatial separation at all. The difficulties of implementing effective binaural rendering at the time have already been highlighted in section 2.2.1 as a limiting factor in auditory display research over this period. McGookin and Brewster (2004) noted that 'due to computational requirements and lack of appropriate hardware, many mobile devices do not have good quality spatial positioning ability'. More recently, binaural sound has become a prominent feature in commercial entertainment media such as virtual reality, 360° video and gaming. These drivers have encouraged focus on Ambisonics as an effective means of rendering 2D and 3D sound on personal computing platforms.

Ambisonic encoding

Ambisonics is a surround sound audio format developed to reconstruct 360° sound fields over four or more loudspeakers. The underlying assumption of Ambisonic audio is to represent, as closely and as evenly as possible, all sound pressure changes occurring across the surface of a sphere at a specified point in space. In its most basic form, First Order Ambisonics (FOA) consists of four channels of sound pressure information (identified as W, X, Y and Z), which correspond to the 0th (W) and 1st order (X, Y, Z) of spherical harmonics. Higher Order Ambisonics (HOA) introduce progressively more orders of spherical harmonics to increase the spatial resolution of the encoded sound field (Nicol, 2018). Figure 2.6 illustrates how precision in the directivity of sound pressure changes captured and/or represented increases with each additional Ambisonic order included in the process.

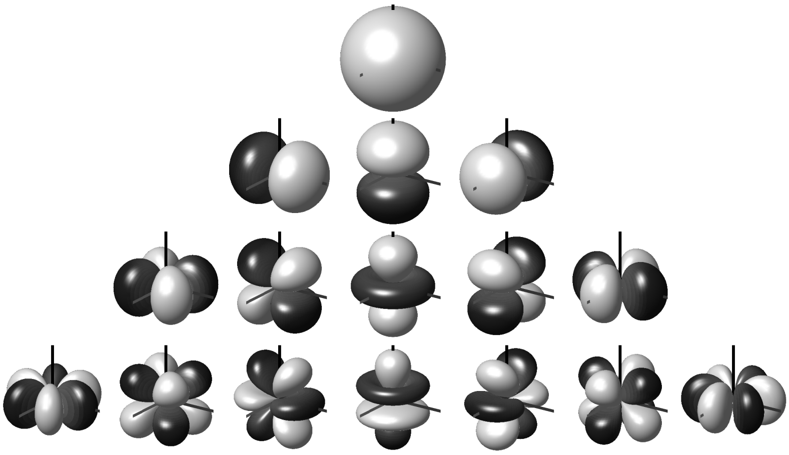


FIGURE 2.6: Illustration of 0th order (top row) to 3rd order (bottom row) spherical harmonics that underpin Ambisonic audio. *Source:* Franz Zotter (CC BY-SA 3.0)[3]

B-Format encoding is the multichannel audio storage and interchange standard developed in parallel with associated Ambisonic recording techniques and microphone technologies, which were originally aimed at transparent and faithful capture of real world sound fields (Fellgett, 1975; Gerzon, 1975). Ambisonic playback is achieved by decoding of B-Format audio into discrete signals that provide direct feeds to individual elements of a loudspeaker array. The optimal speaker array count and configuration is related to the order of the encoded multichannel B-Format signal being played back (Kearney and Doyle, 2015).

Virtual Ambisonics

Virtual Ambisonics describes the technique of rendering 2D or 3D B-Format audio over headphones by synthesising the position of a loudspeaker array binaurally using HRTF sets. In a simple model of virtual Ambisonics, the pairs of HRIR convolutions required to render the spatialised scene (irrespective of the proliferation of sound sources) is equal to the number of virtual speakers desired for the implementation. Additionally, creation of head-tracked scene rendering is made considerably simpler in virtual Ambisonics by more streamlined rotation of the sound field, rather than continual convolution and interpolation between HRIRs for individual source signal locations (Noisternig et al., 2003).

It is now well established that FOA does not adequately reproduce differences in ITD, ILD, nor the monaural or binaural spectral cues required to render spatial scenes with sufficient localisation precision or tonal transparency (Bertet et al., 2013; Thresh, Armstrong and Kearney, 2017; Wiggins, 2017). Improving spatial representation in virtual Ambisonics therefore either requires use of HOA, or more involved decoding processes. Use of HOA necessarily increases computational demands. Optimal application of more sophisticated decoding methods is an active area of research that includes approaches

[3] commons.wikimedia.org/wiki/File:Spherical_Harmonics_deg3.png [accessed 22/5/2022].

designed to improve binaural reproduction of Ambisonic signals at both first and higher orders (Kearney and Doyle, 2015; Nicol, 2018; Schörkhuber, Zaunschirm and Höldrich, 2018).

However, Ambisonics lends itself well to the aim of (re)creating immersive environments founded on representation of defined acoustic spaces and is understandably favoured for alternate reality, gaming and cinematic applications. For example, Facebook (2017) and Google (Gorzel et al., 2019) have both provided audio encoding tools for immersive content production (respectively up to second and third order), whilst games development environment Unity (2019) enables integration with binaural decoding standards based on Ambisonics. Such software is typically used to produce content for dedicated VR systems, games consoles or graphics rendering hardware, but these platforms do also support deployment to mobile devices. Mobile experiences still have a greater focus on immersion, rather than interaction. Users typically access content augmented by 2D or 3D sound fields that can be rotated according to device orientation, at relatively low computational cost.

2.3.2 Vector base amplitude panning and binaural virtual auditory display

In some contrast to the entertainment-focussed immersive media applications discussed above, virtual auditory displays (VAD) are typically multi-stream sonic environments that can be easily aurally segregated by users for information feedback or interaction purposes. To this end, there is a particular priority in spatialised VAD systems for maintaining clarity of sound source signals and accurate representation of their intended locations (Begault, 1994). Both HOA and alternate methods of B-Format decoding introduce mounting computational complexity and cost, which might not be achievable for the requirements of low-powered, portable real-time systems – i.e. embedded computing, wearable technology or even mainstream mobile platforms where only a defined proportion of CPU resource might be available for audio handling.

Vector base amplitude panning

Vector base amplitude panning (VBAP) is an alternate spatial rendering technique that extends standard amplitude panning to enable positioning of individual sound sources in full 2D or 3D surround locations. VBAP uses vector base formulations between adjacent speakers to determine the input gains required for a desired position, relative to the listener, as shown in Figure 2.7 for the case of 2D rendering. Sources are therefore placed in required locations by triangulating the relevant outputs of loudspeaker arrays. The only restriction governing the layout of the array is that each speaker must be located at uniform distance from the listener. Higher density arrays will produce more faithful representation of intended spatial position and tone (Pulkki, 1997).

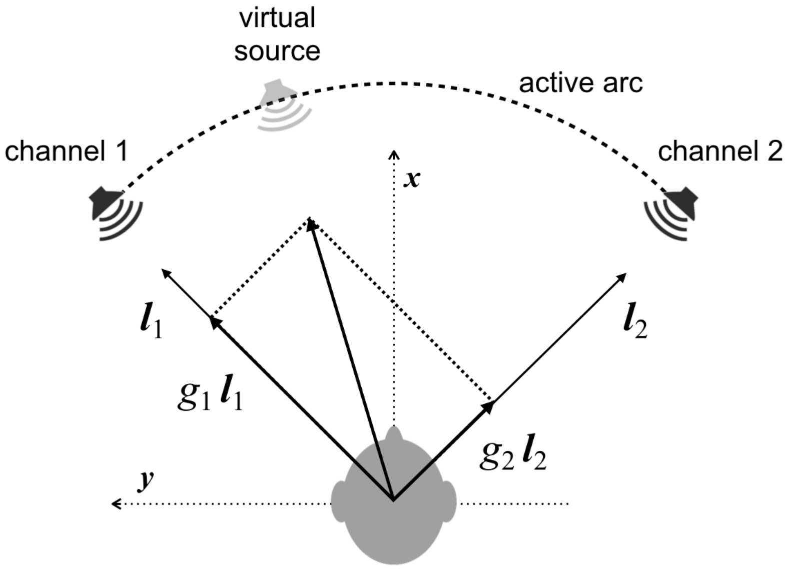


FIGURE 2.7: Illustration of VBAP with two speakers. Speaker feed gain weightings ($g_1\boldsymbol{l}_1$ and $g_2\boldsymbol{l}_2$) are derived from a vector base of speaker locations ($\boldsymbol{l}_1$ and $\boldsymbol{l}_2$) and the desired source position. The approach can be extended using further speakers horizontally and vertically, to situate sound in any directions. *Source*: recreated from Pulkki (1997).

VBAP offers the capacity for improved fidelity of source signal representation compared to Ambisonics, since the least number of speakers required to render a source is always used – i.e. one, two or three – thus minimising phase issues. The approach can therefore be applied to maximise sharpness in sound localisation and clarity of tone within predetermined spatial areas. Moreover, since VBAP positioning is achieved by simple gain weightings for each input signal to every speaker feed, it is a more efficient spatialisation technique to implement than Ambisonics (Pulkki, 1997).

VBAP has its own inherent limitations. Vertically sparse VBAP arrays are known to present errors in elevation cue representation (Pulkki, 2001a), whilst laterally situated triplets distort ITD and ILD cues towards the median plane (Baumgartner and Majdak, 2015). Nevertheless, specific configurations of VBAP speaker layouts have been shown to offer substantially improved horizontal localisation cue accuracy over comparable first and also second order Ambisonics setups (Satongar et al., 2013). Comparative evaluation using auditory modelling suggests that triplet-wise application of VBAP produces considerably more stable lateralisation than either 2D or 3D second order Ambisonics (Pulkki, 2001b).

Virtual VBAP

The principles used in virtual Ambisonics can also be applied to achieve VBAP rendering over headphones. A binaural implementation of VBAP has previously been outlined as the non-diffuse rendering component (i.e. direct and early reflected sound) of the

Directional Audio Coding (DirAC) spatial reproduction method (Laitinen and Pulkki, 2009).

The original focus of DirAC was to present a solution for improving playback fidelity of spatial audio formats over any loudspeaker array configuration. The basis of the technique is to decouple the direct and diffuse components of a B-Format input using time-frequency analysis, thus allowing direct and independent amplitude panning of both elements. This in turn reduces the problem of excessive coherence between loudspeaker inputs that is symptomatic of standard Ambisonic decoding and has been referred to previously in sections 2.3.1 and 2.3.2. Binaural DirAC is therefore a specific application of virtual VBAP that assumes, as its primary application, playback of B-Format recordings. It is also suggested, as an alternative, that DirAC analysis could be applied to other, horizontal-only multichannel formats (such as the Dolby 5.1 or 7.1 standards[4]). These formats could be encoded into either one channel (i.e. low quality, using an omnidirectional or B-Format 'W' signal) or three channels (i.e. high quality, using B-Format 'W', 'X' and 'Y' signals), with accompanying directional and diffuseness metadata for each frequency band. Such encoding would permit playback of the original format over any real or virtual loudspeaker array configuration. Nevertheless, whichever approach is taken, this application of virtual VBAP is founded on a use case of playing-back immersive sound recordings of fixed content for linear listening experiences (Laitinen and Pulkki, 2009; Pulkki, 2007) – much like binaural recording with a dummy head microphone.

The potential benefits, applications and challenges of implementing a direct virtual VBAP approach for VAD – i.e. spatialising, in real time, source signals that continually evolve based on specific interactions of the system user – have not been prominently considered in existing research literature. In particular, the relative merits and constraints of virtual VBAP for this function compared to virtual Ambisonics have not been addressed. For instance, VBAP's flexibility towards loudspeaker array configuration means it is straightforward to place a higher concentration of virtual speakers towards front-facing positions, without any additional computational effort (Pulkki, 1997). Doing so can potentially enable increased positional resolution for locations where we tend to experience greater acuity (Begault, 1994; Blauert, 1997). Moreover, in a binaural context, head-tracking is usually applied by fixing virtual speakers relative to the listener's head position whilst the sound scene itself is rotated (McKeag and McGrath, 1996; Noisternig et al., 2003). Virtual VBAP could therefore enable a defined frontal zone of improved auditory focus, which shifts automatically with head motion to any point in the virtual scene, in a way that virtual Ambisonics cannot.

On the other hand, the fact that VBAP renders each input signal as a discrete point source presents specific problems if anechoic source signals are to be auralised into a virtual acoustic space. For binaural rendering with virtual loudspeakers, this means that the location point of every sound stream – whether direct or diffuse – must ideally be

[4] www.dolby.com/about/support/guide/surround-sound-speaker-setup/ [accessed 5/5/2023].

constantly rotated to its head-tracked position, weighted and summed to relevant virtual speaker feeds before real-time convolution with HRIRs. Spatialising synthetic 3D reverberation for virtual VBAP is therefore a potentially more computationally involved task than in virtual Ambisonics, where 3D room simulation can be first generated synthetically before resulting reflections are encoded and manipulated using just a single set of rotation calculations, applied in the B-Format domain (Noisternig et al., 2003).

Reverberation synthesis with virtual VBAP

Room reverberation is recognised as a critical component of binaural rendering, which is not addressed in detail in this chapter. However, prominent research has established that early reflections alone – in this particular case defined as those arriving within the first 80 milliseconds of the room's impulse response – offer optimal conditions to improve localisation of azimuth (but at some expense to elevation perception) and increase sense of externalisation (Begault, Wenzel and Anderson, 2001). More recent research also indicates that anechoic binaural signals rendered with artificial stereo or even mono reverb could be a perceptually viable approach to room simulation. These conditions compared favourably to both first and third order Ambisonically generated surround reverbs when judged in terms of both overall realism and fidelity of source location (Picinali et al., 2017).

Scattering delay network (SDN) reverberation has been presented as a simplified means of simulating physical spaces, controlled by high-level parameters that determine properties of the room, and source and listener position. A 3D SDN reverb simulation is achieved using just six output streams – one for each surface of a cuboidal room. It has been shown to produce outputs closely aligned with Sabine and Eyring predictions of reverb time, frequency response and room surface absorption (De Sena, Hacıhabiboğlu and Cvetkovic, 2011; De Sena et al., 2015). Further investigation has also indicated that the algorithm synthesises acoustics with perceptually favourable results. In evaluation based on non-interactive simulation of two listening rooms, SDN was judged more natural than three alternative approaches commonly applied in 3D audio for room auralisation: binaural room impulse responses, geometric ray-tracing models, or feedback delay network algorithms (Djordjević et al., 2020). SDN's simplicity, its use of fewest possible output streams and positive subjective evaluation to date make it a particularly suitable approach for use with virtual VBAP spatialisation.

2.3.3 Binaural personalisation for mobile use cases

A further consideration that will determine the spatial fidelity of a binaural system is the extent to which it matches the perceptual profile of individual listeners. The potential benefits to user orientation and sense of immersion from incorporating HRTF personalisation have been highlighted in previous spatial audio experience studies as an area requiring structured investigation (Geronazzo et al., 2019; Vazquez-Alvarez et al., 2016). The quality and degree of separation between sound sources has been identified as a particularly important factor in previous research on virtual auditory display (Brungart and Simpson, 2005; Fazal, Ferguson and Johnston, 2018; Lorho, Marila and Hiipakka, 2001; Vazquez Alvarez and Brewster, 2010; 2011).

Practical obstacles to individualised binaural synthesis

As described in section 2.1.3, HRTFs are sets of location dependent filters that can be applied to simulate spatialisation of sources around a listener. An HRTF set comprises multiple HRIRs, which are measured within the left and right ear of a human or dummy head, using an excitation source placed at incremental surrounding positions. The efficacy of HRTF processing is dependent on both the density of measurements, and correlation between the HRIRs and features of the listener's morphology that affect spatial perception (chiefly the shape of their head, pinnae and upper body) (Begault, 1994; Roginska, 2018). There are numerous challenges in capturing bespoke HRTF data for the purpose of binaural synthesis, including barriers related to cost, time, expertise and specialised resources. Use of a generic or poorly matched HRTF set is liable to impact sound localisation accuracy, discrimination between sources in front/behind and above/below, externalisation (i.e. sense that a sound is emanating from outside the lister's head), and tonal clarity (Seeber and Fastl, 2003).

Subjective HRTF selection in mobile use cases

A growing body of research demonstrates the efficacy of parametric methods for either selecting or simulating best-fitting HRTF sets according to user morphology using advanced computational approaches (Geronazzo, Spagnol and Avanzini, 2010; Meshram et al., 2014; Pelzer et al., 2020; Spagnol, 2020) . However, precise and reliable acquisition of anthropometric head, ear and torso features in the case of mobile system end-users is a nontrivial challenge. User selection of preferred sets from a database of HRTF measurements is recognised as a plausible alternative means of accommodating individual requirements for spatial audio rendering over headphones (Roginska, Santoro and Wakefield, 2010; Seeber and Fastl, 2003). The primary shortcoming of this approach manifests as increased front-back confusion for users (i.e. erroneous perception of sources rendered in the front field as coming from the back) (Møller et al., 1996; Wenzel et al., 1993). However, incorporation of head-tracking is shown to mitigate this problem by allowing subtle or subconscious head rotation to verify the hemispherical orientation of virtual source positions (Begault, Wenzel and Anderson, 2001).

Subjective HRTF selection using absolute or relative judgements

Traditional subjective HRTF evaluation deploys localisation testing to gain a granular view of spatial distortions that occur when a specific set is used by any one listener. In these cases, participants are played a series of sound sources rendered binaurally and asked to make an absolute judgement on perceived virtual locations. Extents and patterns of localisation error are examined to assess the suitability of the HRTF set. This approach has been used to inform understanding of non-individualised HRTFs' limitations (Begault, Wenzel and Anderson, 2001; Møller et al., 1996; Wenzel et al., 1993) and has also been applied to demonstrate how users can potentially be trained, over time, to learn and interpret more accurately the spatialisation cues of generic HRTF sets (Berger et al., 2018; Medonca et al., 2010). One limitation of localisation testing is that judgements do not adequately take into account more qualitative aspects of HRTF suitability, such as sense of externalisation or naturalness.

More recent approaches have relied on participants' relative judgements to evaluate the apparent effectiveness of an HRTF under a range of criteria and conditions. These typically use qualitative scales for listeners to assess the perceived clarity of changes in specific parameters (such as externalisation, elevation, front-back discrimination, sense of direction, sense of distance, etc.) and have used continuous (Schönstein and Katz, 2012), fixed-point (Andreopoulou and Katz, 2016; Katz and Parseihian, 2012) or binary (Andreopoulou and Roginska, 2014; Roginska, Santoro and Wakefield, 2010; Wan, Zare and Mcmullen, 2014) metrics. Evidence suggests that repeatability of such qualitative judgements is contingent on listener expertise (Andreopoulou and Katz, 2016; Andreopoulou and Roginska, 2014; Kim, Lim and Picinali, 2020; Schönstein and Katz, 2012). A key shortcoming identified in these kind of approaches is that they are more reliably applied with expert or familiar users of binaural audio systems (Andreopoulou and Roginska, 2014; Schönstein and Katz, 2012).

Both absolute and relative judgement-based approaches also require subjects to repeatedly identify perceived locations or characteristics of sources placed in various spatial positions, for alternate HRTF sets. This is by its nature time-consuming and typically relies on a laboratory-style format of rapid serial responses to short test signal stimuli (Begault, Wenzel and Anderson, 2001; Møller et al., 1996; Wenzel et al., 1993). Such conditions would be both difficult and inappropriate to recreate for an end-user experience on a mobile device.

Subjective HRTF selection using comparative approaches

Utilising comparative judgements has already been established in psychophysical research as an effective means of assigning rank to any stimuli that must be evaluated according to a subjective perceptual response (Thurstone, 1927). Pairwise comparison has also been previously used by Iwaya (2006) as a proposed means of selecting non-individualised HRTFs. In their study, each participant started with a collection of 32 HRTF sets selected randomly from a pool of 120. They ran an adapted Swiss-style tournament (where a winner is determined using aggregated points accumulation), which eliminated any twice defeated HRTFs – meaning that not all possible pairings were presented to the listener. The comparison task used a one second pink noise burst stimulus presented in an incremental orbit, at locations 30° apart on the horizontal plane (0° elevation).

2.3.4 Summary

Virtual Ambisonics has emerged as the dominant spatialisation approach for immersive media content, but it is not necessarily the optimal format for virtual auditory display applications. Chapter 5 outlines potential requirements and solutions to binaural rendering for interactive auditory display on small devices with processing limitations. It proposes and evaluates a virtual VBAP rendering model against a comparable virtual FOA implementation via computational analysis of the ITD, ILD and spectral errors in both systems. Subjective evaluation of the VBAP renderer is then used to establish localisation error margins that inform the final design of the interactive prototype used in Chapter 6.

Personalisation of the HRTF set used in binaural rendering contributes vitally to the quality of immersive experience. There is no established method for end-user HRTF selection on a mobile device. A holistic, comparative HRTF selection approach is potentially more suited to this particular set of technological and contextual constraints, and the expectations of end-users. In these contexts, the input and feedback interface is limited, at best, to a relatively small touch screen. Further, the time and effort that users would be prepared to invest in any calibration procedure is likely to extend only to a few minutes. Chapter 4 outlines the specification, development and evaluation of an approach for 3D HRTF set selection on a mobile device.

Chapter 3

Binaural auditory display affordances for music exploration

> Back!
> Caught you lookin' for the same thing.
> It's a new thing,
> Check out this I bring.
> Uh, oh, the roll below the level
> 'Cause I'm livin' low,
> Next to the bass, (C'mon!)
> Turn up the radio!
>
> 'Don't Believe the Hype'
> *Public Enemy*

Chapter 1 illustrated how patterns of music consumption have shifted definitively towards on-demand streaming, a significant proportion of which is now pursued on mobile devices. A relatively new and growing context for music discovery and recommendation therefore has the capacity to encompass anything, anytime, anywhere and within a personal auditory space. Chapter 2 outlined core principles of binaural perception and synthesis techniques that can be applied in virtual auditory display systems. It further demonstrated how these areas had previously been a focus of experimental research in the design of music browsing environments, prior to the development and mass adoption of smartphones and audio streaming. However, there is little specific insight into how binaural virtual auditory display might now be applied to the contemporary use case of unfamiliar content being accessed in diverse contexts.

Two areas of existing research on the specific benefits of binaurally presented (non-musical) information were also highlighted in Chapter 2 as being relevant to this thesis: i) comprehension of and ability to recall audio information; ii) auditory segregation of concurrent sound streams. This chapter presents two studies that explore how these perceptual advantages may be deployed in binaurally spatialised music browsing . The first study was conducted in collaboration with BBC R&D and evaluates a model for audio-only exploration of a defined collection of new music, through a binaural prototype environment. The second study presents and analyses a prior experiment undertaken by Dr Rebecca Stewart (second supervisor for this doctoral thesis), which assesses the uses and limitations of concurrent auditory music streams in searching and browsing activity.

3.1 Study 1: An audio-only binaural music discovery prototype

The dominance of digital audio distribution and growth in smart speaker adoption present new challenges for interactive exploration of recorded music. Smart headphones are an emergent consumer technology whose unique affordances have started to be utilised commercially in recent years. This study developed a voice-led interactive application for navigating music collections on smart speaker and smart headphones. It explored potential considerations and opportunities for audio-only interaction with music content in a headphone-based spatial auditory environment, as compared to an equivalent smart speaker experience. The audio interaction design was evaluated via two prototypes, using a mixed methods user experience research construct. Both iterations featured identical content, but alternate modes of presentation. Monophonic playback was used for the smart speaker. Binaural synthesis (i.e. static placement of sound sources in virtual 3D space using head-tracking) augmented the smart headphone experience.

3.1.1 Background: voice-led content exploration with audio-only devices

Discovering unfamiliar content on audio-only devices is problematic because it relies entirely on a user's memory and current mood to motivate listening activity. There are no external triggers to guide exploration. For on-demand content, in particular, voice-led search technologies encourage either known-item requests (i.e. "play me this track/album/artist"), or unscrutinised use of provider-generated collections (i.e. "play me happy/newly released/1980s tracks"). Recent UK music industry analysis of adoption noted concerns that smart speakers could encourage listeners towards less engaged forms of interaction with audio. As acceptance of pre-curated or algorithmically generated recommendations and playlists increases, so listeners might become disconnected from individual works and artists that form such compilations. Two answers to this possible shortcoming are offered. Firstly, it is supposed that use of 'branded' recommendations or playlists with smart speakers caters to types of casual listeners who, in the past, typically used radio as background anyway. Secondly, it is suggested that these devices are simply not designed for browsing. Music discovery will still continue via other means, through which data gathered on user behaviour can be used to populate tailored recommendation lists more effectively. (BPI & ERA, 2018)

Classing voice-led audio interaction as a reductive or secondary experience would ignore three potential opportunities encompassed by this technology:

- As the same industry report goes on to note, speech-delivered search invites the possibility of more verbose and nuanced queries. These could be harnessed to achieve better recommendations if semantic analysis algorithms and metadata structures are suitably designed in future. (BPI & ERA, 2018)
- Interaction design for voice-led technology is still in its infancy. More consideration can be given to how pre-curated or auto-compiled content might be previewed and navigated efficiently with voice-controlled, audio-only devices.
- Voice-controlled devices are not limited to smart speakers. Smart headphones or 'hearables' are an emergent technology that (amongst other features) enable voice interaction for eyes-free control of a connected device. Many manufacturers now offer headphones or earbuds with inbuilt voice interaction capability and, increasingly, some form of layered or augmented reality audio, which blends transmitted with real world sounds using adaptive ambient noise cancellation. [1] Specifically, at the time this investigation was conducted in 2019, the Bose AR platform supported third party development of applications that could leverage motion detection in enabled hardware, offering the potential to deliver responsive binaural audio – or surround sound over headphones (McGill et al., 2020)[2].

The first of these three opportunities – the relationship between recommender systems and voice assistants – was far beyond the limits of this doctoral investigation. However, the second two directly informed the aims of this particular study, which were:

1. *Develop and evaluate an interactive prototype for navigation of audio content by voice.*
2. *Compare a smart headphone versus smart speaker engagement in terms of:*
    - *exploration behaviour;*
    - *recall of system design features and content.*

3.1.2 Prototype design rationale

The design process explored a novel interaction mechanism for exposing content on voice-controlled devices. This section outlines the specific objectives and method that shaped prototype development. It also describes how existing technologies and hardware were used to make fully-interactive simulations of both smart headphone and smart speaker engagements.

[1] www.wareable.com/hearables/best-hearables-smart-wireless-earbuds [accessed 17/2/2022].

[2] www.bose.co.uk/en_gb/products/frames.html [accessed 17/2/2022]. The Bose AR audio augmented reality application development platform has has since been abandoned, but the associated product range continues with a fixed set of capabilities and features.

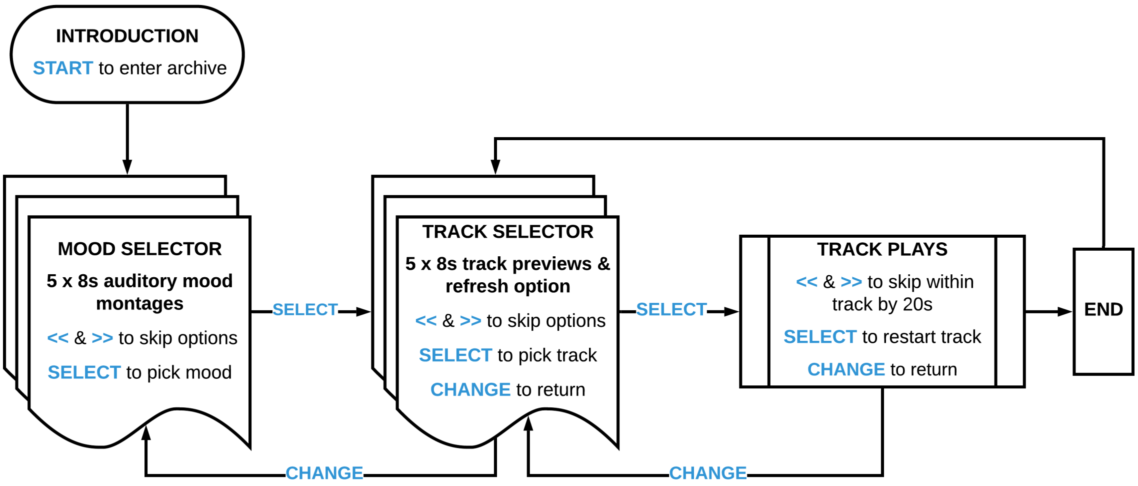


FIGURE 3.1: AAE system overview

Navigation design

The prototype sought to advance established audio-only content navigation methods surveyed in section 2.2.2. However, unlike the design advanced by Hiipakka and Lorho (2003), it was not conceived to offer a completist user experience – i.e. one with tractable pathways to each piece of featured content. The navigation design therefore had to:

- enable seamless previewing of content for rapid exploration;
- devolve a significant degree of agency to the user;
- avoid complex layers of menus;
- encourage engrossing onward journeys commonly associated with discovery platforms like YouTube.

YouTube-style engagement was a conscious conceptual influence, since it accounted for 47% of all on-demand music streaming at the time of the investigation (IFPI, 2019b). The resulting design was termed Auditory Archive Explorer (AAE) and is represented in Figure 3.1.

The prototype is populated with 50 highlights from the BBC's coverage of the Glastonbury 2019 popular music festival. Each of the 50 segments is categorised into one of five moods: *happy*, *sad*, *energetic*, *mellow* or *dark*. Users hear a 30 second introduction using synthesised speech generated from one of CereProc's commercially available voices[3]. This opening summarises the navigation concept and presents the voice commands "start", "select", "forward", "back" and "change". The latter four commands are used to both navigate through menus and control playback of audio tracks once selected, as indicated in Figure 3.1.

[3] www.cereproc.com/en/node/1166 [accessed 17/2/2022].

The 50 audio pieces are accessed via two menu layers (to select a mood and then a track). Seven different auditory icons (discussed earlier in section 2.2.3) are used for different orientation purposes. Four provide feedback confirmation for each of the voice commands. Two more are used to delineate menu items and indicate return to the first of the five presented options. A final one is used to signpost transitions between the navigation and playback contexts. The same synthetic voice adds short announcements for further contextual orientation. Menu options are represented by eight-second auditory previews of content – either short montages with a voice-over (for the five mood categories), or a representative excerpt from a given piece (in the track menu). In both cases five items repeat on a loop, but the track menu includes a 'refresh' option to repopulate the list. A five choice menu format was considered optimal because it: balances the range and cognitively manageable extent of auditory options; provides clear spatial distinction between items (i.e. a middle and inner/outer left/right options); follows the design principle already explored by Lorho, Hiipakka and Marila (2002). Once selected, the track title and artist are announced and segue over the start of playback.

Voice interaction design

Although some consideration was given to the specifics of the voice interaction design, the focus of the research was to evaluate the audio-only mode of displaying and navigating content. The efficacy of the particular voice control mechanics per se were not under investigation. Given this premise, the four navigation voice commands were defined, as far as possible, to:

- be succinct and clear for the purpose of recognition;
- echo functions familiar to visual interface paradigms (i.e. "select") and media player controllers (i.e. "forward" and "back");
- provide semantic coherence with their dual-purpose functions for navigating menus and controlling audio playback.

The inclusion of the "change" command is evidently less intuitive, since it meets these criteria less well and its function is too easily confused with "back" in the menu navigation context. It would have been preferable also to limit the voice navigation commands to just three, which is a common approach found in media playing systems (i.e. back, play/pause and forward). In these cases, it is quite typical for "back" to return to a previous context if triggered within the first three seconds of playback. "Change" was included with these drawbacks and inconsistencies fully in mind, but as a means of enabling the research construct and in the absence of an immediate and more elegant solution.

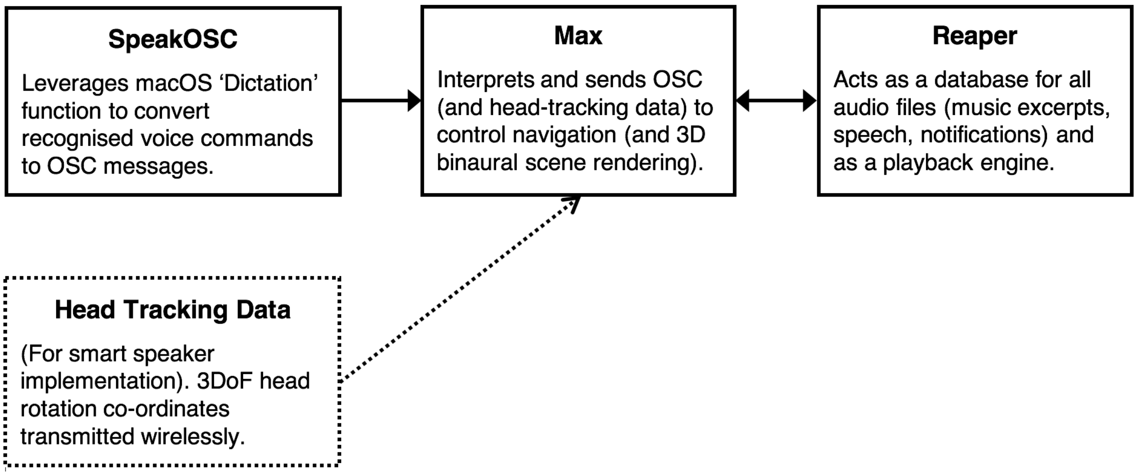

SpeakOSC
Leverages macOS 'Dictation' function to convert recognised voice commands to OSC messages.
Max
Interprets and sends OSC (and head-tracking data) to control navigation (and 3D binaural scene rendering).
Reaper
Acts as a database for all audio files (music excerpts, speech, notifications) and as a playback engine.
Head Tracking Data
(For smart speaker implementation). 3DoF head rotation co-ordinates transmitted wirelessly.

FIGURE 3.2: AAE software architecture

3.1.3 Implementation and engineering

Existing technologies and hardware were used to make a fully-interactive prototype that simulated both smart headphone and smart speaker engagements. This section details some of the software and sound engineering decisions that created the experience.

Prototype realisation

BBC R&D set a timescale of three months for conducting the investigation. This required rapid prototyping of a fully interactive system with defined perimeters of content that was comparable in function to consumer-grade technology. An initial month of experimentation resulted in the implementation architecture represented in Figure 3.2, which details the combination and integration of existing software used to create the interactive experience. Input sensing was handled by the macOS operating system's 'Dictate'[4] feature (for voice input) and low latency wireless EDTracker Pro[5] (for head orientation). The Open Sound Control (OSC) (Freed and Schmeder, 2009) communication protocol was used for cross-platform messaging, starting with the SpeakOSC[6] speech-to-OSC library. All system code was implemented in the Max[7] visual programming language, with audio content and mixing handled by the Reaper[8] digital audio workstation in real-time. A link to the source code repository is provided in Appendix A.1.

[4] support.apple.com/en-gb/guide/mac-help/mh40584/mac [accessed 17/2/2022].
[5] www.brumster.com/index.php/gaming/6-edtracker-pro-software [accessed 17/2/2022].
[6] github.com/dlublin/SpeakOSC [accessed 17/2/2022].
[7] cycling74.com/ [accessed 17/2/2022].
[8] www.reaper.fm/ [accessed 17/2/2022].

Sound design

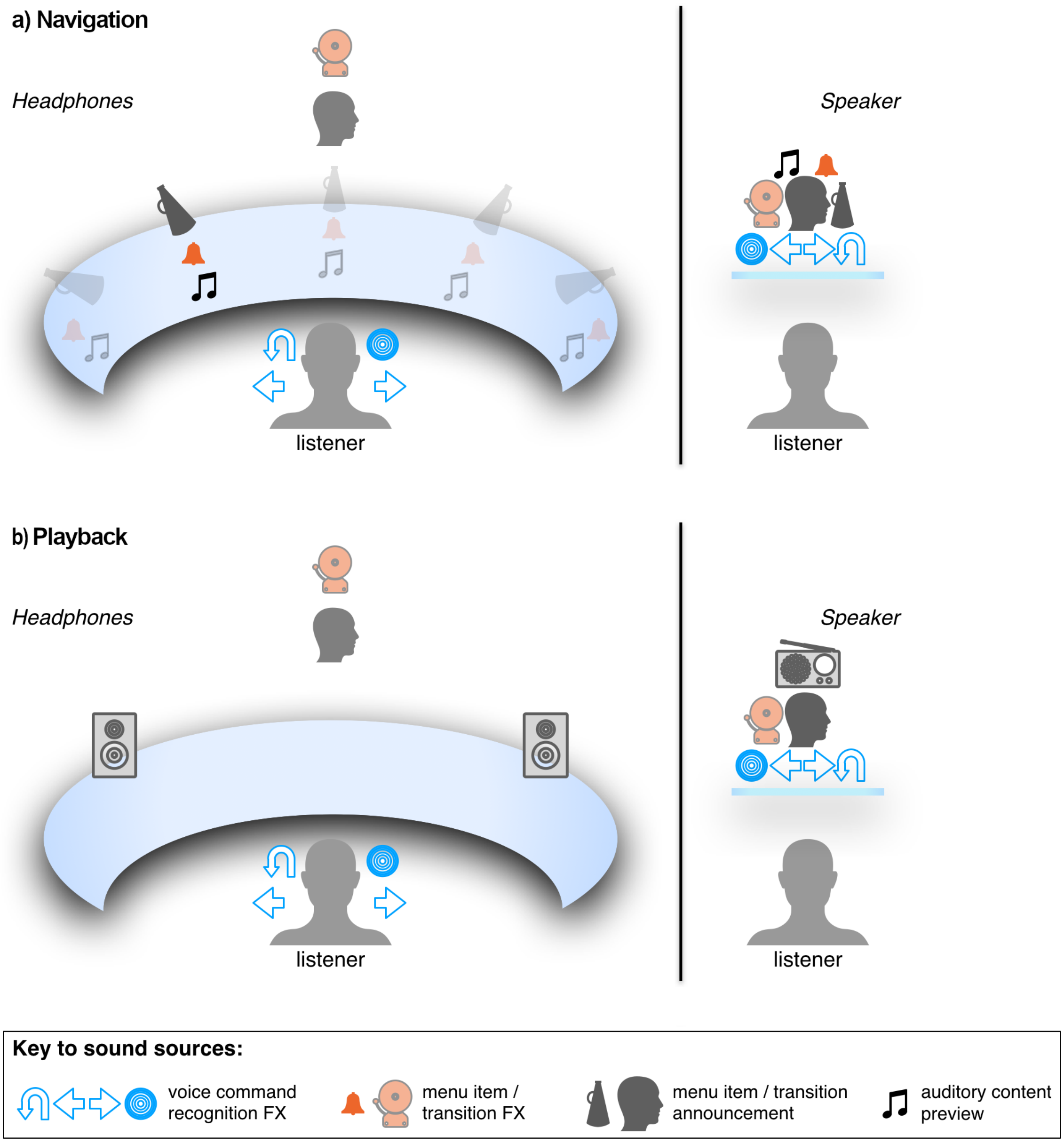


FIGURE 3.3: Sound design for navigation and audio playback in AAE

The sequential arrangement of content for the headphone and speaker simulations of AAE was identical, but the spatial sound production differed significantly. Figure 3.3 illustrates how sound positioning was used to segregate audio information streams in the smart headphone version. In navigation mode (a), menu item previews were spatialised at five positions in the front of the horizontal plane (-85°, -30°, 0°, -30° and 85° azimuth). These locations were based on the optimal separation identified in (Brungart and Simpson, 2005) for non-head-tracked binaural presentation, with one difference. Although clear spatial segregation between options was desirable, it was also judged important to present all available options to have as much virtual parity as possible in the user's attention. Left/right extremes were therefore adjusted by 5° inwards, to be perceived as slightly in front of the user, rather than perpendicular, where they could be regarded as somewhat marginalised.

Voice announcements and navigation transition effects were placed in a binaurally elevated position, to segregate these from the corresponding music excerpts, with the aim of improving intelligibility of both the spoken and musical content. Subjective informal testing determined that a 30° angle of elevation was judged sufficient to audibly separate the voiceover and transition effect from their associated music, but without pushing these two sounds too far overhead the listener. Speech recognition notifications used regular stereophonic playback (i.e. were heard from 'inside' the listener's head), so that these were more clearly associated with their own interactions. The motivation for separating the various sound sources in these ways was to explore the potential user experience benefits identified in the related work discussed in section 2.2.2 (Baldis, 2001; Begault, 1994; Kilgore, Chignell and Smith, 2003). In track playback mode (b), speaker positions were simulated binaurally over headphones, which created the impression of 'externalised' listening to augment the immersive environment.

In contrast, smart speakers typically provide monophonic playback of content. On these devices, sound sources usually have no spatial separation except through a combination of volume level and applied reverberation, for limited impression of relative distance. All audio streams in the smart speaker version were therefore co-located at the same point in space – the position of the smart speaker itself.

Binaural spatial audio rendering

As discussed in Chapter 2, effective binaural rendering is computationally expensive and contingent on a number of considerations. Hardware performance and the complexity of the binaural rendering implementation have an interdependent and fundamental bearing on fidelity and, therefore, spatial realism (Begault, 1994). Quality of experience is also dependent on the unique anatomic and cognitive makeup of individuals, so tends to be highly subjective (Blauert, 1997). At the time of this investigation, Bose AR was a prominent commercial development platform for authoring smart headphone experiences. The technical constraints in its software and associated hardware were used as a yardstick for optimising the prototype. By design, the AAE binaural implementation therefore:

- runs virtual Third Order Ambisonics (Fellgett, 1975; McKeag and McGrath, 1996; Noisternig et al., 2003; Thresh, Armstrong and Kearney, 2017) in Reaper using AmbiX plugins (Kronlachner, 2014) – equivalent to the most complex 3D audio algorithm available to Bose AR;

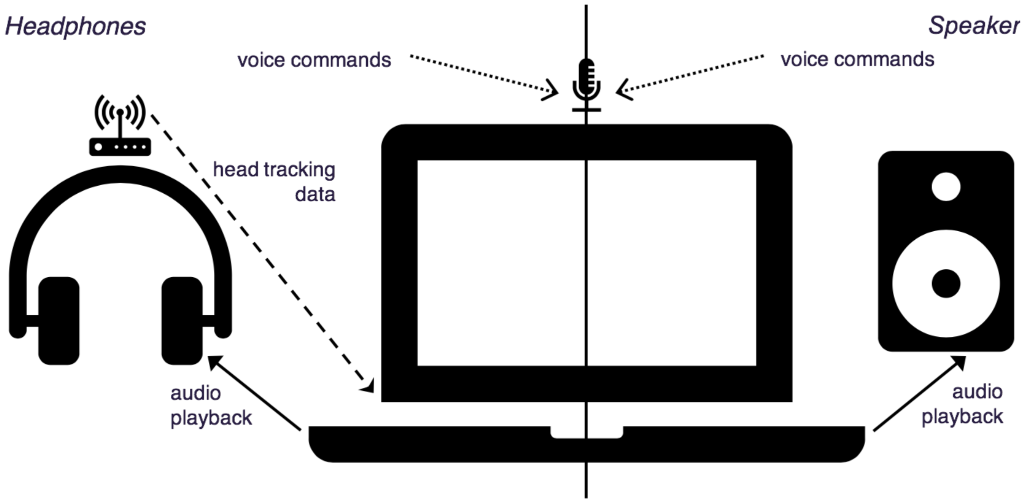


FIGURE 3.4: Hardware configuration for AAE smart headphone/speaker experiences

- uses the AmbiX 24 speaker room impulse response set[9] (at 2048 sample length with early room reflection data) for real-time rendering, without additional reverb that would be computationally challenging for mobile platforms;
- has 68 milliseconds (ms) of known system latency in digital buffering, with a likely total beyond 100ms – probably below but near the stated 196-246ms of Bose AR[10].

In short, the AAE prototype was purposefully designed with a spatial audio resolution and head-movement response time that approximated the then or near future high-street technology.

Smart technology simulation

Likewise, the smart headphone and speaker experiences were simulated with hardware that approximated the capabilities of consumer devices, as illustrated in Figure 3.4. The onboard microphone of a 13-inch 2017 model MacBook Air provided input for voice capture and all software described ran on the same device. For the headphone version, open-back wired Sennheiser HD650[11] were paired with the wireless EdTracker Pro for motion detection. The speaker version was delivered through a Zamkol ZK606[12] with wired connection.

3.1.4 User research design

Both versions of the prototype were evaluated in the BBC R&D user experience testing environment, using the processes described in this section.

[9] www.matthiaskronlachner.com/?p=2015 [accessed 17/2/2022].
[10] developer.bose.com/guides/bose-ar/end-end-system-latency (requires login) [accessed 17/2/2022].
[11] www.sennheiser-hearing.com/en-UK/p/hd-650/ [accessed 12/6/2022].
[12] www.zamkol.com/portal/article/index/id/132?lang=en-uk [accessed 11/6/2022].

	Headphones (10)	**Speaker (12)**
Age	23-38	23-34
Gender	6 female 4 male	5 female 7 male
First language	5 English 5 another	6 English 6 another
Voice assistant usage	6 infrequent 4 regular	5 infrequent 7 regular

TABLE 3.1: Participant profiles recruited for AAE user evaluation

Participant recruitment

Participants were recruited via an open email call to doctoral students across QMUL's Schools of Electronic Engineering and Computer Science, and to students at Imperial College's White City campus. Twenty-two participants were selected from 38 respondents who registered their interest online, to achieve desired balance across the criteria in Table 3.1. Headphone participants either had no previous exposure to binaural audio listening (eight people) or had only experienced the technology on a few occasions (two people). Ethical approval for the study was granted from the Queen Mary University of London (QMUL) Ethics Committee (reference 2187). Participants were given £30 in gift vouchers as compensation for the time and travel costs of attending their session.

Study format

Participants undertook an individual one-hour study session, comprising:

1. *Pre-task interview* (approx. 10 minutes)

   A short introductory discussion investigated participants' existing behaviour in music exploration and discovery. This conversation was designed to prompt participants to reflect on their current patterns of interaction when discovering new music, before experiencing AAE. The direction of the conversation was prepared and guided by BBC R&D user experience researchers, to aid interpretation of the findings in point 5, which are not included as part of the work presented here. Following this dialogue participants were given a high-level verbal summary of the prototype concept.

2. *Orientation* (approx. 10 minutes)

   Some training and supervised use of the system was required before participants completed the evaluation task. Those undertaking the headphone experience were exposed to a one minute audio demonstration. This presented a direct comparison between spatial positioning in standard stereo playback and sound placement using head-tracked binaural audio. Those undertaking the speaker experience received a short tutorial on projecting their voice with sufficient loudness and clarity to be recognised over system playback, using the keyword "testing". In either mode, participants were then given a short time to navigate the system freely. They were provided with a written prompt reiterating the four navigation commands and informed only that these enabled interaction with the system menus, and controlled track playback. They were given a minimum of 1 min 30 secs (but never more than 2 min 00 secs) to ensure they successfully completed two or more voice commands. This pre-exposure phase also included the 30 second narrated system introduction:

> *Welcome to the Glastonbury 2019 audio browser. Browse the performance highlights to suit your mood, using just your voice. You can use four voice commands to navigate the browser: "select", "forward", "back", "change". Say "start" to begin browsing or wait to hear this information again.*

Participants were invited to ask questions after the pre-exposure phase, but no further instruction or indication was given on how or when commands were to be used, or the actions they effected.

3. *Evaluation task* (15 minutes)

   Participants were given a copy of the following instructions, which they also retained for reference during the task:

   > *You have 15 minutes to explore the archive as far as you can and find six new tracks that you like. Use the pen and paper provided to make your list as you go. Make a note of the track names and artists that you choose and anything particular you liked about each one.*

   A final opportunity to ask questions was provided before starting. Again, no guidance was given about how to operate the system or access track and artist names. When ready, participants were left alone to complete the task with the prototype, which stopped responding automatically after exactly 15 minutes.

4. *Post-task questionnaire* (approx. 10 minutes)

   Participants were given a short questionnaire to assess the prototype's usability and evaluate user experience.

5. *Post-task interview* (approx. 10 minutes)

   A short closing discussion with BBC R&D user experience researchers explored participants' responses to interacting with the prototype in greater depth. The findings from that enquiry are not included as part of the work presented here.

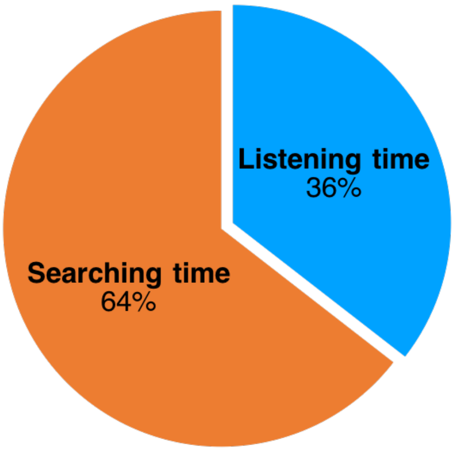


FIGURE 3.5: Aggregate AAE activity division for all participants

Data collection

Data was gathered from participants using four sources:

- automated real-time logs of their voice command interactions with AAE during the task;
- scores from a system feature recall test in the post-task questionnaire;
- ratings on usability from the post-task questionnaire;
- hand-drawn mental models of their interactions with the system.

3.1.5 Results

Data analysis was focussed on the two research objectives identified at the end of section 3.1.1: i) evaluating AAE's voice-led navigation mechanism and ii) comparing the two smart technology implementations in terms of usability and participants' recollection of system design features and content.

Use of voice-led discovery overall and between groups

Figure 3.5 illustrates the average activity of all 22 participants viewed collectively. Approximately two thirds of all time was used to browse content and a third was dedicated to listening. Table 3.2 shows that, in their 15 minutes, participants on average previewed around half of the available content and selected six tracks for further listening, but tended to fall short of their assigned target to note six new tracks they liked. In addition, all participants accessed more than one mood category. All but one participant listened to multiple tracks and 19 listened to five or more.

	Mean	Median	Standard Deviation
Previews heard *(50 tracks)*	24.05	25	6.77
Track listens *(50 tracks)*	5.64	6	1.79
Responses given *(6 track target)*	4.27	5	1.64

TABLE 3.2: Average AAE activity for all participants

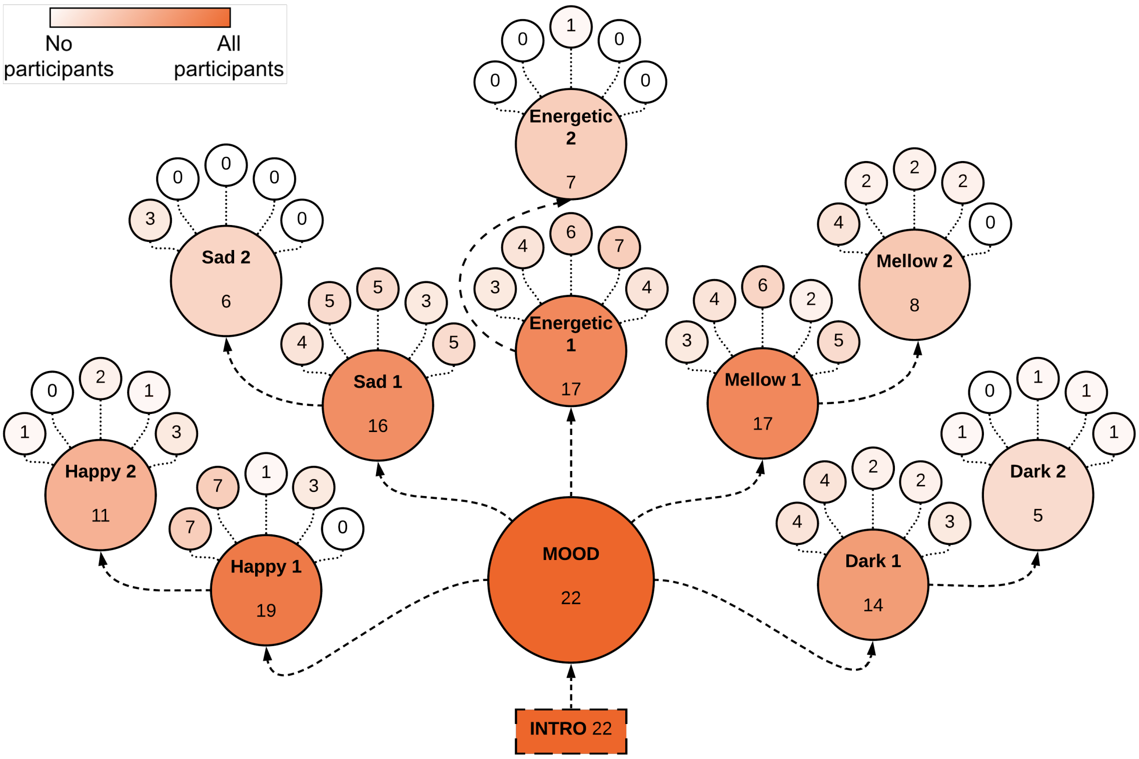


FIGURE 3.6: Heatmap of all participant visits by AAE content area. Numeric values and the depth of shading show how many participants visited, at least once, the main mood selection menu (large circle), mood track menus (medium circles) and individual tracks (small circles).

Figure 3.6 shows the combined reach of participant activity during their 15 minute task. Each circular node represents a location in AAE, spanning from the primary mood selection menu (1 large), through mood track menus (10 medium) and individual tracks (50 small). The value within circles specifies how many participants visited that point in the system at least once during their session, which is also reflected more visually across the diagram by the depth of orange shading. It can be seen clearly that all 22 participants passed through the introduction announcement and the mood selection menu. From there, participant journeys took individual paths and no other menu or track was visited

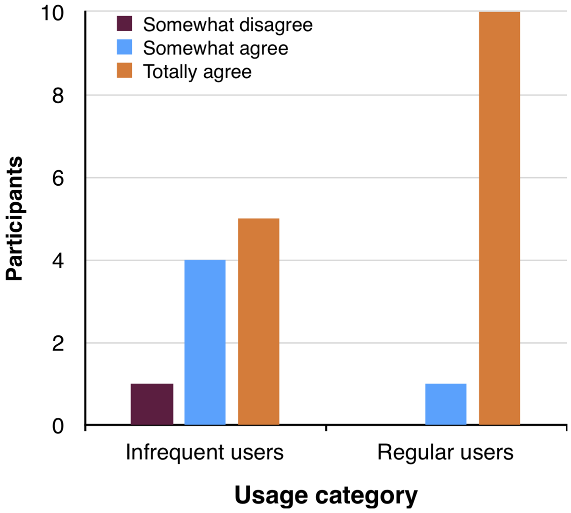


FIGURE 3.7: 'I was able to discover something new' AAE ratings by platform

by everyone. The mood track menu visited by most participants is the first set of *happy* content, which attracted 19 users. The highest number of participants previewing any one song was seven, which is the case for options one and two in the first set of *happy* content and the fourth track in the first set of *energetic* pieces.

Visit counts are seen to reduce at deeper locations – i.e. outer points in the diagram – and this is to be expected. These destinations are farther removed from the entry point and, as such, can only be reached by selecting the option to refresh at the end of the corresponding first mood track menu. For example, it is only possible to access the second set of *sad* options by progressing to the end of the first *sad* menu and using "select", when prompted, to hear more options. Reaching the secondary layer of mood track menus and songs therefore required users to persist on exploring deeper into a particular mood, rather than returning to the main menu to explore an alternative mood category. However, a more notable trend is that activity was fairly well balanced in terms of menu item precedence. Participant visits were quite evenly distributed when examined left-to-right, at all levels in the diagram. This pattern indicates that exploration seems to have been pursued relatively freely and was not unduly influenced by the arrangement of mood categories.

Prior experience with voice assistant technology appeared to influence how favourably users viewed AAE in its capacity for content discovery. Regular users included all those who self-reported daily or weekly voice assistant interaction, infrequent were those who declared their usage to be monthly or never. Regular users were significantly more likely to provide favourable responses to the statement in Figure 3.7 (Mann-Whitney U-test p = 0.045). However, equivalent disparities were not found between prior voice assistant usage groups in their responses to any of the other four qualitative usability ratings (the precise wording of which are listed separately in Figure 3.10). Additionally, there was no

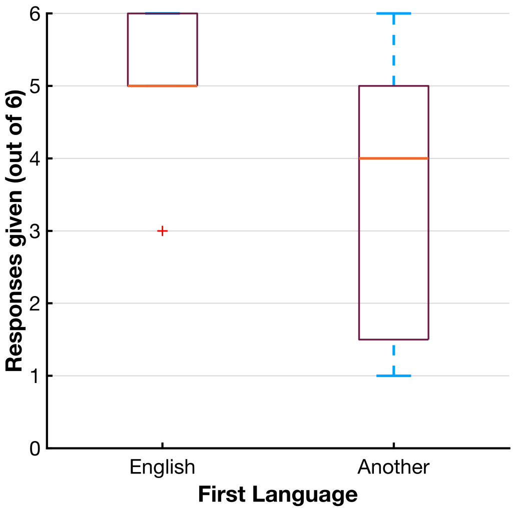


FIGURE 3.8: Distribution of the number of tracks provided at the end of the AAE task, by participant's first language. Participants were asked to find six unfamiliar tracks that they liked. Box plots indicate the 25th-75th percentile range and median value, which are significantly different between groups.

notable difference in the extent of task completion (i.e. number of written track choices given) between regular and infrequent user groups.

Participants with English as their first language were significantly more likely to progress further with completing the task (Figure 3.8, two-way ANOVA $p = 0.038$). However, there was again no notable discrepancy found between the two first language groups in their actual interactions with the prototype. The group without English as a first language tended to register a comparable count of voice commands and encountered similar proportions of preview content and full track playback. However, counterintuitively, there were no significant differences identified between language groups in their five qualitative usability ratings (Mann-Whitney U-test), which they were asked to judge following their experience (i.e. – the questions listed separately in Figure 3.10).

Exploration behaviour and usability ratings by prototype

No significant differences were found in any of the interactions (t-test), the task completion rate (Mann-Whitney U-test), or self-reported usability ratings (ANOVA) between users of either implementation. Figure 3.9 shows how similar the headphone and speaker participant groups were in their use of the respective versions of AAE over 15 minutes. The outcomes of either group's sessions – i.e. the average number of tracks listened to, previews heard and responses supplied – was closely matched. Speaker users appear, in aggregate, to have been consistently and marginally more proactive in their behaviour, using all four voice commands more frequently, on average, than the headphone group. However, in no instance was this trend to a statistically significant degree.

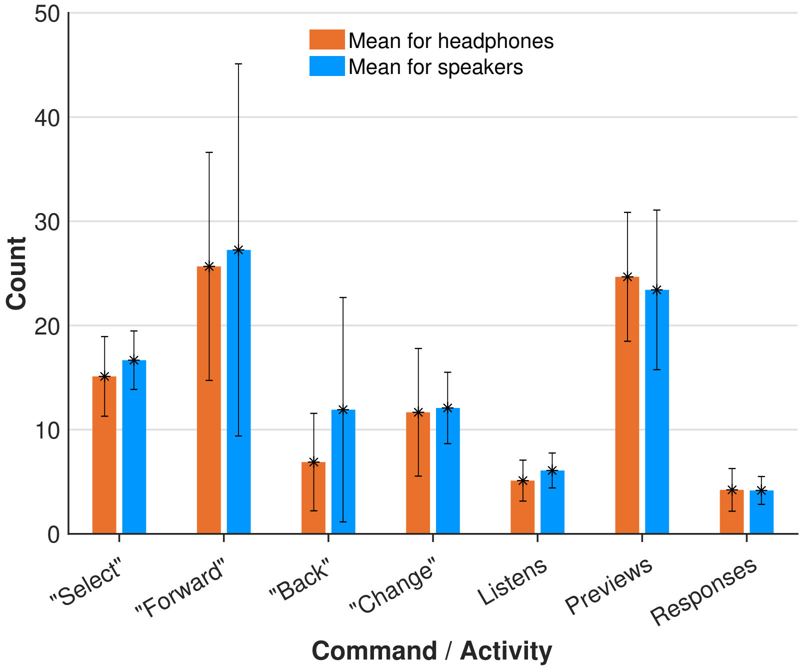


FIGURE 3.9: Participant use of AAE by platform. The chart shows the mean count (bars) and standard deviation (error bars) for use of each voice command, tracks listened to, tracks previewed and responses provided.

Likewise, Figure 3.10 illustrates that opinion expressed through usability ratings coalesced very comparably between users of either version. Both the absolute and relative values between rating statements are mirrored for the headphone and speaker responses. The headphone user group agreed slightly more strongly, in aggregate, that they were able to complete the task, navigate with ease and discover new content. The group was also more conclusive that the system was not difficult to use and did not require a particularly steep learning curve. The speaker group registered a less positive response against each of these criteria, most notably expressing slight disagreement in their capacity to complete the task. However, although the headphone ratings are consistently more favourable in this instance, again this is never to any statistically significant extent.

Recall of system design features and content by prototype

Following the evaluation task, participants were asked to recall details of the main mood menu that they encountered. Three separate questions were used to test recall of the number, name and sequence of mood categories (*happy*, *sad*, *energetic*, *mellow*, *dark*) in the primary menu of the prototype. The maximum available score was 12 and the scoring system was applied as follows:

- Two points were available for correctly identifying the number of mood categories. Zero score was awarded for an incorrect answer, or an answer that provided two

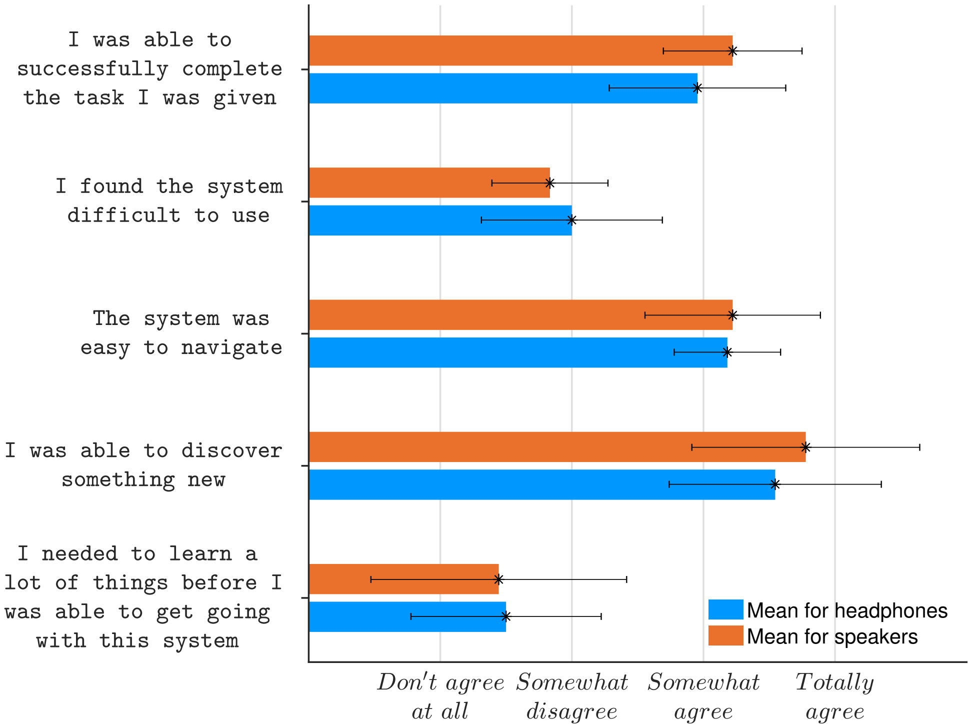


FIGURE 3.10: Participants' self-reported usability ratings for AAE, by platform. Participants were asked to indicate their agreement with five statements after completing their session. The chart shows the mean responses (bars) and standard deviation (error bars) for each question.

or more digits in response (i.e. one participant reported "5 or 6", so no points were applicable in that instance.

- A maximum of five points was available for correctly listing each of the mood categories. Phonetically spelt names that were recognisably equivalent to a category were credited (e.g. "melo" was accepted instead of *mellow*), but synonyms were not (e.g. "upbeat" was not recognised in place of *energetic* and "hard" was not accepted in place of *dark*).

- A maximum of five points was available for correctly identifying the order of mood categories. One point was awarded for every correctly positioned category. To assess instances where a participant had identified only four mood names in response to question one and two, the missing category was first inserted into its correct position within their answer to question three. Any names offered by the candidate that were correctly located following this insertion were then awarded a point. For example, "happy, sad, dark mellow" scored four points, because all provided names are in correct respective locations after *energetic* is inserted at position three. In contrast, "happy, energetic, melow [sic], sad" scored just one point, because only *happy* is correctly located once *dark* is inserted at position five.

Headphones	Speaker
- physical space and entities	- hierarchical structures
- portrayals of personal experience	- flowcharts / decision trees
- narrative explanations	- process definitions
- curves and circles	- hard angles

TABLE 3.3: Characterisation of AAE mental model illustrations, by platform

Of the twenty-two participants, twelve scored full marks and only five scored less than 9, with 5 being the lowest score registered. There was no significant difference found (t-test) in users' ability to recall the makeup of the mood menu between either version of AAE.

Participants were also asked to provide a mental model illustration of the system they had experienced ("draw a visual representation of the system you interacted with"). Submissions were viewed as a collection and discussed jointly with two BBC R&D senior user experience research professionals. Examining and comparing all of the headphone experience representations on one pinboard against the speaker collection on another, we spent approximately 30 minutes in this group of three defining and agreeing characteristic differences between the two sets of drawings. The first phase of analysis involved brainstorming of common characteristics to either set of drawings. Any one of the three researchers would offer an observation and, if the other two were in agreement, this would be noted on a whiteboard. The second phase was undertaken once this process was exhausted and no further suggestions were forthcoming. In this process, the lists and wording of the noted observations were jointly reviewed, refined and consolidated into a final form by the same three researchers.

The resulting sets of descriptions in Table 3.3 were not all evident universally, but they represent the consensus overview – arrived at by the three researchers in the process described – on the general character of the two collections, which was quite clearly distinct. A sample section from one of each group is given in Figure 3.11 to show instances where these traits were exemplified perhaps most strongly. The left hand shows a depiction that superimposed a physical journey onto the headphone experience; the right hand shows part of a logic diagram interpretation of the speaker experience. The remainder of the mental model illustrations are shown in Appendix B.

3.1.6 Discussion

The research aims identified at the end of section 3.1.1 are revisited here to analyse and interpret these results.

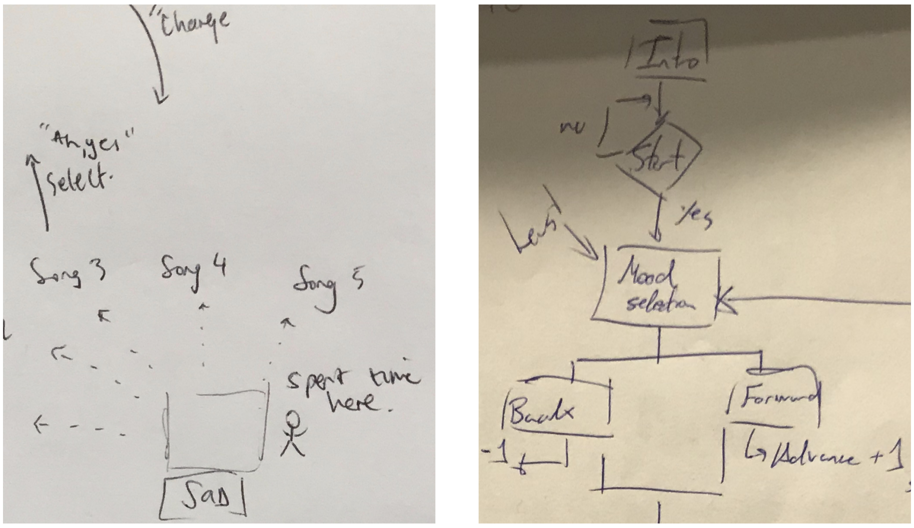


FIGURE 3.11: Example AAE headphone (left) and speaker (right) mental model illustration excerpts

Interactive voice-led discovery

Patterns of interaction suggest that, as a cohort, users had no real difficulties in navigating the system to discover new content and striving for the required response, even if they were unable to fulfil this in the allotted time (Figure 3.5 and Table 3.2). Participants were exposed to AAE for a very limited time and without any operating instructions. Both groups of users nevertheless felt confident skipping through menus to identify content of potential interest "forward" by far the most used command), but more rarely having to repeat an option ("back" used the least). The fact that "select" was used noticeably more frequently than "change" emphasises that at least a proportion of users discovered the extended uses of the former – i.e. to refresh a track menu list and/or to restart playback from any position during a piece (Figure 3.9). This quantitative data is confirmed in the self-reported usability ratings, where all responses average within the affirmative (where the statement is favourable) or negative (where the statement is pejorative) thirds (Figure 3.10). In summary, both versions of AAE seemed to enable users to onboard themselves and pursue a specific time-bound task straightforwardly and with a good degree of success.

The range and concentration of exploration presents a similarly positive outlook. This demonstrates that the system design supported exploration across all sections of the content, seemingly without any undue effect from the order in which categories or tracks were presented (figure 3.6). Though there was slightly greater traffic through the *happy* category, this could be ascribed partly to initial trial-and-error experimentation with voice commands and orientation. It could also be reasonably supposed that the slightly lower level of traffic through the *dark* route might be due to the more specialist appeal of that category.

Interestingly, the prototype's potential to aid music discovery was more apparent to those familiar with (current limitations of) accessing audio content through voice assistants (Figure 3.7). With no significant variation in the two groups' actual interaction with the system or completion rates, it would seem the benefits of voice-led exploration like AAE are more evident to those with experience of current smart assistant interaction design.

A similar interpretation can be made around the role of users' first language, where native English speakers showed significantly higher completion rates (Figure 3.8), despite there being no other notable disparities in behaviour or rating between these two categories of participant. If first language was in itself a barrier to successfully interacting with or comprehending the mechanics of the navigation, some divergence in the frequency of voice commands, volume of content visited, or qualitative evaluation of usability might also be expected in parallel. The absence of these accompanying trends points towards a possibility that language fluency did not present a barrier to engaging with the system itself, but that deciphering the spoken artist names and track titles was indeed more challenging. This is further supported by information gathered from post-task interviews, which revealed a number of comments that the pace, volume and accent (a Scottish male) of the narrator made them difficult to discern in some cases. It is worth noting within the context of these findings that transcribing artist and track names is a relatively artificial construct included for the purposes of the research study, which would not typically be pursued in a real-world content discovery journey.

Platform comparison

Results from the study suggest that the spatial arrangement of content and system notifications did not have any measurable effect on patterns of interaction, activity or task success. Nevertheless, some observations related to the two areas of examination established for the second research aim — exploration behaviour and recollection of system design features and content – are worth noting.

- *Exploration behaviour*

  Aspects of the study design could have presented potential limiting factors in exploration behaviour. It is possible that the 15-minute engagement (plus 1m30s–2m pre-exposure) was too short to establish full fluency with the system. Users will have spent a good proportion of their allotted task time to continue self-orientation with the voice commands and system structure. If differences in exploration behaviour were to emerge between the two prototype versions, it's possible they might only present when a base degree of interaction proficiency is established by users. Likewise, it is possible that the latent nature of voice interaction itself could limit users' ability to take advantage of any added perceptual orientation afforded by the headphone version. In that prototype, spatial cues communicate where, in a menu list of five, the user is currently located. If the interaction method was instantaneously responsive (as in most media playing technology), this would potentially allow skipping forward or back to previously heard menu options by relying on (memory of) its virtual position as an anchor. So, delayed responsiveness in voice interaction itself could have been a further limiting factor on users' ability to exploit any potential navigational benefits in the headphone version.

- *Participant recall of system design features and content*

  Participants' scores for recalling the main mood menu structure also suggest that spatialisation did not significantly aid memory or focus, when compared to monaural presentation. However, the qualitative summary in Table 3.3 suggests that, overall, the binaural auditory environment had a markedly different effect for headphone users in the character of their interaction and connection with the content. It is notable that this contrast in perceived experiential effect between either system resulted with a consumer-grade, non-personalised binaural rendering implementation.

3.1.7 Study 1 summary

A design for voice-led discovery of music that seeks to progress current interaction design has been outlined and evaluated for two smart technologies. Participants were seen to navigate through a variety of content, without any prominent precedence bias and with a balance of activity in line with expectations, given the task they were set. Aggregate interaction patterns and usability ratings also suggest that users were able to self-orient and navigate confidently and accurately, with some use of more advanced features. Regular smart assistant users more readily recognised the usefulness of the approach. Those without English as a first language appeared to find interpreting spoken content more challenging, but not use of the voice input interface itself.

There was no evidence from this experiment to suggest that spatial presentation of menu options influenced either interaction behaviour or storage and recall of system information from users' short-term memory. The task duration and voice interaction mode could have presented limitations on users' capacity to maximise the added binaural orientation cues. However, data gathered from users' mental model system representations strongly suggest that the smart headphone implementation did create an evidently different type of interactive experience.

3.2 Study 2a: Binaural presentation of concurrent music streams

Simultaneous auditory review of music pieces is a potentially useful tool in sound interaction design for music discovery and recommendation. It is relevant to any such pursuit that may involve rapid review of search engine or recommender results, wearable devices with restricted or no visual interface, or use cases requiring eyes-free operation. The experiment presented in this section provided insights into the perceptual considerations for navigating recorded music using spatially separated concurrent audio. A headphone-based test environment for music searching and browsing using binaural synthesis had been realised by Dr Rebecca Stewart, prior to the start of the doctoral research presented in this thesis. The test interface allows multiple music tracks to be played back simultaneously, using 2D binaural rendering. Task-oriented evaluation had been used with 22 participants to measure the influence of four system variables, against response accuracy and duration. Neither the study design, research methodology, nor the data from that research had been previously reported in any form. Full documentation and analysis of that investigation was conducted as an initial exploration of the extent to which concurrent presentation of music streams could feature in the design of a binaural auditory display system.

3.2.1 System design

Section 2.2.3 outlined why the case for investigating audio-led navigation of music content with concurrent streams is more advanced than ever before. The methodology used in this experiment was designed to explore specific factors in searching and browsing music using simultaneous sources. It sought to measure the influence of three variables – two of which were common to those examined in the existing body of work surveyed and one that extends the scope of prior research:

- the use or absence of visual references to aid sound source localisation
- the use of varied numbers of concurrently displayed sources
- the use of single or mixed-genre sound sources

Genre is of interest since it can be viewed as a broad approximation of similarity that could be a factor in source signal disambiguation. This element is in lieu of the specific, modifiable earcon attributes pitch and instrumentation discussed by McGookin and Brewster (2004) and auditory icon action/object overlap addressed by Brazil and Fernström (2006) and Brazil, Fernström and Bowers (2009), discussed earlier in section 2.2.3.

The bespoke code that implements the audio rendering, generation of random user trials and logging of participant responses was built using Python. The test graphical user interface (GUI) was produced using openFrameworks [13], which communicates with the audio and testing journey engines via Open Sound Control (Freed and Schmeder, 2009). The design assumes that a complete search and browsing system would incorporate current music information retrieval (MIR) technology, with algorithms that allow users to retrieve a subset of audio files from a collection of any size. The user request could be as

[13] openframeworks.cc/download/ [accessed 17/2/2022].

simple as searching by artist name, or a more complex query involving any number of semantic descriptors that are analysed and interpreted by a separate recommender system. MIR systems for retrieving audio files from a database may be built on one or more types of information such as metadata, social tags, user behaviours or the signal content of the audio, but in any case a ranked list of songs is returned by an intermediary algorithm (Pampalk, Dixon and Widmer, 2004). In the controlled study described here, participants used the test interface to complete a series of music search- and browse-based tasks while a selection of parameters are varied.

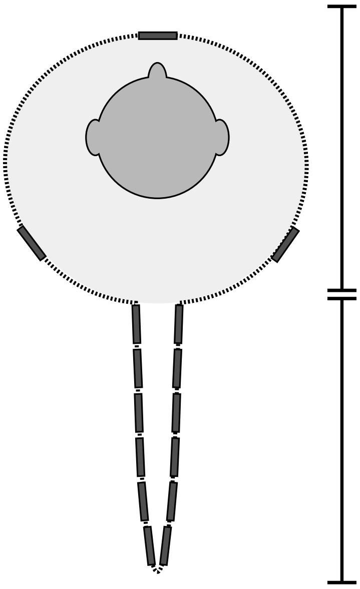
Auditory interface

A computationally simplified approach to binaural synthesis is used for the test interface. Elevation angle is restricted to 0°, such that all sources are positioned around the horizontal plane bisecting the listener's ears. Distance is also uniformly simulated at the proximal edge of what is considered in acoustics as the far field (Roginska, 2018), under anechoic conditions (i.e. without any simulation or synthesis of room reflections). The test interface presents a number of music tracks (referred to from here on as 'songs') at evenly distributed positions around the horizontal plane.

At the start of an individual listening trial, 15 songs are randomly selected and arranged into a list that can be regarded as a virtual ellipse, illustrated in Figure 3.12. The participant only hears a portion of that ellipse, in which a specified number of concurrent songs are positioned in equidistant increments 360° around their head. The participant can rotate manually through the spatialised list in real time. When any song is rotated beyond the position immediately behind them (i.e. +/- 180° azimuth) it falls out of the auditory scene and is replaced by the next one in the elliptical queue of 15. Once they have cycled through the entire collection of 15, an operating system bell auditory icon is played to indicate their return to the start of the list.

A GUI accompanies the test auditory interface to: i) enable comparison with and without visual indications of song locations; ii) provide participants with a familiar input interaction mechanism.

Two versions of the GUI exist. Version one features two visual representations of the active songs placed within the virtual auditory space (shown in Figure 3.13). The first visual prompt is presented in the large grey rectangle from a first-person viewpoint, whereby the songs seem to rotate around the listener with a direct correspondence between their auditory location and presence on the screen in front. When the sound source is in the front field (i.e. within +/- 90° azimuth) a blue square is displayed within the grey area relative to its orbital location. The second visual prompt is given in the small grey circle from a third-person bird's eye perspective, as if the user is looking down on themselves from above. This enables the participant to see where songs to either side or behind (and therefore off-screen) are located. A number displayed within this circle



The number of songs allowed to play back concurrently (3 in this illustrated case) are spread evenly over the horizontal plane bisecting both ears, with the front song slightly louder than the songs to either side. The shaded area represents the extent of the auditory scene.

The rest of the songs in the list are in a circular queue and not heard.

FIGURE 3.12: Visual representation of the virtual auditory environment.

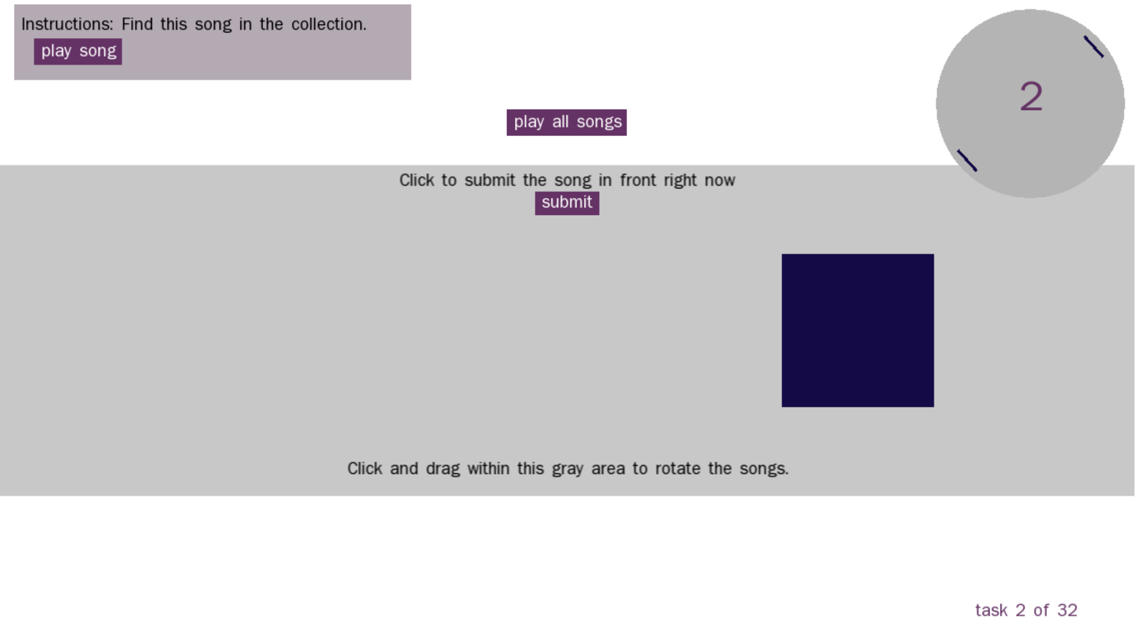
further shows how many songs are playing. In the version without the visual prompts, there is no graphical representation of the songs anywhere on the screen.

Both versions of the GUI have identical interaction mechanics. The participant interacts with the large grey rectangular area in the centre of the GUI, using a mouse to rotate the spatial image clockwise (drag right) or anticlockwise (drag left). In addition to interacting with the songs, there are GUI elements for displaying the experiment status. The upper left corner shows instructions for the active trial and the lower right corner shows the current progress.

Binaural implementation

The testing system was developed in 2012 and uses first order horizontal-only virtual Ambisonics without reverberation to render the interactive binaural scene. Although this is the simplest implementation of virtual Ambisonics, it remains the case that processing capabilities are restricted in many mobile computing contexts – either due to hardware constraints or because only limited capacity might be available for audio handling at any given time. Since this doctoral research addresses potential applications for interactive binaural audio on mobile devices, the data remains relevant to the overall enquiry contained within this thesis. The baseline reference of horizontal-only FOA is therefore examined here with the well documented perceptual limitations discussed in section 2.3 held in consideration. A potential method for improving on FOA rendering is outlined in Chapter 5.

Prior to pursuing the tasks reported in this study, each participant undertook an optimisation process to select their preferred of six different HRTF sets, drawn from two



Instructions: Find this song in the collection.

Click to submit the song in front right now

Click and drag within this gray area to rotate the songs.

task 2 of 32

FIGURE 3.13: Accompanying graphical user interface for the auditory environment (with visual location prompts included).

publicly available databases (Algazi et al., 2001; Warusfel, 2003). The procedure and outcomes of this process are reported fully in Chapter 4. Having completed the selection routine, all of the searching and browsing trials subsequently faced by each participant were rendered using their favoured set. As described earlier in this section, the test interface uses manual manipulation of the sound scene rather than head-tracking. To assist with front-back disambiguation in the absence of head-tracking, the intensity of sources in the auditory environment is exaggerated to mimic visual attention and focus. Songs located towards the sides and rear of the listener are attenuated. So, a song located at φ degrees relative to 0° azimuth (directly ahead) has a gain of:

$$g(\varphi) = e^{-\varphi^2/k} \tag{3.1}$$

where k is a constant value 10000.25, which scales the attenuation to the precise extent desired across a 180° range.

3.2.2 Experimental design

To evaluate the viability of binaural display of concurrent songs and associated presentation factors, participants were timed and tested against a series of 64 trials with the test interface.

Task description and variables

Variable	Conditions
Number of concurrent songs	1 2 3 4
Music collection genre	Genre 1 (Hip-hop) Genre 2 (Jazz) Genre 3 (Rock) Mixed
Visual Prompts	Off On

TABLE 3.4: Search and browse task variables and their conditions

Each individual trial is randomly categorised as either a searching or browsing task. For search tasks, the participant is presented with a target song – which they are able to preview without any binaural spatialisation applied – and asked to find that same song within the collection of 15 presented, as described in section 3.2.1. For browse tasks the participant is asked to choose a song that either: has a strong beat; you do not like; you like; you would listen to on your morning commute; you would listen to at the gym. No metadata is displayed for any of the songs in either the search or browse tasks. In both cases, participants submit their response by clicking on the 'submit' button seen in Figure 3.13. For search tasks, there is an additional 'play song' button to allow preview of the target song. For browse tasks, this button does not exist and instead the instruction in the top left corner states the judgement they should apply when making their selection, for example: "Choose a song that ... you would listen to at the gym".

Three further variables are built-in to the configuration of each trial, whether it is a search or browse task. The variables and their conditions are stated in Table 3.4. The number of songs played concurrently is allocated randomly and ranges from one to four. The makeup of the 15 songs presented is also determined randomly as either all jazz, all rock, all hip-hop or a mixed collection drawn from five styles – comprising the same three genres plus latin pop and electronic. The song selections and their sequence is also drawn randomly from a custom library. Appendix C explains the resources and process used to curate the music catalogue in detail. (In summary, pre-existing familiarity with any of the 15 songs used in a trial playlist was likely to influence performance. To control for this latent subject variable, an automated process to select songs with low popularity rating was devised and implemented to ensure lowest possible chances of recognition across participants.) Finally, visual location prompts are either displayed or hidden and this state is determined by random pre-allocation of the participant to one of two groups, as explained in section 3.2.3.

3.2.3 Participants and study protocol

Twenty-two volunteers participated in the study and were paid for their time at the standard rate set by New York University (NYU), which was the commissioning institution for Dr Stewart's original data collection exercise (as prefaced here in the 'Details of collaboration' on page ii). Ethical approval for the use of participants in the research was applied for and granted through the University Committee on Activities Involving Human Subjects, NYU. Subjects were drawn from a combination of departmental staff and students at NYU and QMUL and from open public calls in New York and London. Eight participants identified as female and fourteen as male. Twenty-one of the volunteers' ages were distributed across five groupings ranging from 18-24 to 40-44 and one was aged over 60. All participants self reported that they did not have a hearing impairment.

At the beginning of their session each participant was assigned randomly to either group A or B. Group A conducted the first 32 of their trials with visual prompts on and the following 32 tasks with visual prompts off. The inverse sequence was applied to group B, who conducted their first 32 trials without visual prompts, then the following 32 with visual prompts. Each participant then completed the steps outlined in Table 3.5 during their research session.

3.2.4 Results

Section 2.2.3 identified how the specific area of concurrent spatial music presentation has been under-explored in previous research. The design of this study sought to evaluate the three key considerations that emerged from related work. As such, the research scope – in terms of variables and conditions – was broad, seeking to identify presentation factors in spatial auditory display of music that require closer investigation. In view of this, the rationale for data analysis is first outlined and explained in detail. Specific variables are then subsequently analysed in turn against search task data. Search task data is a more powerful source, statistically, since it allows analysis of trial response times in conjunction with accuracy (i.e. correct versus incorrect search submissions). Browse data cannot be verified in the same way, but as a user behaviour it is also a common and important use case in digital music discovery and recommendation. A final analysis of the same variables against browse data response times is then made in light of the search task analysis.

Description	Duration
Session One	60 mins
1. Consent form to outline the study and gather personal profile data 2. HRTF selection demo video – 1m42s 3. HRTF selection procedure (as outlined later in section 4.1) 4. Trial demo video either with visual location prompts (group A) or without (group B) – 2m30s 5. One practice search and one practice browse task 6. 32 trials, either with visual location prompts (group A) or without (group B) and each randomly generated: • as either a search or browse task • with concurrent playback of one to four songs • using a playlist compiled from one of the four genre definitions 7. Part one of a short questionnaire (not included as part of the work presented here)	
Break	5-10 mins
Session Two	45 mins
1. Verbal explanation of the second session, highlighting the change in visual prompts 2. 32 more trials, either without visual location prompts (group A) or with (group B) and using the same random generation as in session one 3. Part two of a short questionnaire (not included as part of the work presented here)	

TABLE 3.5: Research study session protocol

Data cleaning and analysis approach

The participants – identified alphabetically from here as A-Z (excluding I, O, U and W) – collectively provided 701 search and 707 browse task responses. Search task trial data points ranged from 28-34 per participant and browse tasks from 30-36 per participant. This variation is due to the randomised presentation of search or browse type tasks that made up each participant's total of 64. They did not repeat the same trials either with themselves or between each other. For every trial, the 15 song playlists were compiled randomly each time from the related pool of available tracks and, in the case of a search task, the target song and its list position was also determined by random chance. Nevertheless, a high degree of uniformity in the pattern of responses is apparent across participants. This statistical consistency is demonstrated in detail first, to support subsequent aggregation of individual data points in the analysis of independent variables.

Close examination of search task response times revealed two data points that can safely be judged as accidental. In one instance, the response time was measured as 0 seconds, in the other it was measured as 1 second. In both instances the response was incorrect and the target song was not present within the active auditory scene. Under these circumstances, it is highly likely that the responses submitted were the result of an accidental double-click on the 'submit song' button, which led to unintended answers. These two data points were removed from the analysis, leaving 699 search task responses.

It is known that human response time (RT) measurements tend not to follow a normal distribution, but present as a right-skewed data that is log-normal. It is generally accepted that parametric tests can be correctly applied to response time data that is judged to fit log-normality (Lachaud and Renaud, 2011). In their study assessing participants' ability to identify auditory stimuli, Lorho, Marila and Hiipakka (2001) only included correct trial answers in search response time comparisons. Thus search RT analysis here includes correct answers only – unless otherwise stated – and is represented by 413 of the 699 valid search task data points. An Anderson-Darling test of collective RTs was consistent with the null hypothesis of log-normal distribution, with a value of p = 0.607. Log-normality of correct response time distributions is further illustrated and confirmed in Figure 3.14. Tests of RT distribution by individual participants only identified A's responses as clearly contravening log-normality, with confidence of p = 0.026. So, participant response time data shows strong patterns of log normality when taken in aggregate, and by individual participant. All subsequent ANOVA tests were applied to log transformed RT data.

A chi-square test of accuracy in search tasks showed no significant variation in accuracy between participants (χ^2 = 22.1; p = 0.453). One-way ANOVA of individuals' RTs found discrepancy involving five participants (F = 3.71; p < 0.001). Post-hoc Tukey-Kramer analysis showed that RTs for D, M and Y (faster mean) diverged from those of C and S (slower mean) with p values all at <= 0.001. (Participant M additionally differed significantly from F, G and Q, whilst S also differed from H, but in these cases at p values ranging from 0.013 to 0.048.) So, overall, there were no significant differences in response success between participants and significant variances in RT were associated with just 5 of the 22 individuals involved. The potential effect of an imbalance in time taken to

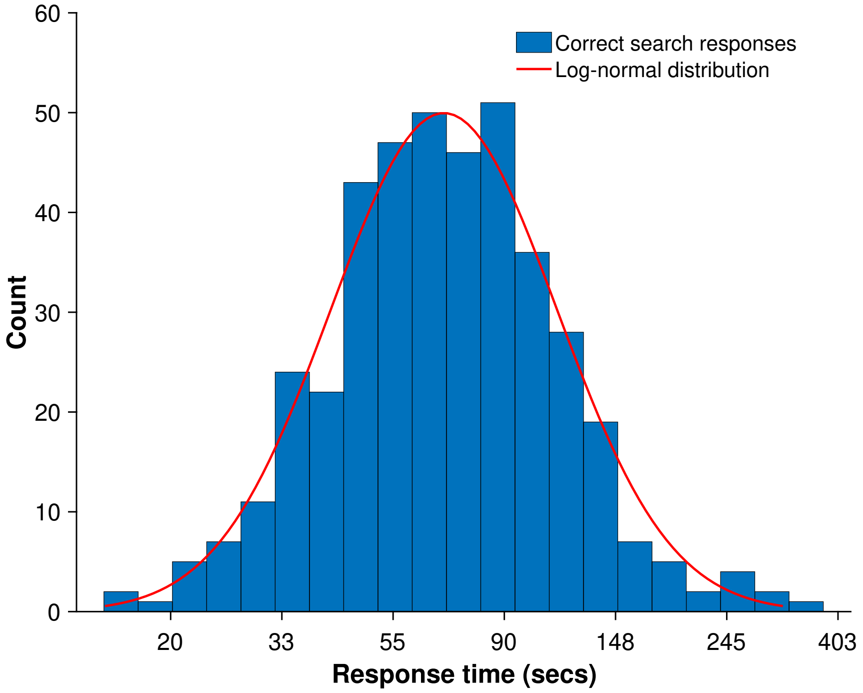


FIGURE 3.14: Distribution of all correct search task response times. Response time bins are plotted against a log transformed scale. The line of fit confirms log-normality of the data distribution.

respond correctly for a minority of participant data, between study groups A and B, is shown later in Table 4.2 and discussed in Chapter 4.

As outlined briefly in Table 3.5 and fully in Chapter 4, participants pre-selected their preferred of six HRTF sets drawn from two publicly available databases (Algazi et al., 2001; Warusfel, 2003). All six sets were represented within the 22 participants' selections. No significant variation in search accuracy ($\chi^2 = 5.31$; $p = 0.379$) or RT ($F = 0.76$; $p = 0.5826$) was observed between the HRTF sets used. So, no particular HRTF set can be viewed as having influenced either aspect of participant performance.

Search tasks analysis

Analysis of search task data addresses the three variables identified at the start of section 3.2.1. Axis scales are kept uniform to aid comparability throughout.

Owing to a software error, genre state information was logged for only 10 of the 22 participants, which reduced the data set to 321 responses and 197 correct answer RTs. Despite the reduction in data size for this variable, Figure 3.15 strongly suggests that there was no significant difference in performance between the four genre categories in terms of accuracy ($\chi^2 = 2.26$; $p = 0.519$) or RT ($F = 1.57$; $p = 0.198$). Effects were even weaker between mixed and single style accuracy ($\chi^2 = 0.54$; $p = 0.461$) and RT ($F = 0.01$; $p = 0.903$). Genre was therefore precluded as a factor in further analysis of the search task data.

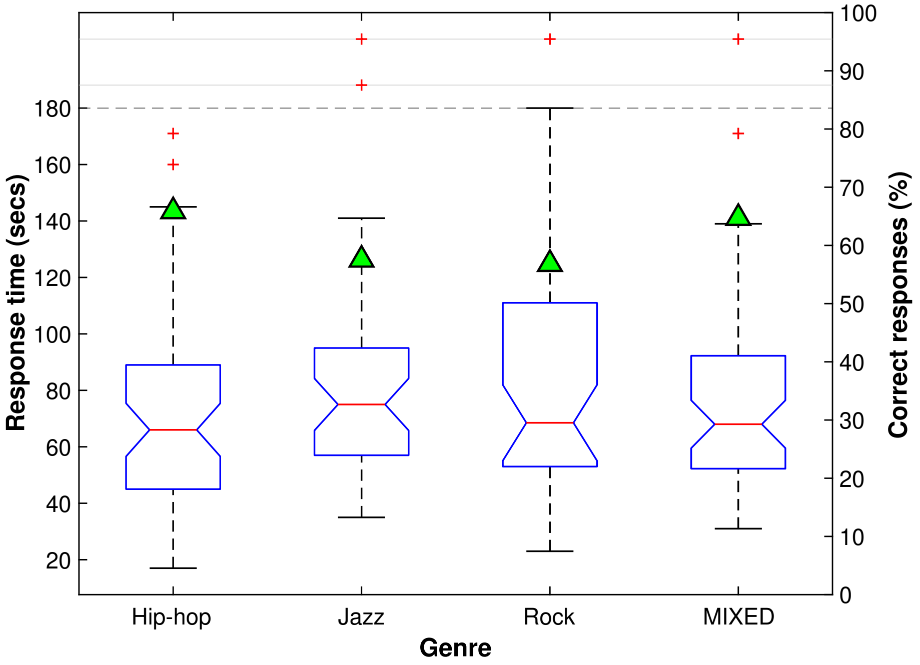


FIGURE 3.15: Response time distributions for all correct search answers, by genre. Box plots indicate the 25th-75th percentile range and median value – no differences are significant. ▲ indicates the percentage of correct responses returned for each genre (plotted against the right axis).

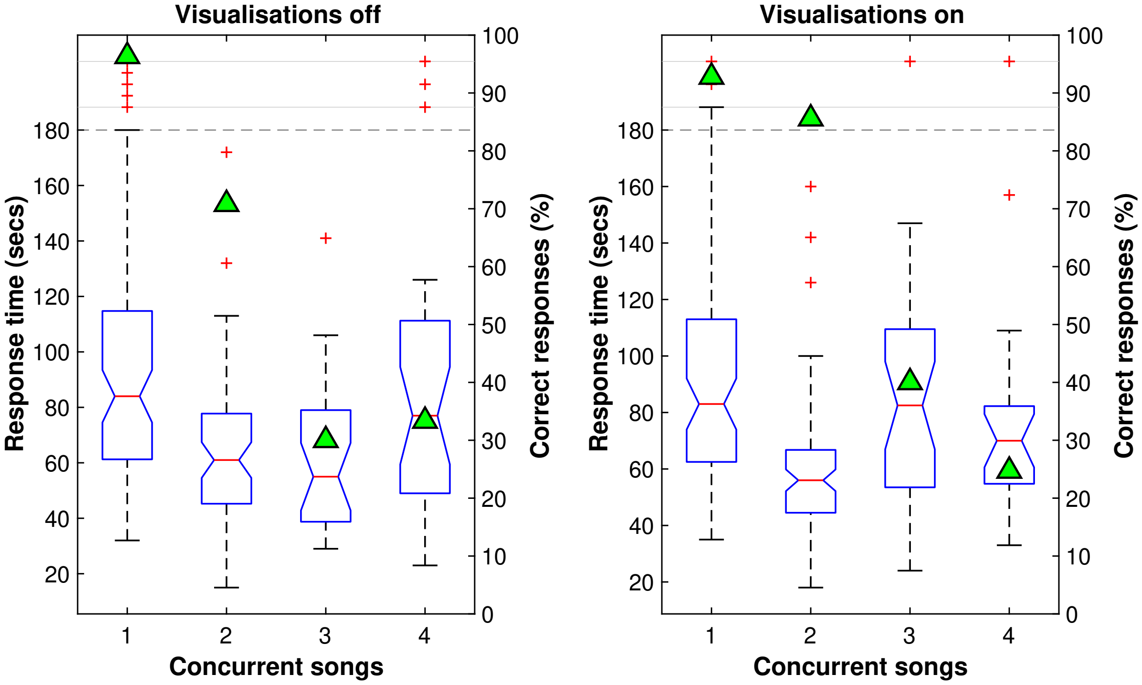


FIGURE 3.16: Response time distributions for all correct search answers, by number of concurrent songs. The separate plots show data for either visualisation state. Box plots indicate the 25th-75th percentile range and median value – significant differences exist within either plot between groups where box notches do not overlap. ▲ indicates the percentage of correct responses for each state (plotted against the right axis).

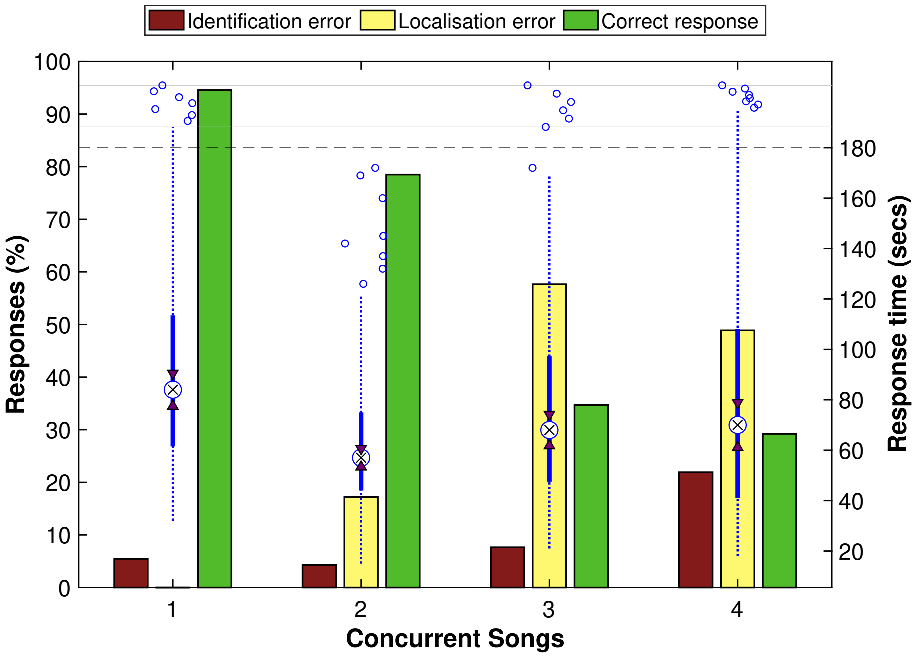


FIGURE 3.17: Proportions of all search outcomes by concurrency, with distribution of all 'in-scene' response times (against right axis). ⊗ indicate median response times.

Figure 3.16 compares the eight possible system states experienced by participants. In general, these snapshots suggest that response accuracy falls off sharply as concurrency increases above two simultaneous sources. Use of two concurrent sources was quickest in cases where visual prompts were active, but three simultaneous songs enabled quicker responses when visual prompts were deactivated. Use of four concurrent sources benefitted neither accuracy nor speed, whether visual prompts were present or not. Separate chi-square tests (for outcome) and a three-way ANOVA (for RT) evaluate the main effects of three variables related to auditory display: concurrency, visual prompt state, and order of exposure to the visual prompts (detailed in section 3.2.3). As anticipated and as suggested in Figure 3.16, the number of simultaneous songs impacted significantly both accuracy ($\chi^2 = 222.31$; $p < 0.001$) and RT ($F = 16.59$; $p < 0.001$). However, visual location prompts were not found to be a significant factor in either search accuracy ($\chi^2 = 1.99$; $p = 0.159$) or RT ($F = 0.09$; $p = 0.77$).

Response errors fall into two categories: i) identification errors – where the target song was not present in the active scene heard by the listener at the moment of their response; ii) localisation errors – where the correct song was present at the time of answering, but it was not positioned within the required frontal area. Error type i) can be regarded as a failure in musical recognition, whereas type ii) can be interpreted as a consequence of difficulties in listener orientation with the binaural scene. The latter type is grouped with correct answers as "in-scene" responses, whereby the target song has been recognised

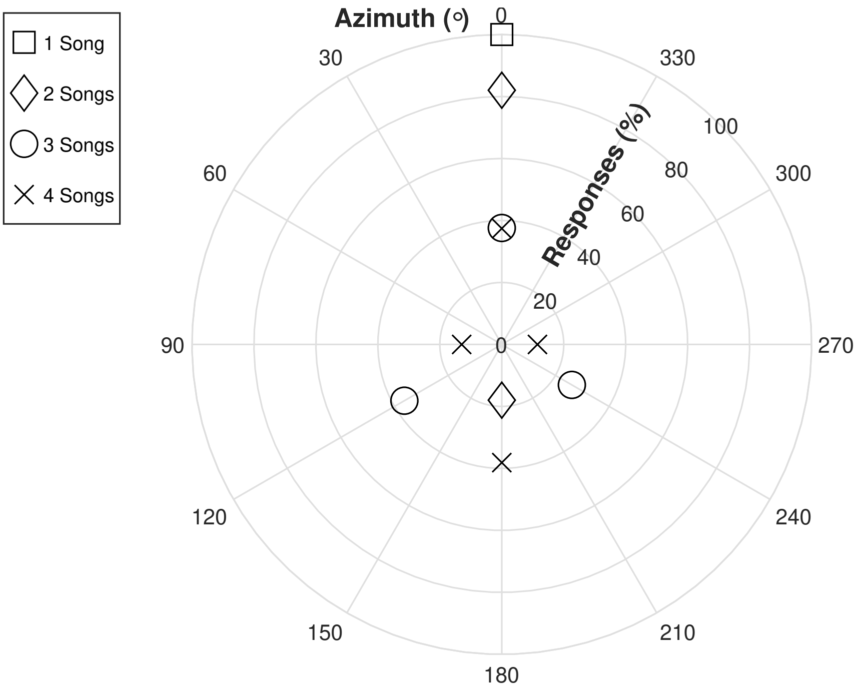


FIGURE 3.18: Distribution of target song locations for all 'in-scene' search responses, by concurrency

but not necessarily positioned successfully (i.e. in-scene total = correct responses + localisation errors). Analysis shows that visual prompts became inconsequential to in-scene success ($\chi^2 = 0.06$; $p = 0.811$) and remained a weak determinant for RT ($F = 0.14$; $p = 0.705$). Figure 3.17 illustrates the distribution of all response types by concurrency, accompanied by average in-scene RTs (which constitute 630 of the 699 data points). Total in-scene responses show a high degree of consistency between one, two and three concurrent songs ($\chi^2 = 1.87$; $p = 0.392$), although a high residual of 22% identification error persists for four simultaneous sources. Variance of in-scene RT by concurrency level ($F = 14.45$; $p < 0.001$) is of a similar magnitude to its influence on wholly correct answers alone.

Figure 3.18 provides a graphical representation of the target song's relative location for all in-scene responses. Only a minority of responses for two concurrent songs had a problem with front and rear binaural orientation. For three sources, front-back confusion resulting in placement of the target in the rear field (closest to 240° or 120°) was actually more frequent than correct song positioning. For four song scenes, reversal of the front and rear locations was as common as correct positioning, whilst some confusion also occurred when the target was at lateral locations (nearest to 90° and 270°).

The strength of individual outcomes from the HRTF selection tournament process is quantified in Chapter 4 as either a *tied win* (where a random winner is chosen between two or more tied HRTF sets), *slight win*, or *clear win*. This categorical classification is indicative of HRTF selection confidence and, therefore, potential suitability or "fit" of the chosen HRTF set to the participant.

Figure 3.19 shows the influence of visual prompt exposure order. No effect is found

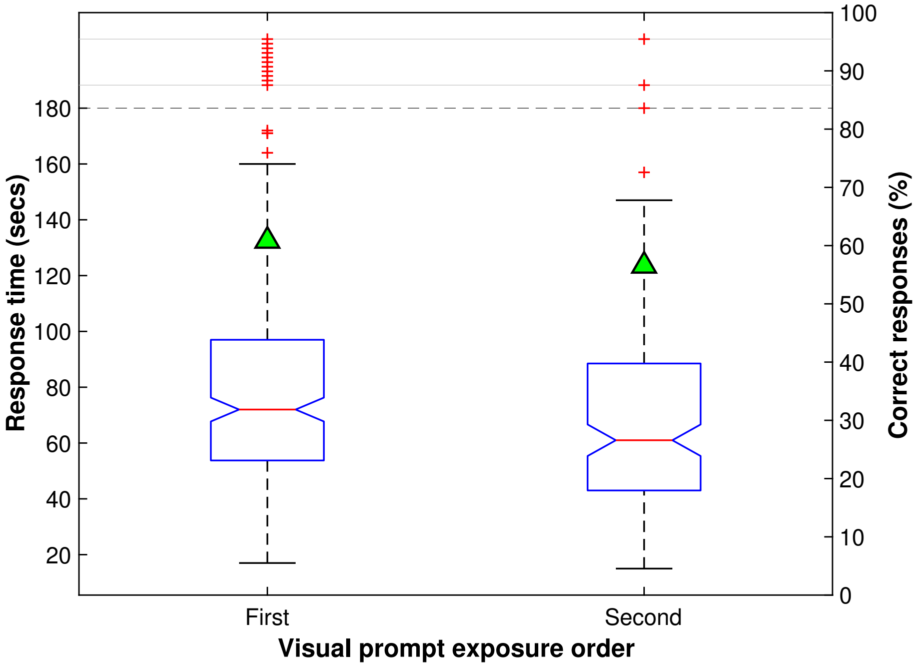


FIGURE 3.19: Response time distributions for all correct search answers, by order of visual prompt exposure. Box plots indicate the 25th-75th percentile range and median value, which are significantly different between groups. ▲ indicates the percentage of correct responses for either order (plotted against the right axis).

in accuracy (χ^2 = 1.28; p = 0.259). However, RT analysis from the three-way main effect ANOVA (alongside concurrency and visual prompt state) does indicate a relationship with group membership. Group B, which was exposed to visual location prompts in the second half of their listening trials, was significantly quicker, in aggregate, in their RT to all search tasks (F = 10.8; p = 0.001).

Browse task analysis

The browsing task objectives in the study design were inherently and necessarily subjective. Response accuracy, as such, is not verifiable and therefore RTs cannot be categorised or analysed according to correctness. Furthermore, Anderson-Darling analysis rejected the null hypothesis for log-normality of the 707 response time data points with high confidence (p < 0.001) and for all but six participants when considered individually. Browse response time distributions is Figure 3.20 further illustrates that the distribution of browse response times does not correspond with log-normality. Subsequent analysis therefore includes all 707 RT data points and relies on non-parametric analysis of participant means by variable to aid interpretation.

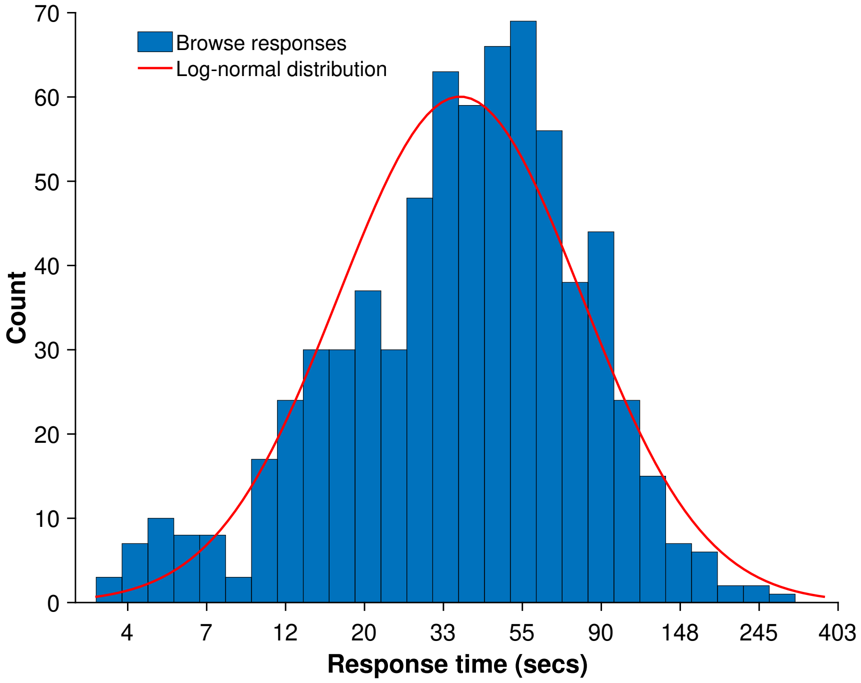


FIGURE 3.20: Distribution of all browse task response times. Response time bins are plotted against a log transformed scale. The line of fit indicates that the data distribution does not adhere to log-normality.

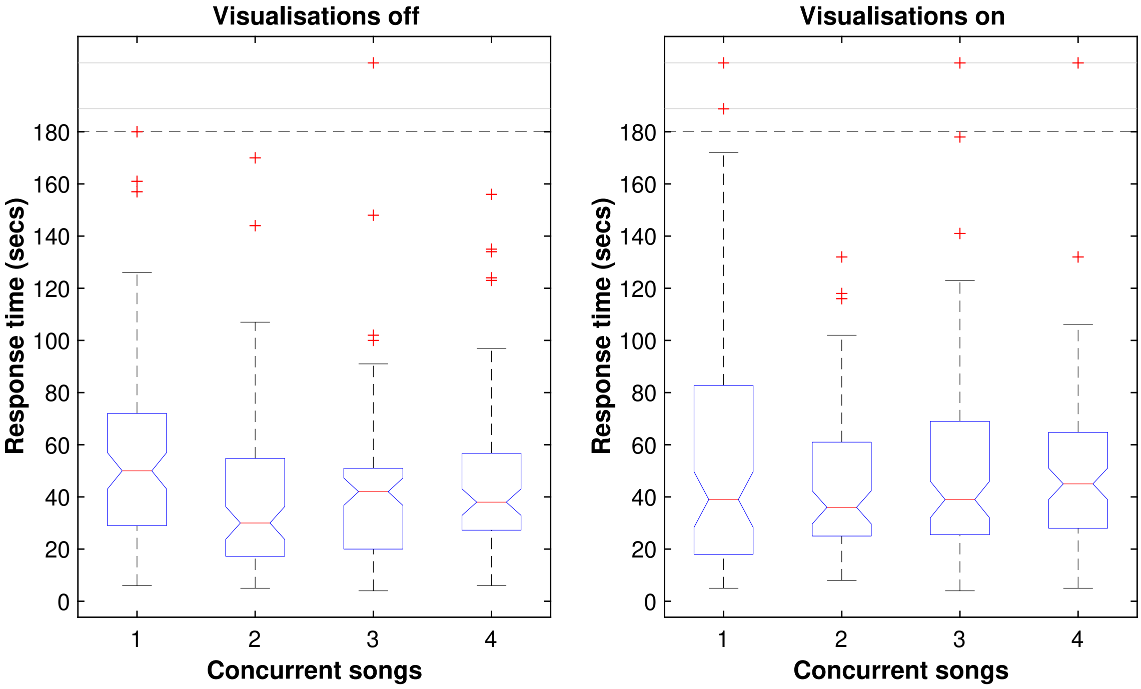


FIGURE 3.21: Response time distributions for all browse tasks, by number of concurrent songs. The separate plots show data for either visualisation state. Box plots indicate the 25th-75th percentile range and median value, which are significantly different only between the one and two concurrent songs groups in the 'Visualisations off' plot.

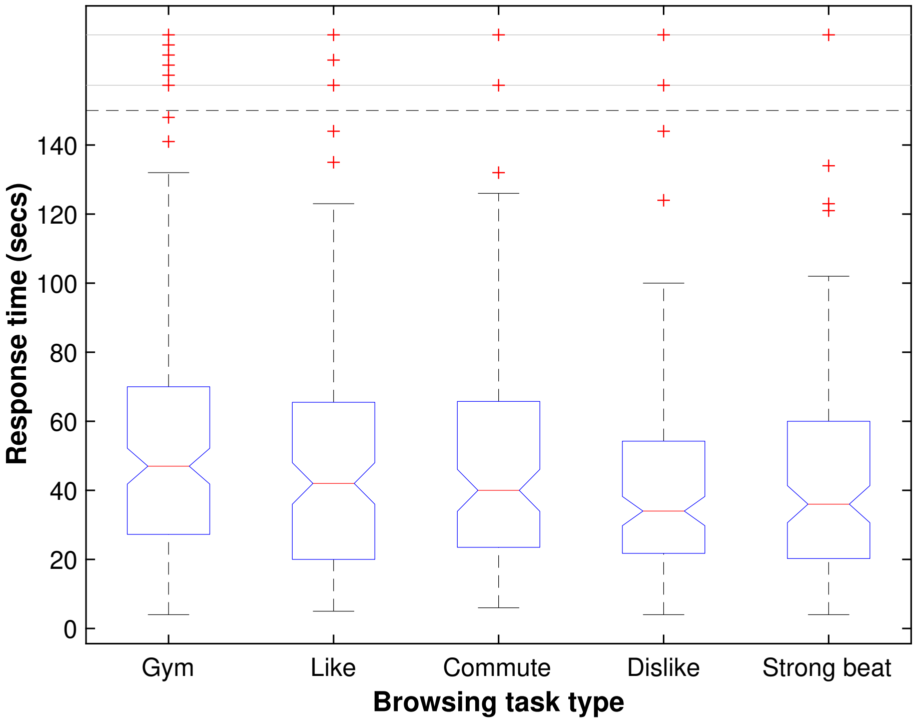


FIGURE 3.22: Response time distributions for all browse tasks, by task type. Box plots indicate the 25th-75th percentile range and median value, which are significantly different only between the "gym" and "strong beat" groups.

Figure 3.21 shows that, without a single correct answer, all configurations of browse task resulted in substantially quicker responses with medians ranging from 30-50 seconds, compared to 55-84 seconds for search tasks (Figure 3.16). Furthermore, smaller variance from the main effect of the four concurrency levels shows lesser influence on RT (Friedman test $\chi^2 = 4.85$; $p = 0.183$). However, closer examination reveals that number of simultaneous sources retains a significant effect in browsing when visual prompts are off ($\chi^2 = 10.23$; $p = 0.017$), but not when on ($\chi^2 = 4.47$; $p = 0.220$). Post-hoc Tukey-Kramer analysis identifies significant difference between one and two concurrent songs in the visual prompts off state ($p = 0.042$). The presence or absence of visual prompts in itself is again found to have no main effect (Mann-Whitney U-test $p = 0.322$).

Some verification of the (lack of) effects found for concurrency and visual location prompts can be achieved through comparison of browse task types. Figure 3.22 shows the distribution of all 707 browse RTs according to these categories. The order of categories, left to right, reflects a decreasing average in time taken to complete each variety of browse task. The sequence can also be viewed to correlate with increasing challenge, where the activity-specific requirement of finding a track the participant would "listen to at the gym" (leftmost) requires a more nuanced judgement than the more open and immediately discernible task of identifying any song with a "strong beat" (rightmost). Significant difference in mean participant RTs between categories is found (Friedman test χ^2 = 10.07; p = 0.039), specifically between the "gym" and "strong beat" variants (Tukey-Kramer p = 0.034). The existence of this distinction – and between two categories for which it might be most expected – supports to some extent the presumption that participants were engaging in earnest with the requirements of browse requests. Had respondents been answering haphazardly or casually, we would expect to see less distinction between browse task type RTs.

Unlike with search tasks, no relationship was found between study group A and B. Mean response speeds between those starting with visual prompts on was found to be similar to those who began without (Wilcoxon rank sum test p = 0.548).

Genre data was again restricted, this time to 317 data points. Little influence was evident between the hip-hop, jazz, rock and mixed style states, where a Friedman test of participant mean RTs within these categories revealed no influence (χ^2 = 1.32; p = 0.724). The difference in average RT between all single- and mixed-genre trials was more notable, but still non-significant (Wilcoxon rank sign test p = 0.131).

3.2.5 Discussion

The three auditory environment variables presented in section 3.2.1 are revisited to discuss results from the experiment.

Visual references to aid sound source localisation

The very low influence of visual prompts over in-scene search task answers confirms expectations; that visual location anchors do not assist in the musical recognition aspects of spatialised concurrent exploration. Much more striking, however, was the lack of any direct effect from graphical representation in aiding user orientation with sound sources. Inclusion of visual sound source positions only had a slight and statistically non-significant benefit on search task localisation accuracy (i.e. correct response rate). Furthermore, the influence was measured to be very small and non-significant when considering speed of either correct search task responses, in-scene search task answers, or browse tasks. All of these trends suggest that a binaural, concurrent music exploration system could be designed and operated as an audio-only experience, as effectively as one with a visual interface. This provides strong support for the principle of eyes-free music exploration design in relation to three potential areas of need:

1. contexts requiring rapid aural review of returned subsets from unfamiliar music catalogues;

2. exploration of audio content via GUI-less wearable computing;

3. scenarios for which or users for whom reliance on use of visual attention is not possible.

The potential effect of exposure sequence to visual prompts is addressed in Chapter 4, section 4.1.2.

Numbers of concurrently displayed sources

Analysis of correct search task responses shows very clearly that use of two concurrent streams was substantially quicker and only 7% and 25% less successful (depending on visual prompt state) than using a single source. More importantly, when considering all in-scene responses, differences in identification error from one to three concurrent sources was negligible and non-significant, whilst average RT for two songs was even quicker. It follows that improving clarity of source positioning could reduce absolute error with two and three sources and potentially have further RT benefits as well. On the other hand, it is also clear that four sources presented a point of sonic saturation in which identification began to suffer substantially. These findings suggest two outcomes of note for further investigation into concurrent music browsing techniques. Firstly, a working assumption can take three sources as the upper limit of perceptual benefit for concurrent, spatialised music presentation. Secondly, improving user localisation of binaurally presented sound sources is necessary to optimise accuracy of two and three song presentation.

Although the visual prompt on state was conceived to potentially reduce localisation confusion in concurrent presentation, as highlighted in section 3.2.5, it did not have a significant main effect on any measure of accuracy or RT across search and browse tasks. Figures 3.12 and 3.16 indicate that this is most likely due to an interaction effect where visual location prompts benefitted some levels of concurrency, but was detrimental for others. Figure 3.21 shows further that the visual prompt on state contributed to longer RTs for browse tasks, also reducing differences between concurrent levels to non-significance. These inconsistencies indicate that visual prompts might have distracted from, rather than improved, participant focus on the individual sound sources, which further advocates for an audio-only approach.

Use of single or mixed-genre sound sources

Neither the genres used in the experiment, nor a mixed- versus single-style playlist were found to hold any bearing over participant searching or browsing performance. Due to an error in the initial study software's logging of the genre variable, only 45% of data points could be analysed, so statistical power was inevitably reduced for this factor. However, the effects of all genre configurations on search accuracy and RT, and differences between the three single-style browse times, were particularly weak. Whilst there was a more noticeable effect of mixed-genre playlists eliciting faster responses than single-genre, this was still not significant and quite likely a result of users more quickly discounting songs from their less preferred styles when browsing the list of 15. So, despite the reduced data set from which to infer answers, there nonetheless seems sufficient information to conclude that genre configuration did not influence participants' performance with the test interface.

Yet, it is known – from earlier research in concurrent earcon and auditory icon system design – that similarity between audio signals has an impact on listeners' ability to disambiguate simultaneous sound streams. It is quite feasible, therefore, that genre categorisation alone is an insufficient means of attributing similarity of signal content between concurrent songs. The makeup of audio signal content in ensemble recordings is determined by a number of musical factors: key, tempo, rhythmic pattern, instrumentation, playing style, to name just a few. Whilst genre is partly determined by these more observable characteristics – which constitute the makeup of the audio signal – style categorisations are ultimately socially attributed and contested. Thus it is quite possible to have songs classed within the same genre, but with substantial contrasts in tonality, tempo, rhythmic idiom, instrumental arrangement or performance character. Future enquiry into the influence of source similarity in concurrent music exploration might need to evaluate a combination of specific musical features, to define and gauge the the influence of sonic uniformity in more detail. The analysis, quantification and representation of such elements in music exploration interface development is a well-established and vast area of research amongst the music information retrieval (MIR) community (Knees, Schedl and Goto, 2019).

3.2.6 Study 2a summary

This experiment establishes a clearer understanding of the potential for using concurrent spatial display of music tracks for content exploration purposes. It revisited themes explored in previous work within the field to understand the current research context for this model of audio interaction. Three important variables were identified to specify a wide but defined focus of investigation, which also informed the study design and approach to data analysis. Two clear outcomes are evident from the analysis of "known-item" search data, on which future research could be supported. First, exploration with two or three binaurally presented concurrent music tracks was seen to be quicker than using a single stream, with listeners' ability to identify the required song not significantly impacted. Second, no evidence was found to suggest that visual representations of sound source positions improved either response accuracy or speed. Analysis of browse task response times suggest that these trends also present for more open-ended use cases. On the basis of these findings, future research on concurrent spatial music exploration can be founded with some confidence on a design principle of audio-only interaction, using up to three simultaneous sources.

A third and more minor insight revealed no patterns of influence between concurrent presentation of songs from different styles of music, or when comparing single- verses mixed-genre presentation. This outcome prompts a need for further focus on perceptual evaluation of simultaneous music streams where the definition of source uniformity/contrast is based on (a combination of) more granular musical similarity metrics. This consideration is addressed, in part, through the design of the final study outlined in Chapter 6.

3.3 Chapter summary

In this chapter, two approaches to applying binaural spatial auditory display in music discovery and recommendation have been assessed to inform an initial view of their potential.

Evaluation of the AAE prototype showed that the voice-led design for exploring a defined library of content supported effective browsing and discovery, in both the standard smart speaker format and the binaural spatial auditory environment. Significant differences in user experience were found for those with English as a first language, and those with higher levels of prior smart assistant exposure. (In both cases these differences in aggregate data are from use of both platforms). Qualitative data also indicated marked contrast between speaker and headphone participants' conceptualisation of their engagement. These patterns, plus possible limiting factors that voice-led interaction placed on fluid navigation of the spatialised auditory scene, indicate that spoken input control is not suited to maximising the benefits of binaural auditory display for music recommendation and discovery. For this reason, the research that follows adopts touch surface input interaction mechanisms rather than speech.

Findings from the experiment on concurrent spatial auditory music browsing indicated that incorporating a visual interface does not significantly aid overall accuracy or speed of selection. Though searching with two concurrent streams was the quickest performing configuration, analysis demonstrated that improved binaural rendering fidelity and inclusion of head-tracking has potential to enable statistically similar levels of selection accuracy for up to three concurrent sources. This issue is the subject of Chapter 5, which concerns methods for effective implementation of real-time binaural rendering on mobile systems. Additional outcomes suggested that improving participants' capacity to focus on the aural scene – either by first exposure without accompanying visual prompts, or through personalised rendering — may further benefit response time. Methods for end user personalisation via HRTF selection are addressed next in Chapter 4.

Finally, it is also worth stating that steps to improve the perceptual quality of a user's engagement would influence more than just the accuracy or speed of their interaction. Presenting a consistent, convincing, immersive auditory music exploration experience also has potential to influence the character of their interactive journey (as suggested by participant mental models of AAE), or even affect browsing decisions made compared to a conventional music library GUI (Barrington, Oda and Lanckriet, 2009). Decision-making and quality of experience therefore form part of the eyes-free spatialised music exploration final prototype evaluation in Chapter 6.

Chapter 4

HRTF selection for interactive auditory display of music content

> I close my eyes and here it comes again.
> I can hear music,
> I can hear music.
> The sound of the city, baby,
> Just don't disappear.
> I can hear music,
> Sweet, sweet music.
>
> 'I Can Hear Music'
> *The Ronnettes*

Personalised binaural rendering – whereby the perceptual cues determined by an individual listener's morphology are synthesised as closely as possible – remains an ongoing challenge. As discussed in section 2.3.3, although a variety of approaches to HRTF customisation do exist, these are contingent on a number of factors, including user expertise, duration, specific hardware or available computational resource. At the time of writing there was no established approach that could be considered suitable for end-users of a mobile computing device. However, it is clear that an audio augmented reality system deploying binaural synthesis for music discovery would benefit from a proven method for users to select the non-individualised HRTF set that works best for them personally.

The following considerations for judging the efficacy of any selection system are defined here for use throughout this chapter:

- *Reliability* – Does the system return an identifiable preference in a significant majority of cases? Given that this thesis concerns a non-critical leisure activity pursued by an end user, a 90-95% success rate will be regarded as desirable.

- *Validity* – Does the returned HRTF provide sufficient spatial rendering fidelity for the user? For the music exploration, audio-only use case that concerns this thesis, rendering fidelity is defined as the capacity to discriminate between both the content and virtual origins of different sound sources within a scene.

- *Usability* – Can the system be operated equally successfully by any user, irrespective of listening expertise?
- *Efficiency* – Is the overall time taken to complete the selection process of an acceptable duration? For the purposes of single-time calibration of a recreation-focussed system, no more than ten minutes will be considered desirable and under five minutes would be ideal or preferable.

4.1 Study 2b: User HRTF selection for 2D interactive audio

4.1.1 Experimental design

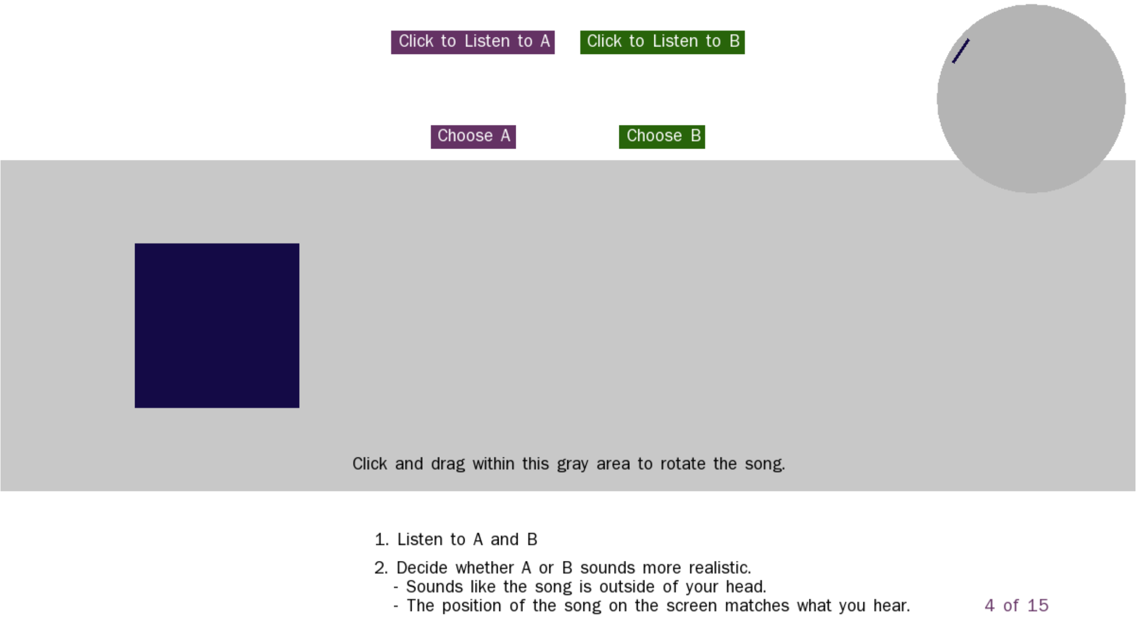
This experiment was conducted in conjunction with that outlined in section 3.2 (i.e. Study 2a), by Dr Rebecca Stewart and prior to the start of the doctoral research presented in this thesis. The test interface used to gather responses is based on a similar design to Study 2a, but presents the listener with pairs of HRTF sets and asks them to select a preference in each case. Rather than either absolute or relative parametric judgements (as discussed earlier in section 2.3.3), the method uses an interactive, holistic evaluation to determine, for each pair, which provides the most realistic effect for the listener. The outcomes of each selection round are then used to sort the collection into a final ranked order. Although tournament-style selection has been used previously by Iwaya (2006), this approach differs in that it exhaustively iterates over every possible pairwise HRTF set combination and uses recorded music tracks as stimuli within an interactive system.

Task description

The experiment evaluates a proposed design for end-user selection of a preferred, non-individualised HRTF set for binaural synthesis from a collection of HRTF sets. Participants compare pairs of binaurally synthesised spatial renderings of a single song presented over headphones. For each pair, either render is of the same song but convolved with one of two HRTF sets implementing horizontal-only virtual FOA (Noisternig et al., 2003). The participant can choose either HRTF at any time whilst they are listening and rotate the song's virtual position around their head using an interactive interface. Head-tracking is not used during any part of the experiment.

HRTF selection tournament structure

Six HRTF sets were identified for use in the study by Dr Stewart at the time of devising the experiment in 2011. Three human-measured HRTF sets (with the identification codes *1014*, *1022* and *1028*) were used from the publicly available *LISTEN* database (Warusfel, 2003). Three human-measured sets (with the identification codes *12*, *15* and *58*) were used from the publicly available *CIPIC* database (Algazi et al., 2001). These sets are the



Click and drag within this gray area to rotate the song.

1. Listen to A and B
2. Decide whether A or B sounds more realistic.
   - Sounds like the song is outside of your head.
   - The position of the song on the screen matches what you hear.

4 of 15

FIGURE 4.1: Graphical user interface for the 2D HRTF selection tournament

three from each collection that were found to perform best under horizontal plane evaluation by Roginska, Santoro and Wakefield (2010). Findings from that research identified these specific sets as those most clearly favoured for their sense of externalisation and front/back discrimination at 0° elevation. They authors noted, in particular, that although the subset from the *LISTEN* database seemed less preferred in aggregate than *CIPIC*, there was a minority of participants who had a clear preference for those HRTF sets. Options from both *LISTEN* and *CIPIC* were therefore included in the 2D HRTF selection study to cater for possible individual preferences towards different HRTF measurement databases (Roginska, Santoro and Wakefield, 2010).

Use of six HRTF sets results in 15 pairwise comparisons. Pairs are determined by a round robin tournament structure, where every one of the six available HRTF sets meets each of its five opponents once using randomly generated scheduling. A round robin tournament makes it possible to determine, definitively, which HRTF set(s) within the group of six are preferred by a participant. It is purposefully distinct from the approach used by Iwaya (2006), as it establishes a ground truth and data set that can be used to model and evaluate less comprehensive tournament formats.

For each tournament round, the participant is asked to choose the HRTF set that provides the better spatial effect. To do this, they are instructed to think holistically about the realism of the spatialisation, giving specific attention to sense of externalisation and accuracy of localisation. Figure 4.1 shows the graphical user interface (GUI) and text prompt used to elicit responses.

The interface is comparable to that described in section 3.2.1 (to collect responses for Study 2a) and also uses an identical binaural rendering implementation – i.e. anechoic, virtual FOA without head-tracking. However, for Study 2b, only one song is presented at a time and visual prompts are always active. The participant interacts with the grey area in the GUI in the same way, using a mouse to rotate the spatial image clockwise (drag right) or anticlockwise (drag left). When the sound source is in the front field (i.e. within +/- 90° azimuth) a blue square is displayed within the grey area relative to its orbital location. The circle in the upper right corner of the window illustrates the active song's spatial position from an overhead perspective, which updates in real-time with user interaction. When shifted beyond +/- 180° azimuth, the active song transitions out of the auditory scene and one of two other songs is rotated into the environment in the concurrent direction. The aim in including this type of interactive mechanism within the comparison process was to simulate features of potential multimedia applications, where user control and spatial scene rotation might (ideally) be integral to the listening and sound localisation experience.

The same three songs are presented for each of the 15 pairings. Each song is of the same musical genre (latin pop, as determined by the process outlined in Appendix C), is edited to fade at one minute and plays back on a repeated loop. As stated in Table 3.2, before starting the experiment a video is shown to the participant, which demonstrates how to interact with the GUI and fully explain the task they need to complete. This includes the instruction:

> *Listen to A and B. Rotate the song and decide if A or B has a better 3D audio effect. Consider whether the song sounds outside your head, if it sounds like it really goes behind and in front of you, and if where you hear the song matches where you see the song on the screen.*

Participants are also informed that this is a calibration process and that there is no right or wrong answer, only their own personal preference.

As shown in Figure 4.1, the navigation and selection software enables A/B comparisons by allowing a seamless switch between HRTF sets without changing the song's virtual position in space or playback point. After 15 rounds the HRTF set chosen most frequently as the winner of a tournament round (up to a maximum of five times) is identified as the participant's preferred set. If there is no clear winner, one HRTF set is randomly selected from the tied top results.

Auditory navigation trials

After choosing their winning HRTF set, the participant is presented with the 64 auditory navigation trials that are described and analysed in section 3.2 (i.e. that which formed the basis of Study 2a). At this stage it is important to highlight that a poorly matched HRTF would be expected to make the majority of auditory searching and browsing tasks described in section 3.2 more challenging for participants. With two, three and four concurrent songs, less accurate binaural rendering would introduce greater front/back confusion and reduce clarity in sound source localisation. This would be compounded in the 50% of tasks where a visual representation of sound source locations is not displayed. Such conditions could be anticipated to either increase selection error (reduce success), or make disorientation more likely (lengthen response time).

Participants

The same 22 volunteers participated in the experiment as described in section 3.2.3 – i.e. stages 2 and 3 of session one detailed in Table 3.5. As part of the consent form, each subject filled in a personal profile questionnaire. This form requested information about their prior musical experience and music listening habits, which is reported in section 4.1.2.

4.1.2 Results

The following analysis is divided into four parts. First, results from this experiment are examined in isolation to identify how often an HRTF set was chosen in individual tournament rounds, when it was ranked as an overall winner and the strength of each outcome. Second, the HRTF selection strength is examined against participants' performance in the search task outcomes from Study 2a. Third, an overview of participant listening habits and musical experience is presented in relation to HRTF selection strength. Fourth, a comparison of the round robin results is compared with the projected outcome of a knockout tournament, to see how far results would differ under a less comprehensive selection process.

HRTF selections

Figure 4.2 shows that when the individual tournament match outcomes from each participant are viewed collectively, one HRTF set performs clearly below chance level (*CIPIC 58*, with 22.7%) and another notably above (*CIPIC 15*, with 63.6%). One other set performs marginally under chance level (*LISTEN 1014*) and the rest slightly over. A Friedman test confirms there is a significant difference in the mean overall popularity of *CIPIC 58* compared to *LISTEN 1022* ($p = 0.011$), *LISTEN 1028* ($p = 0.007$), *CIPIC 12* ($p = 0.032$) and *CIPIC 15* ($p = 0.002$). *LISTEN 1014*, on the other hand, is not significantly different in its mean ranking to either *CIPIC 58* or any of the top four sets.

To analyse the degree of certainty or strength of the winning HRTF set two indices were used: winning HRTF set score and winning HRTF set margin. The winning score is the number of tournament round wins attained by the selected HRTF set, which has a value of either 3/5, 4/5 or 5/5. The winning margin is the difference in tournament round wins between the selected HRTF and the second place set, which has a value of

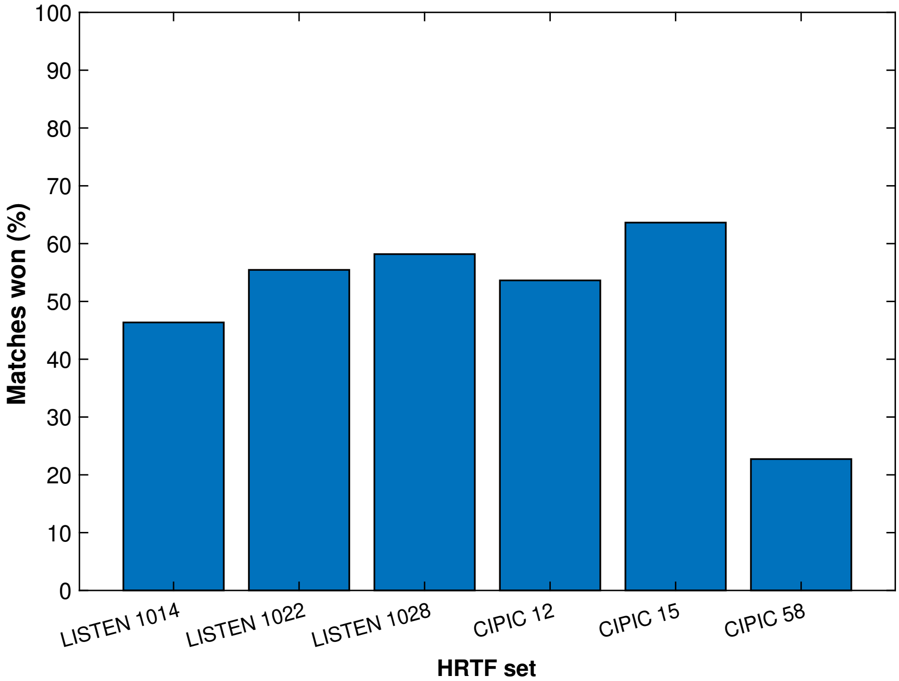


FIGURE 4.2: Percentage of matches won by each HRTF set in the 2D selection tournament

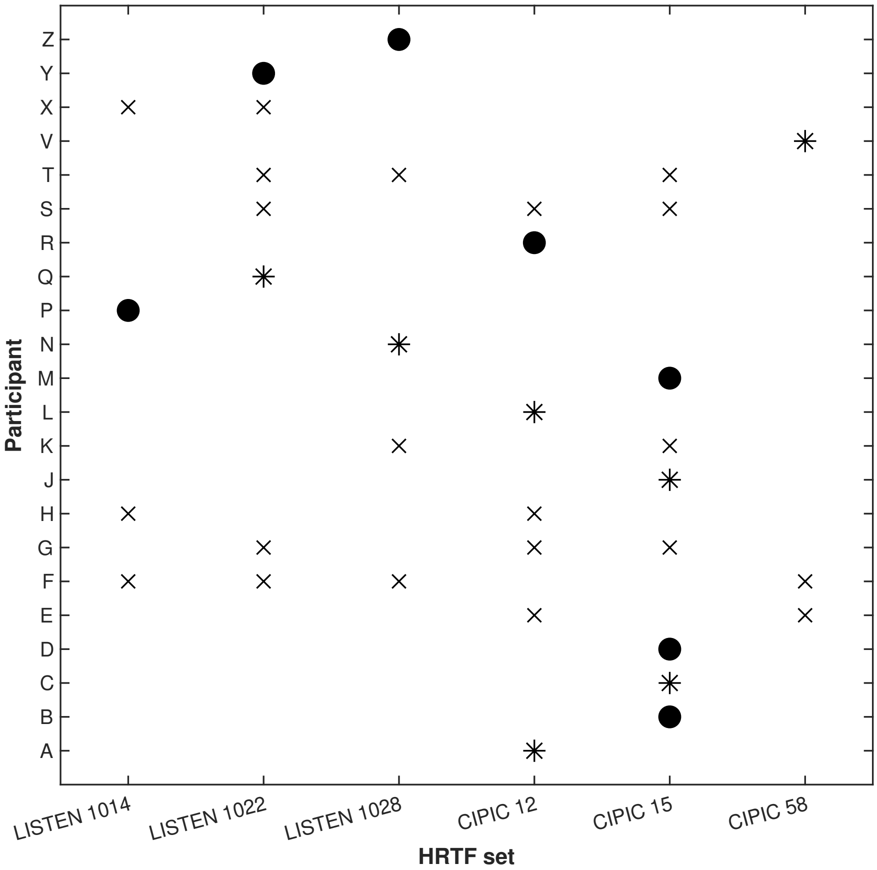


FIGURE 4.3: Breakdown of winning HRTF sets in the 2D selection tournament (x = tie, $*$ = winner by margin of 1, • = winner by margin of 2).

	Margin			***Totals***
	0	**1**	**2**	
	(tied win)	***(slight win)***	***(clear win)***	
Score of 3/5	1	0	0	***1***
Score of 4/5	7	2	0	***9***
Score of 5/5	0	5	7	***12***
Totals	***8***	***7***	***7***	

TABLE 4.1: Distribution of HRTF selection results in the 2D selection tournament

either 0 (in the event of a tie), 1 or 2. The distribution of both measures is shown in Table 4.1. For all but one participant, the competitive selection process resulted in a winning score of either 5/5 or 4/5.

The aggregate performance of each HRTF set (found in Figure 4.2) is mirrored in the outcomes for individual participants (in Figure 4.3). For instance, *CIPIC 15* was the most frequently chosen set in all individual matches and the most commonly chosen as an outright or joint overall winner (five and four times, respectively). Both of the least selected HRTF sets overall were still outright winners in one case (*CIPIC 58*, for participant V and *LISTEN 1014*, for participant P). Furthermore, in each of these instances, the HRTF set was an undefeated winner, scoring 5/5 in the selection process.

The instructional video told participants that the HRTF selection process would take approximately 10 minutes and in practice the average time taken to complete it was a mean of 11.4 and median of 9.5 minutes.

HRTF selection strength and user performance

Chi-square tests between the number of correct search task responses per participant and the three winning score groups of the HRTF competition process (3/5, 4/5 or 5/5) show no significant effect (χ^2 = 0.028; p = 0.986). ANOVA tests of correct search task response times (RT) do show significant difference between the same groups (F = 6.9; p = 0.001). However, when comparing winning score to the winning margin as indicators of selection strength, winning score is a less useful metric for two reasons reflected in Table 4.1. First, membership of the categories is particularly imbalanced at 1, 9 and 12 participants in each group. Second, the 4/5 category contains a number of tied results whereby the HRTF set used was subsequently selected randomly amongst the top tied winners. For these reasons, the winning margin is instead identified as the preferred index for selection strength and is the measure referred to in the rest of this section using *tied win*, *slight win* and *clear win* as the grouping categories.

Chi-square tests of correct search task responses per participant between the three winning margin groups of the HRTF tournament selection also show no significant effect

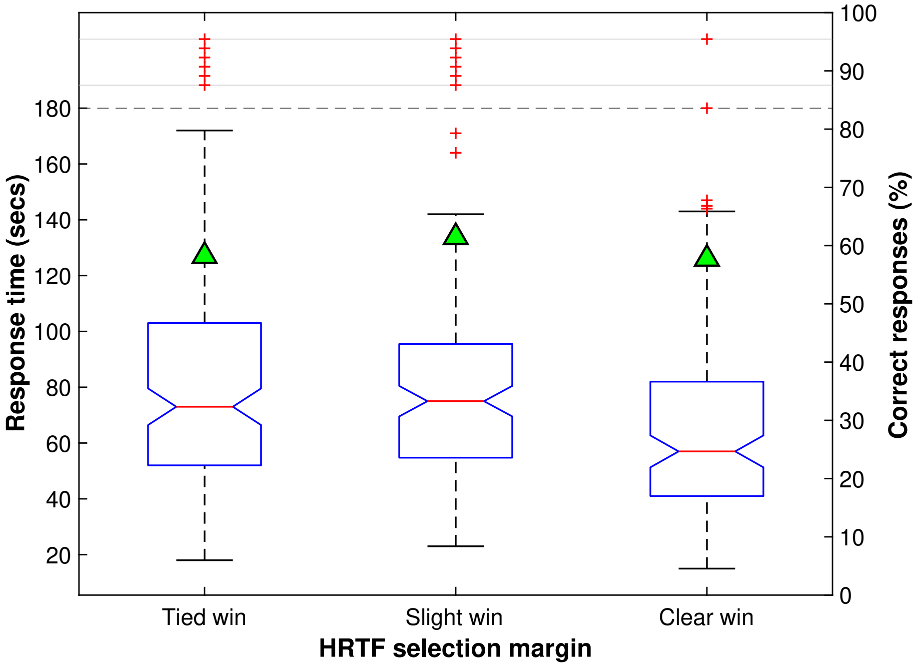


FIGURE 4.4: Response time distributions for all correct search answers, by 2D HRTF selection margin. Box plots indicate the 25th-75th percentile range and median value, which are significantly different for the *clear win* group compared to *tied win* and *slight win*. ▲ indicates the percentage of correct responses for each margin (plotted against the right axis).

($\chi^2 = 0.688$; $p = 0.709$), which is reflected in Figure 4.4. However, a significant difference in RT is evident amongst participants who selected their HRTF with a *clear win*. This group was quicker, on average, by 15 seconds or more than either the tied win or slight win groups ($F = 10.15$; $p < 0.001$). This trend is in line with expectations – i.e. that participants with the strongest HRTF selection outcome tended to respond quicker. In effect, the group that experienced a more clearly preferable personalisation of their search tasks may consequently have benefitted from improved disambiguation of sources and their apparent locations.

However, the data displayed previously in Figure 3.19 and here in Table 4.2 show that this apparent correlation could have a secondary association. Table 4.2 highlights the disproportionate occurrence of participants who chose their HRTF set as a clear winner in visual prompt exposure order Group B (4/9), compared to Group A (3/13). Further, the five participants identified as having significant differences in RT deepen the three-way association between speed, visual prompt order and HRTF selection strength. The *clear win* HRTF selection strength category contains the three participants with faster correct RTs, whilst *tied win* and *slight win* feature those two with slower correct RTs. When HRTF selection strength is substituted into the three-way ANOVA alongside concurrency and visualisation state (as outlined in section 3.2.4), it is seen to have a comparably significant effect on RT ($F = 10.59$; $p < 0.001$) to visual prompt exposure order ($F = 10.8$; $p = 0.001$).

	HRTF Selection Strength		
	Tied win	*Slight win*	*Clear win*
Group A (vis. first)	3 *inc. S (slower)*	7 *inc. C (slower)*	3 *inc. Y (faster)*
Group B (vis. second)	5	0	4 *inc. D (faster)* *inc. M (faster)*

TABLE 4.2: Distribution of participants by visual location prompt order and HRTF selection strength categories. (Five participants with significant differences in their average correct response times are identified).

HRTF selection strength and user expertise

The makeup of the selection strength groups was also cross-referenced against personal listening habits and musical experience. The breakdown in Figures 4.5 and 4.6 show fairly even representation of all winning margin groups across levels of listening and musical training. Chi-square tests confirm no significant difference in the makeup of winning margin groups against either listening habits (χ^2 = 6.40; p = 0.380) or musical expertise (χ^2 = 6.72; p = 0.567).

Alternate tournament format

A round robin tournament is an exhaustive competitive tournament structure, as all possible combinations are evaluated. This is in comparison to knockout tournaments where only consecutive winners are matched against each other. The latter format results in fewer comparisons and therefore a shorter overall tournament. A knockout tournament was simulated to project whether this shorter format would result in different winning HRTF sets than the results of the round robin format.

Each participant's tournament was re-run using the same sequence of HRTF pairs and results as in the live experiment, but any losing HRTF set was eliminated from future rounds. Subsequent comparisons that involved an eliminated HRTF set were then ignored. This tournament model reduced the number of comparisons per participant from 15 to just five – with each round removing one of the six HRTFs from the competition.

This re-projected format resulted in a different final selection for just six of the 22 participants. Furthermore, for five of these six, the knockout winner had also been a joint winner in the round robin (i.e. had won the same number of comparisons) and had not been designated as the chosen HRTF merely due to random selection between top tied results. Thus only one participant would have actually been allocated a weaker choice under this identification method. In that instance, the participant would have been

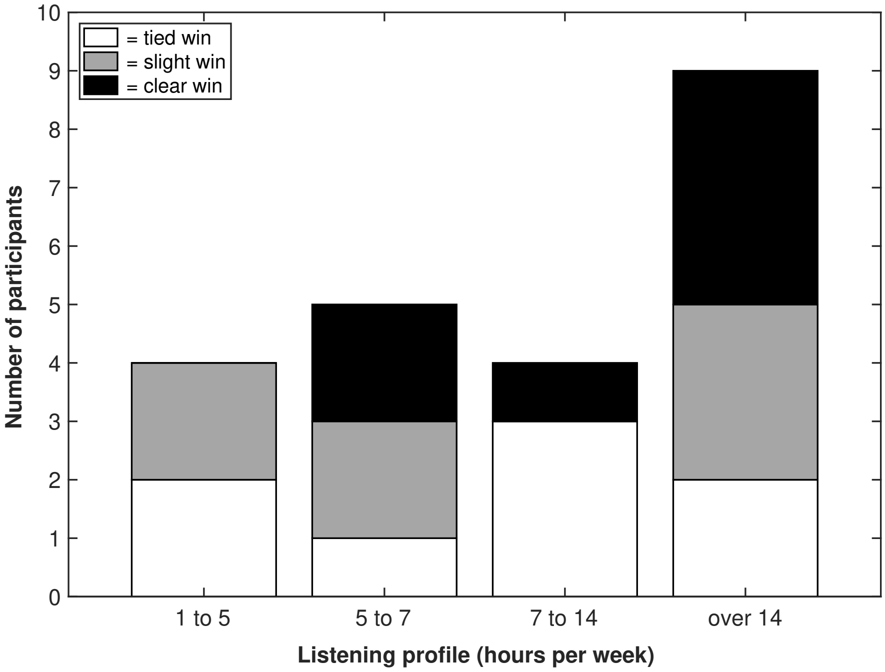


FIGURE 4.5: Makeup of 2D HRTF selection winning margin groups, by participant listening profile

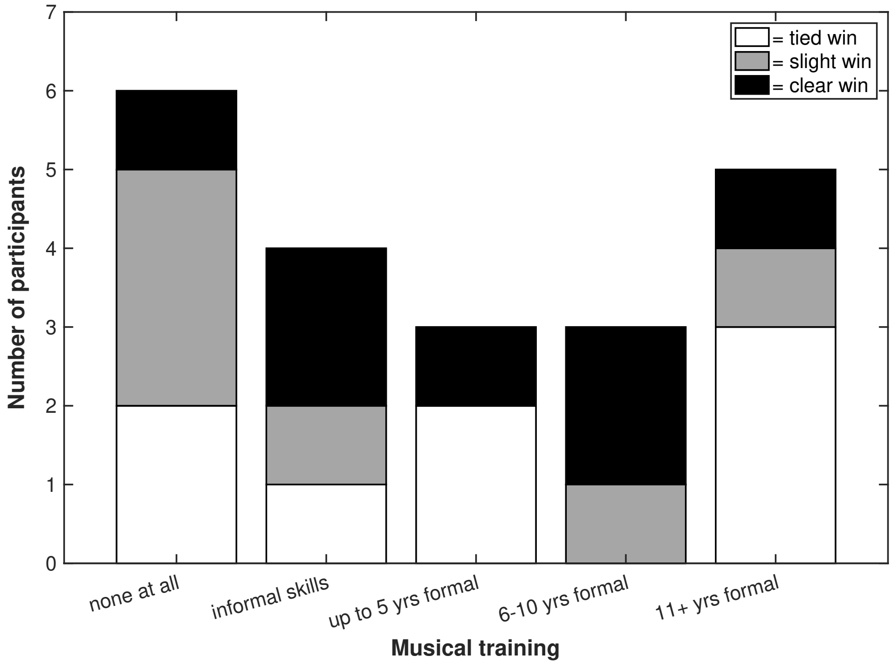


FIGURE 4.6: Makeup of 2D HRTF selection winning margin groups, by participant musical training. (Note that one participant did not respond.)

allocated an HRTF that they had deemed favourable in only 2/5 comparisons, rather than the winning score of 4/5 that had resulted from their completed round robin tournament.

4.1.3 Discussion

The four criteria put forward at the beginning of this chapter are revisited to evaluate the success of the HRTF selection system.

Reliability

Participants frequently showed demonstrable preference for an HRTF set through the comparative selection process, with 95% of tournaments resulting in a score of 4/5 (either as an individual or joint winner) or 5/5. Moreover, the distribution of competition results for HRTFs overall and by individual participants confirms that less preferred sets, in the context of a larger population, can nevertheless be well-matched for specific individuals. This pattern is consistent with previous research (Roginska, Santoro and Wakefield, 2010) and suggests that the requirements of listeners whose best fitting HRTF set is less commonly chosen in aggregate can still be successfully matched under this type of selection system.

This data attests to the effectiveness of the pairwise comparison mechanism for consistently identifying personally preferred HRTF sets. Had the system been less effective, a higher proportion of unclear outcomes would be evident. In fact, there is only one instance of either a joint winning score of 3, or a four-way tie (participant F, in both cases). Furthermore, there were only three further cases of a three-way tie (participants G, S and T).

Validity

The findings presented here show that the relative strength of the HRTF match had no significant influence on user success in the subsequent auditory retrieval trials. This is the case if participant selections are analysed either by winning score, or winning margin. This data provides some verification as to the validity of the proposed HRTF selection process for task-oriented applications, such as in the one devised for this research. I.e. all strengths of selection returned by the system (*tied win*, *slight win* or *clear win*) provided a fit that produced similar levels of accuracy.

However, further data uncovered in combination with the analysis of Study 2a shows that either the user's strength of HRTF set selection, or the sequence of their exposure to visual location prompts, was associated with significantly improved correct search task RTs. Overlap between these two factors by participant made it difficult to isolate their relative impacts. Nevertheless, there is a commonality of principle between better HRTF personalisation and initiation to the system without visual prompts; both serve to promote improved user focus on the auditory scene. It should be noted that visual prompt exposure order was a wholly independent variable in the study design, whereas HRTF set selection strength was derived as an indicator to enable some comparison between certainty of "fit". More detailed and dedicated verification of the suitability of selected HRTF set is therefore required to evaluate the validity of this general approach to binaural personalisation.

This study also demonstrates a holistic, interactive approach to subjective selection of non-individualised HRTFs. In doing so, it contributes to wider discussions on how quality of binaural rendering systems can be assessed using means that extend beyond measurements of localisation accuracy alone (McKenzie, Murphy and Kearney, 2018; Nicol et al., 2014; Reardon et al., 2017; 2018; Simon, Zacharov and Katz, 2016).

Usability

There is no evidence from this participant group to suggest that either personal listening habits or musical experience influenced their ability to make a more or less decisive HRTF selection via the method tested.

Efficiency

Preliminary investigation indicates that the round robin tournament format could be shortened and still achieve very similar outcomes with potentially only small impact on reliability. Reducing the process by a third (and in theory from around 10-11 minutes to 3-4 minutes in duration) only altered one outcome substantively.

4.1.4 Study 2b summary

Together, Study 2a and b measured the success of a mechanism for selection of a non-individualised HRTF set based on holistic comparative judgements, by users of an interactive binaural system without head-tracking. This approach has been shown to result in consistent identification of optimal HRTFs — whether or not a singular preference is ultimately identified by the process. The outcomes of the selection method have been tested with a task-oriented simulation. Results show no significant effect between the different strengths of HRTF preference that resulted and task competence. Moreover, neither personal listening patterns nor musical training appear to influence the strength of HRTF choice that presents from using this selection mechanism, demonstrating its potential applicability to both novice and expert users.

A clear, but three-way association between search task response time and exposure to HRTF selection "fit" / initial use without visual prompts was also uncovered. The overlap that resulted between these categorical variables meant that causation is unclear in this instance. However, the trend points to a likelihood that improving participants' focus on the binaural scene – whether this is through personalisation of the rendering, or an emphasis on auditory training – will support sound source disambiguation and localisation, with concomitant impact on the speed of interface use.

Overall, these findings support the case for an interactive and iterative calibration process that allows users to choose an optimal non-individualised HRTF set. Projection of simpler tournament structures further shows that more efficient selection procedures might be used in future systems to gain selections with a similar level of certainty, but more rapidly and thus to potentially incorporate greater choice of HRTF sets. However, the holistic judgement drawn out from participants in this design solution only caters for 2D binaural display. Further verification of the validity of HRTF selections that result from interactive comparative judgement would also be beneficial. Section 4.2 now explores how holistic evaluation of a 3D binaural effect could be elicited, both in terms of the design of stimulus material and the nature of the pairwise comparison asked in a mobile computing use case.

4.2 Study 3a: User HRTF selection for 3D interactive audio

Part a of Study 3 develops and evaluates a method for user selection of preferred HRTF sets via a custom binaural rendering system for mobile, interactive spatial audio. It presents a further response to the limitations of prior work on HRTF selection for end-user audio-only contexts, as discussed in section 2.3.3. It extends the approach taken in Study 2b, taking into account the findings from that analysis. In addition, the method conceived for 3D HRTF comparison addresses the further limitation of mobile device constraints. The proposed method is assessed against the same set of criteria established at the start of this chapter, i.e. that the solution must demonstrate:

- *Reliability* – clear and consistent selection outcomes
- *Validity* – spatial fidelity that is fit for purpose
- *Usability* – potential benefit to any end user
- *Efficiency* – a duration acceptable for the use case

4.2.1 Considerations and approach for 3D audio-only mobile contexts

Adapting the 2D holistic approach described in section 4.1 for 3D judgement of HRTFs in mobile use cases presents three challenges:

1. Thorough comparative judgement of 3D involves a much wider range of spatial positions and trajectories, all of which must be adequately explored to make valid selections.

2. Mobile and 'hearable' devices have either restricted or no visual display, meaning that judgements should be made without reference to dynamic graphics or complex interfaces that would be necessary for audiovisual interaction.

3. A clear indication of selection repeatability (i.e. *reliability* is necessary in light of the above complexities.

The prototype used to conduct this research deploys the virtual VBAP (Pulkki, 1997) rendering system developed on an open-source embedded Linux platform detailed and evaluated in section 5.1. It provided the head-tracked, audio-only, mobile approach to binaural synthesis required for this experiment. In summary, the approach uses eight HRIR pairs to simulate a sparse, pseudo-spherical virtual loudspeaker array. Individual sound sources are positioned via VBAP signal processing prior to binaural encoding. Five virtual speakers are located on the horizontal plane (0° elevation) at 0°, -60°, 60°, 120° and -120°. Three further speakers are placed around the median plane (0° azimuth) at 90° (directly above), -45° (below front) and -135° (below back).

Participants wore a battery powered embedded device running the VBAP binaural synthesis renderer, study journey and response logging software. They used a networked smartphone to submit responses and progress through both part a and part b of the the study. The role of the smartphone in both instances was to record participant responses in the testing system via a touch screen, with a bespoke GUI – it did not have any involvement or influence on the binaural rendering. Sennheiser HD650[1] headphones provided audio playback without any equalisation for binaural synthesis applied. Head-tracking was not enabled for part a of the study (i.e. for the HRTF selection process). A link to the source code repository is provided in Appendix A.2.

4.2.2 Participation

As stated, the HRTF selection method was devised and evaluated against the four criteria in section 4.1. Participants undertook three identical study sessions (to help assess *reliability*) in an isolated and acoustically dampened environment. In each of the three separate sessions, part a comprised of the HRTF selection procedure (to test *usability* and *efficiency*) and part b consisted of a follow-up objective localisation task (to measure *validity*), which is detailed in section 5.2.

Twenty-one participants (aged 25-45, 6 female and 15 male) were recruited on to the study, which was approved by the Queen Mary University of London (QMUL) Ethics Committee (reference 2038). Each session was approximately 45 minutes, with a minimum of 48 hours between sittings. Participants received £20 in compensation for their time at the end of their third session. All participants were recruited via an open email call to staff and doctoral students across QMUL's School of Electronic Engineering and

[1] www.sennheiser-hearing.com/en-UK/p/hd-650/ [accessed 12/6/2022].

Computer Science. No hearing impairments were declared, other than from two participants who reported occasional and slight tinnitus that neither regarded as pronounced. A questionnaire was used to collect information on musical training, headphone listening habits and prior exposure to binaural audio.

4.2.3 Experimental design

For the 3D selection study, a comprehensive tournament structure used 21 pairwise comparisons between an optimised shortlist of seven human-measured HRTF sets from the *LISTEN* database (Warusfel, 2003) identified by Katz and Parseihian (2012). The makeup of this shortlist differed from the 2D selection study in number (seven, rather than six), database source (exclusively *LISTEN*) and specific sets used (only set *1022* featured in both the 2D and 3D selection study). However, the work by Katz and Parseihian (2012) emerged after the 2D HRTF selection data was collected. They systematically curated the pool of seven *LISTEN* HRTF sets from the complete database of 46, specifically for the purpose of supporting practical approaches to subjective user selection. The subset is considered to offer an optimal balance of the the most convincing binaural simulation for a range of 45 test participants, using the fewest possible options. It has been used by Kim, Lim and Picinali (2020) in another prominent study investigating selection repeatability.

Another notable investigation into HRTF selection repeatability used stimuli with fixed trajectories that were not responsive to head-tracking (Schönstein and Katz, 2012). The test pursued here followed a similar approach, but used content derived from recorded music, rather than test tone signals. The comparison stimulus was compiled from excerpts of an anechoic recording of Haydn's Trumpet Concerto in Eb performed on unaccompanied cornet (Hansen and Munch, 1991).

Comparison trajectory

The trajectory in Figure 4.7 addresses considerations 1 and 2 specified in section 4.2.1. The horizontal plane orbit consisted of a single sustained note lasting approximately two seconds. Five short bursts in four static positions on the median plane used an even four-note phrase of around one second. The overall stimulus was a little under ten seconds. A key feature of the trajectory was that it passed through all eight virtual speaker locations, which takes advantage of VBAP's amplitude panning approach to spatialisation. Although the trajectory covered just 363 of 64,082 potential coordinates, this small minority focussed on the eight fundamental points from which all locations are rendered (consideration 1). The trajectory was also refined and judged to be sufficiently short and simple enough to enable purely internalised A/B auditory comparison, without reference to dynamic or interactive graphics (consideration 2).

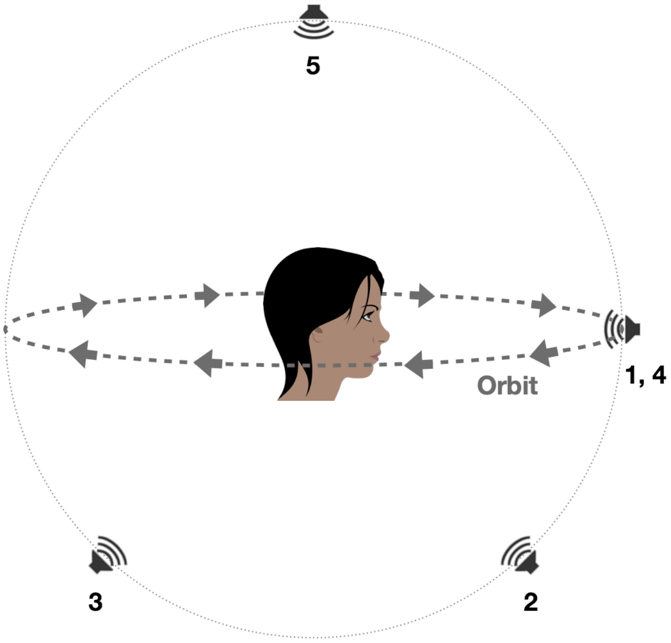


FIGURE 4.7: Trajectory for the virtual VBAP 3D HRTF comparison, including horizontal plane orbit and five short bursts at four median plane locations

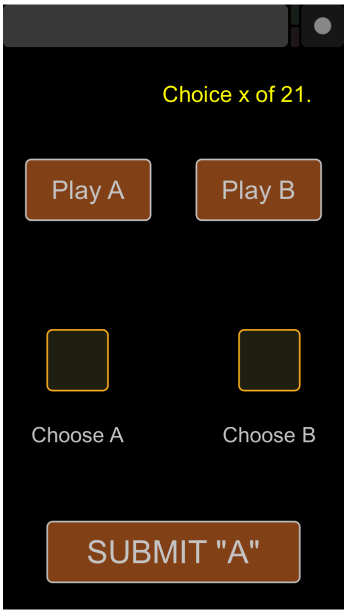


FIGURE 4.8: Smartphone user interface for 3D HRTF preference submission

Selection process

Participants used the GUI shown in Figure 4.8 to compare trajectories and submit preferences. They were also given the diagram in Figure 4.7 and an accompanying instruction:

> ***Which has the more convincing 3D effect, excerpt A or B?***
>
> *When comparing A and B, you may wish to consider:*
>
> - ***horizontal accuracy** during the orbit and at the four static central positions*
> - *sense of **spread** between front/back and up/down positions*
> - *sense of **distance** or how "outside of your head" the sounds seem*

Participants completed one example response to check their understanding of the task before starting. For each response, both the time elapsed and outcome of each comparison was logged automatically. Participants were allowed to listen to either trajectory as many times as they wished, but were forced to listen to A and B at least once in their entirety, before response buttons became enabled.

Both the sequence of comparisons and the order of A/B pairs were randomised for each participant, at all three sittings. For each of their three sessions, the HRTF sets chosen and rejected most often were designated as the preferred and least favoured options. In the event of a draw one of the tied sets was picked at random. These two designations were then used as the best and worst fitting HRTF sets in participants' subsequent localisation test for part b (detailed in section 5.2).

4.2.4 Results

Between the 21 participants, 63 HRTF selection procedures were completed. For each session, the outcomes of a participant's comparisons were translated into rank order based on the number of times each HRTF set was selected (a maximum of six and minimum of zero occasions). Tied HRTF sets were given a shared ranking at the highest jointly occupied position. So, for example, a ranking list of 1,2,3,4,4,4,7 reflects three HRTF sets gaining a score equal to fourth place. This section presents the overall outcomes of the selection process, detailed examination of the consistency of participant responses between sessions, followed by outcomes of a simulated knockout tournament.

HRTF selections

For each of the 63 selection procedures, there were 21 tournament matches (or pairwise comparisons) completed by the given participant. Figure 4.9 shows that there was very little variation between HRTF sets in the outcome of all matches taken in aggregate. Each HRTF set won a share of its matches that was close to chance level of 50% (i.e. 189 of the 378 matches in which it featured). A repeated measures Friedman test confirms no significant difference between the popularity of each HRTF set ($\chi^2 = 7.75$; $p = 0.257$).

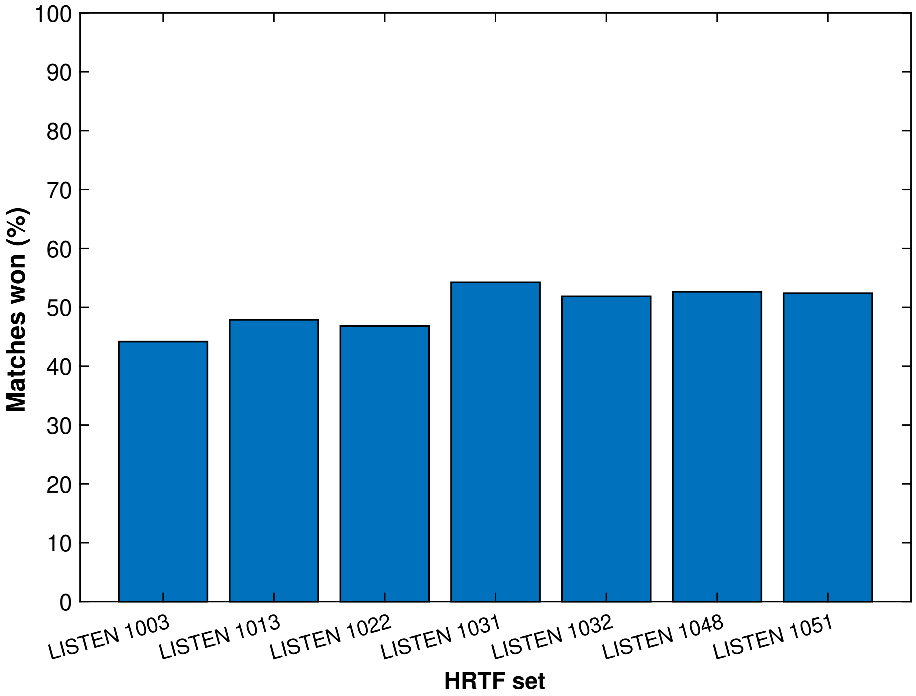


FIGURE 4.9: Percentage of matches won by each HRTF set in the 3D selection tournament

Although individual match outcomes show no greater or lesser overall preference for any one HRTF set, Figure 4.10 indicates that the 63 round robin tournaments produced a less even outcome between the choices, which ranged from 5-12 wins each. As discussed in detail in section 2.1, the cues that inform binaural perception are highly personalised and a product of individual ear, head and torso morphology – especially so for elevation perception, discrimination along the *cone of confusion* and sense of externalisation. However, previous research – and, indeed, analysis of data from Study 2a in Figure 4.2 – has uncovered trends that suggest some human-measured HRTF sets may be more amenable than others to satisfactory use by a wider section of the population. This is likely to manifest in cases where an individual's measured HRTF is closer to the average or dominant morphological characteristics of a given group (Armstrong et al., 2018; Katz and Parseihian, 2012; Kim, Lim and Picinali, 2020; Roginska, Santoro and Wakefield, 2010). In the case of this study, *LISTEN 1031* was the most frequency selected HRTF across all comparisons, although its status in that regard was not statistically significant. It was also one of the four best performing in terms of tournament wins (which includes tied results settled by random selection), along with *LISTEN 1003*, *LISTEN 1032* and *LISTEN 1051*. These suggestions that *LISTEN 1031* may have had widest appeal in the design of this rendering system and study, although not conclusive, turned out to be relevant in informing later phases of research. Those considerations are discussed further in conclusion of this chapter and in Chapter 6.

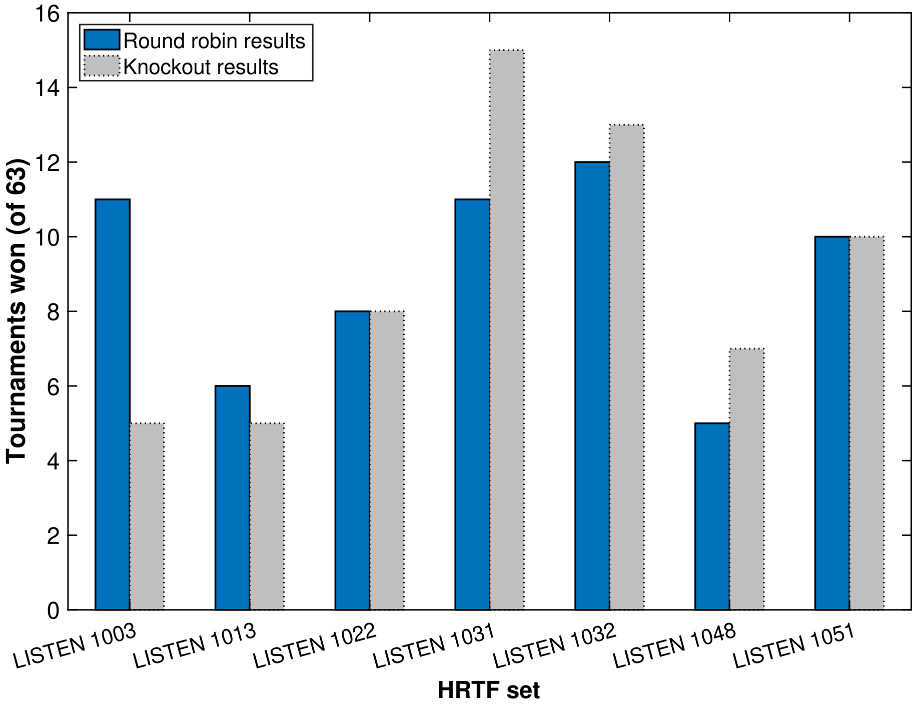


FIGURE 4.10: Outcomes from the two 3D HRTF selection tournament formats

Table 4.3 shows equivalent data to that shown in Table 4.1. On this occasion, a significant minority of outcomes (11) were achieved with the lower score of 4/6 and clear wins constituted only 17% of outcomes. For the 2D selection method used in Study 2b, winning margin was derived as an index to indicate the degree of certainty in the preferred HRTF selected by each participant, as reported in section 4.1.2. For the 3D selection process, the three repeated sessions were instead used to verify selection confidence (each of which was then further validated using the localisation task reported in section 5.2.

The average time taken across all sessions was a mean of 13 and median of 11.8 minutes. This is a marginal increase from the mean of 11.4 and median of 9.5 minutes required for the 2D selection process used in Study 2b, as reported in section 4.1.2.

Intra-class correlation measurement

Intra-class correlation (ICC) is a statistical approach used for measuring consistency between different raters to verify the robustness of a rating system (Hallgren, 2012). ICC has been used previously to evaluate the reliability of repeated HRTF set ratings expressed by the same raters (Kim, Lim and Picinali, 2020). The HRTF selection reliability established for each participant via ICC is presented in Table 4.4. Calculation of ICC was achieved using the R statistical computing package, according to the guidance and numerical outcome classifications provided in (Hallgren, 2012), where: less than 0.0 represents lower than chance levels of agreement (systematic disagreement); between 0.0 and less than 0.4 is an above chance (but weak) level of agreement; from 0.4 to less than 0.6 indicates

	Margin			***Totals***
	0	**1**	**2**	
	(tied win)	***(slight win)***	***(clear win)***	
Score of 4/6	11	0	0	***11***
Score of 5/6	9	27	2	***38***
Score of 6/6	0	5	9	***14***
Totals	***20***	***32***	***11***	

TABLE 4.3: Distribution of HRTF selection results in the 3D selection tournament

Systematic Disagreement			**Weak Agreement**			**Fair or Good Agreement**		
C	-0.306	(n)	*A*	0.349	(n)	*B*	0.429	(n)
D	-1.080	(o)	*G*	0.186	(n)	*E*	0.743	(p)
H	-0.095	(o)	*K*	0.075	(n)	*F*	0.437	(n)
J	-0.418	(o)	*O*	0.342	(p)	*I*	0.715	(n)
L	-0.840	(p)	*P*	0.380	(o)	*N*	0.726	(p)
M	-0.795	(n)	*Q*	0.142	(o)	*R*	0.648	(o)
T	-1.151	(p)	*U*	0.258	(o)	*S*	0.510	(o)

TABLE 4.4: HRTF selection reliability values and category for each participant (A-T), including binaural experience indicator (n = none; o = occasional; p = practised)

fair agreement; between 0.6 and less than 0.75 shows good agreement; 0.75 and beyond constitutes excellent agreement.

Information provided by participants regarding their background with music and audio was analysed against the groups in Table 4.4 using chi-square and Pearson correlation tests. No relationship was evident between the selection consistency groupings and either level of musical training (χ^2 = 8.88; p = 0.353), weekly headphone listening (r = 0.141; p = 0.541), or level of prior exposure to binaural audio (χ^2 = 2.72; p = 0.843). Each participant's self-declared experience with binaural audio is shown in Table 4.4 for reference.

Alternate tournament format

The alternate knockout tournament simulated for Study 2b (as described in 4.1.2) was applied identically to the results of the 3D selection process. The knockout format reduces the procedure from 21 to just 6 pairwise comparisons, with each round eliminating one HRTF set. Of the 63 comparisons, 15 had a substantial change in outcome when simulated as a knockout (i.e. resulted in a selection with a lower score than the result of the round robin format). Figure 4.10 shows the net difference in the knockout tournament outcome for each HRTF set compared to round robin selection processes. The knockout simulation resulted in an even wider variation of tournament wins between HRTF sets, ranging from 5-15 each.

4.2.5 Discussion

The four criteria for assessing an HRTF selection system put forward at the start of this chapter are revisited to evaluate results for the described method.

Reliability

Headline data suggests that the reliability of the proposed 3D HRTF selection mechanism may be lower than for the 2D method. Only 83% of selections were made with the two highest possible scores of 5/6 or 6/6, compared to 95% for the equivalent 2D HRTF selection outcomes. Moreover, the distribution of individual comparison outcomes viewed in aggregate in Figure 4.9 shows no differentiation in popularity, whereas the final round robin selection outcomes in Figure 4.10 shows wide variation. This is in contrast to the outcomes for 2D HRTF selection in Figure 4.2 and Figure 4.4, which showed some significant differences in individual pairwise comparisons that were also mirrored in overall selection outcomes.

Despite these indications of a lower level of certainty in participants' 3D HRTF selections compared to the method for 2D HRTF selection, there is some evidence indicating a degree of repeatability. Table 4.4 shows that a third of participants demonstrated a fair to good level of consistency in the rankings that resulted between their three HRTF selection sessions. A further third showed some tendency towards repeating their patterns of selection beyond chance level. The final third returned sets of rankings that actively diverged from each other to a greater degree than chance level. Although absolute values and proportions of participants between these groups do not indicate a mechanism that could yet be described as reliable, for a significant minority it was possible to attain outcomes that were repeatable to an acceptable level.

Validity

The validity of preferred HRTF selections was evaluated directly in part b of the study through a localisation test. In summary, that data shows apparent improvement to the accuracy of azimuth localisation in the upper hemisphere with participant's preferred HRTF set, compared to their least preferred. This trend was significant along the horizontal plane (i.e. for sources at 0° elevation). In contrast, evidence suggests that vertical localisation accuracy improved over time, rather than with HRTF preference. Section 5.2.2 and 5.2.3 report and discuss these outcomes in detail.

Usability

It is notable that the two most reliable raters judged themselves to be practised in binaural listening (participants E and N). However, statistical analysis found no significant relationship between ranking consistency and either musical training, headphone listening habits or prior binaural exposure. Moreover, some of those with only occasional (participants R and S) and even no binaural experience (participants B, F and I) were able to achieve fair or good levels of repeatability. All evidence therefore suggests that the approach was similarly effective for all profiles of user, irrespective of any experience that could have a determining effect on capacity to evaluate binaural rendering critically. This contrasts with many previous approaches, which have reported greater consistency with participants categorised as expert listeners (Andreopoulou and Katz, 2016; Andreopoulou and Roginska, 2014; Kim, Lim and Picinali, 2020; Schönstein and Katz, 2012).

Efficiency

The average completion time of between 12 and 13 minutes for the round robin selection process is outside the acceptable duration identified at the beginning of this chapter. However, as stated previously, the simulated knockout tournament reduced the number of required comparisons from 21 to 6 and resulted in either the same selection or an alternative of equal rank (i.e. a different HRTF that was tied in first place from the round robin) for 48 of the 63 sessions. A direct theoretical switch to knockout format therefore implies a procedure that is potentially only 3.7 minutes (or 29% of the round robin length) but with outcomes that are equivalent 76% of the time. In practice, however, there could actually be benefits to both accuracy and consistency by shortening the comparison process in this way. A six step process would not be as fatiguing and confounding as undertaking 21 pairwise comparisons of stimuli that, in many cases, are very difficult to distinguish. In this regard the option of a knockout tournament format has an alternative potential to yield more robust outcomes that would require dedicated investigation.

4.2.6 Study 3a summary

Overall, it is to be expected that the 3D HRTF selection procedure would perform less successfully than the 2D counterpart from Study 2b, for three main reasons:

1. Using a set of seven HRTFs requires six more comparisons, which presented greater scope for fatigue or distraction to occur amongst participants.
2. The nature of the 3D judgement requires consideration of a further dimension (height), which increased complexity and therefore likelihood of misjudgement.

3. Real-time visual representations of the sound source location were not included as these would be difficult to render clearly on a mobile device, which meant participants were making purely internalised judgements on a trajectory's apparent path.

Despite these additions to the cognitive load in the 3D comparison process, selection outcomes arrived at the two highest possible scores (5/6 or 6/6) in 52 of 63 sessions. This is just 7% short of the 90% threshold specified at the start of this chapter. It is less clear why the 2D selection process enabled marked rejection of one HRTF set by a majority of users, whereas the 3D method resulted in chance levels of preference across all options when the 1379 pairwise comparisons are viewed in aggregate. One possibility is that the nature of the optimisation process undertaken by Katz and Parseihian (2012) resulted in a list of seven that is more evenly balanced in its relative appeal to a pool of different individuals.

Importantly, by conducting comparisons between individual participants' ranking sessions, this study has revealed a degree of repeatability in the method amongst a substantial minority of users, which also appears agnostic to expertise. Given the holistic nature of the comparison judgement (simultaneously considering azimuth, elevation and externalisation) and duration of the overall selection process, there is clear and substantial potential for further development towards even more reliable usage. Important areas of focus towards this goal would be to:

- identify how the stimulus, trajectory or written guidance outlined in section 4.2.3 could be further improved to focus the comparison process;

- explore whether a knockout tournament improves selection validity and repeatability by reducing the length of the process and thus enabling more concerted user concentration.

4.3 Chapter summary

The efficacy of binaural music recommendation and discovery experiences will be impacted by the degree to which rendered scenes fit the perceptual profile of individual users. To date, there is no established method for personalising the HRTFs deployed for end-users of audio virtual or augmented reality in mobile contexts. When looking towards end-user applications, the process needs to: result in clearly identifiable preference(s) for the desired proportion of cases (be reliable); return an HRTF set that provides a faithful spatial image (be valid); be equally effective for expert and non-expert listeners (be usable); not burden the user through excessive and lengthy tests (be efficient).

Outcomes from Study 2b established the viability of a general approach that fulfilled these aims for selection of a preferred HRTF set for 2D rendering. Study 3a built on those findings to address the same problem for 3D rendering on a mobile device. Its first iteration has shown promising outcomes against the same evaluation criteria. The method was also assessed more comprehensively over repeated sessions to gauge repeatability and included a structured localisation test for more direct verification of the approach's validity. The latter is discussed in detail in section 5.2. The potential areas for improvement identified above in section 4.2.6 were originally planned for inclusion in the final study detailed in Chapter 6. However, as explained later in section 6.1.1, due to the COVID-19 pandemic that final study was conducted remotely and it was not deemed logistically possible to include an HRTF selection procedure as an additional preliminary step. Instead, the most frequently chosen HRTF set in Table 4.9, LISTEN 1031, was used for each participant involved that last piece of research.

Chapter 5

Mobile 3D virtual auditory display design and evaluation

Trumpets and violins I can hear in the distance,
I think they're calling our names.
Maybe now you can't hear them, but you will,
If you just take hold of my hand.

'Are You Experienced?'
Jimi Hendrix

It was argued in section 2.3 that, although Ambisonics is an increasingly dominant audio format for binaural immersive media, it is not necessarily the optimal approach for rendering interactive virtual auditory displays (VAD). A virtual speaker based implementation of vector base amplitude panning (VBAP) was advanced as a possible alternative. The first part of this chapter details a virtual VBAP implementation for 3D head-tracked binaural rendering on an embedded Linux system. The technical performance of virtual VBAP is evaluated alongside a First Order Ambisonics (FOA) approach on the same platform, using analysis of localisation cue error against a human-measured head-related transfer function (HRTF) set. This virtual VBAP implementation was the system used to undertake the 3D HRTF selection procedure already reported as Study 3a in section 4.2.

The second section of this chapter reports the findings from Study 3b; the localisation test that followed the 3D HRTF selection procedure. The aims of the test were twofold: first, to assess the validity of the 3D HRTF selection method and, second, to establish user error margins when the virtual VBAP system is deployed as an interactive system. The outcome of the first aim was summarised earlier in section 4.2.5 and is laid out in more detail here, along with findings related to the second aim.

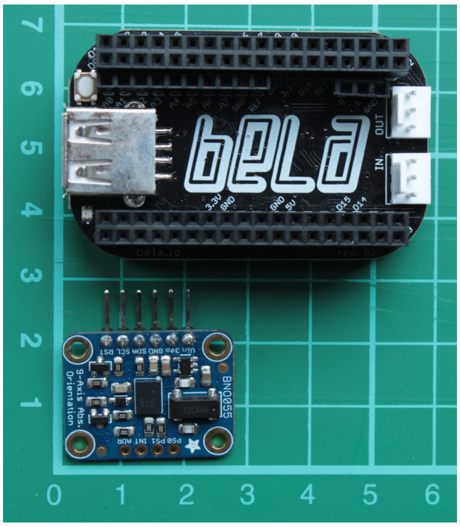


FIGURE 5.1: The Bela Mini embedded Linux platform (top) and Bosch BNO055 IMU (bottom). The grid illustrates component dimensions (in centimetres).

5.1 Binaural virtual auditory display system and measurement

This section describes a system developed for mobile, interactive binaural auditory display. It details the hardware used for real-time audio processing and head-tracking, the software implementation of virtual VBAP and first order Ambisonics rendering systems, and the methodology used for evaluation. The outcomes of the technical assessment in terms of localisation cue error in both approaches are presented, followed by discussion of of these findings' significance for real-time, 3D, binaural VAD applications.

5.1.1 Implementation

Two real-time head-tracked binaural systems were developed on the same embedded platform: a virtual VBAP renderer and a virtual FOA renderer. The hardware used and software design are outlined in detail below, along with the methodology used for comparative evaluation.

Hardware

Bela is a commercially-available, open-source, embedded Linux platform for low-latency audio and sensor processing. The software described here was run on the compact 'key-fob' sized Bela Mini model (Figure 5.1), built to achieve a high degree of portability. The Bela Mini runs a 1GHz ARM Cortex-A8 processor and has two channels of audio output (Bela, 2018). Audio processing on the full sized Bela (which also runs a 1GHz ARM Cortex-A8) has been shown to perform at sub-millisecond levels of round-trip latency (McPherson, Jack and Moro, 2016). C++ was used to programme both the VBAP and

Ambisonics systems outlined below. A link to the source code for the VBAP implementation and measurement is provided in Appendix A.2.

MrHeadTracker is a low-cost plug-and-play system for incorporating head-tracking on Arduino embedded computing platforms. It uses the Bosch BNO055 nine degrees of freedom sensor with on-board processing to measure, compute and output quaternions or Euler angles (Figure 5.1). The device performs with a refresh rate of 100Hz and is shown to have an angular standard deviation between 0.5° and 2.5° (Romanov et al., 2017). The second supervisor for this doctoral research, Dr Rebecca Stewart, had ported the *MrHeadTracker* code for integration with Bela (Stewart, 2018).

VBAP binaural rendering software

Sections 2.1.3 and 2.3 explained why and how an HRTF set is used to synthesise the position of virtual speakers and generate a binaural signal. The IRCAM LISTEN database was chosen as the source from which to select a human-measured HRTF for this investigation (Warusfel, 2003). As explained earlier in section 4.2.3, this collection was favoured because it has been systematically rationalised into an optimal shortlist of seven HRTF sets that, as a group, is judged to present the highest strength of preference for the largest proportion of users (Katz and Parseihian, 2012). LISTEN HRTF set 1013 was randomly chosen from the optimised shortlist for use in the analysis that follows here. It should be noted that the convention used within the LISTEN database for expressing azimuth angle differs from that expressed in section 2.1, instead incrementing continuously from 0° anticlockwise through 360°, as shown in Figure 5.2. This co-ordinate format is used throughout the rest of section 5.1.

3D VBAP allows any combination of three or more loudspeakers placed equidistantly from the listener to render spatial audio. A working constraint of eight virtual speakers (therefore 16 individual convolutions with left and right HRIRs to render the binaural scene) was established for realising this auditory environment. The aim was for the system to render the effect of pseudo-spherical surround sound within that constraint. Full surround rendering was also necessary to enable direct objective comparison against a virtual FOA implementation. Similarly, the limit of eight virtual speakers established a degree of parity in the computational resource demanded from the binauralisation processing for both VBAP and FOA (i.e. 16 convolutions in each case), which made direct cross-evaluation of the two spatialisation algorithms' performance more viable.

As described further in section 2.3, the virtual VBAP layout deliberately concentrated more loudspeakers towards frontal locations. It was hoped that improving resolution in this zone would benefit interactive auditory display development planned for the final research phase. The resulting position of each virtual speaker is given in Figure 5.2. The five placed on the horizontal plane (at 0° elevation) are spread in 60° increments, but with a 120° gap at the rear. A single speaker at the zenith enables upward triangulation. Since

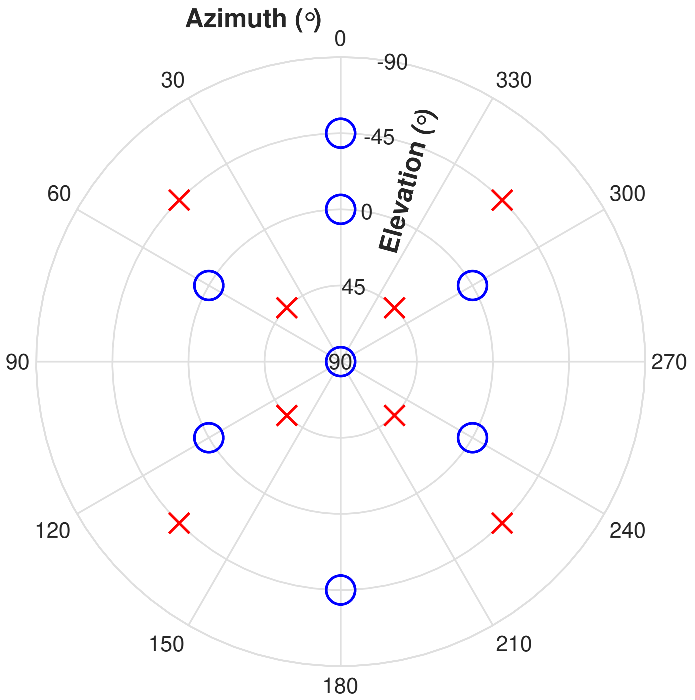


FIGURE 5.2: Representation of the VBAP and FOA systems' virtual speaker positions. ○ = VBAP system; × = FOA system.

the HRTF set does not feature a measurement at the nadir (as is typically the case), two placed at -45° allow downward triangulation.

A flow diagram of the virtual VBAP system implementation is shown in Figure 5.3. VBAP weightings for the virtual loudspeaker layout have been precalculated using the code developed by Politis (2015). The resulting lookup table of speaker feed gains for every possible angular position at 1° increments is loaded as a matrix on startup. Standard resolution audio files (16 bit 44.1 kHz) are called and streamed from any predefined azimuth/elevation co-ordinate. Head-tracking readings are refreshed at 86Hz (every 512 samples) to update the position of each sound source. A buffer of 2048 samples is used to comfortably meet processing deadlines for Fourier transformations and frequency domain convolution, which is in addition to the default Bela system buffer size of 16 samples. The Bela digital-to-analogue converter is known to introduce a further 21 samples of delay (McPherson, Jack and Moro, 2016). Maximum system latency is therefore 2597 samples, or 59 milliseconds, which is within the 75 millisecond response time advocated by Suzuki, Yairi and Iwaya (2007).

Ambisonics reference software

To evaluate the VBAP software's performance, the same hardware was used to render a virtual FOA environment. The virtual FOA reference system was implemented through supervision of a Masters research student, Teodor Radu. This system adapted the *libspatialaudio* C++ encoding/decoding library (Digenis, 2017). The *libspatialaudio* library

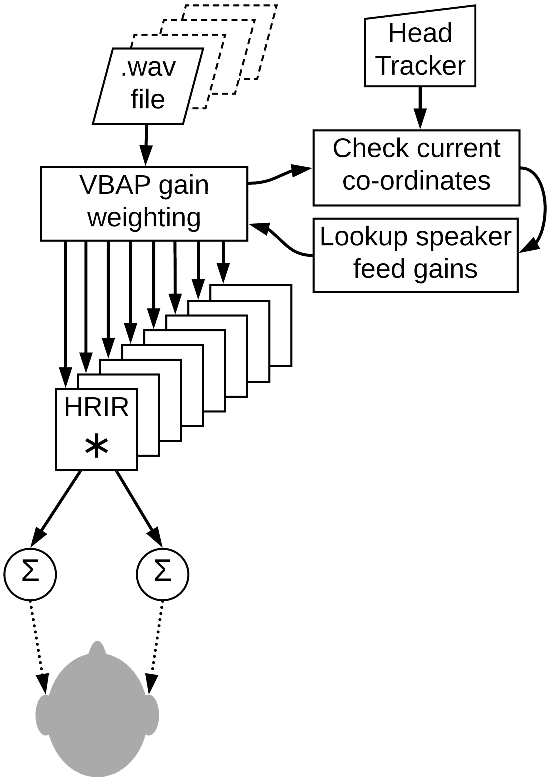


FIGURE 5.3: Flow diagram of the virtual VBAP implementation

adopts MaxRe weighted and All-round Ambisonic Decoding algorithms for psychoacoustically optimised source representation, which has been shown to improve rendering fidelity (Zotter and Frank, 2012). Head-tracking was included by using BNO055 output data to rotate the B-Format sound field using on board functions. The library renders 3D FOA binaurally using virtual speakers placed in a cube arrangement with a chosen HRTF set. However, the processing required to run this code in its original form proved too computationally intensive for Bela. Instead, binauralisation was achieved using an Ambisonics-to-binaural optimisation, which calculates direct transfer functions for each B-Format channel, the principle for which is outlined by McKeag and McGrath (1996). A cube virtual loudspeaker arrangement was retained for the resulting implementation, because it was the configuration used in the Google VR Audio FOA binaural decoder at the time of devising the analysis[1]. That decoding approach was already being widely used in the Google VR SDK, YouTube360/VR and the Omnitone browser-based spatial audio rendering library, so could be viewed as a de facto standard for immersive media applications. The virtual FOA cube array therefore constituted a legitimate reference implementation against which to evaluate the relative performance of virtual VBAP. HRIRs from LISTEN 1013 were again used to generate binaural B-Format, which then only required eight convolutions (two for each of the channels W, X, Y and Z) to render 3D FOA

[1] https://www.york.ac.uk/sadie-project/GoogleVRSADIE.html [accessed 15/2/2023]

at virtual speaker locations defined in Figure 5.2.

Localisation cue error measurement

Section 2.1 summarised how auditory perception of spatial location can be quantified by three measurements: interaural time difference (ITD), interaural level difference (ILD) and the monaural and binaural spectral cues introduced by filtering from pinnae and (to lesser extents) head and torso reflections. HRTFs encode into HRIRs the ITD, ILD and spectral shaping experienced by an individual at given spatial locations. Whilst ITD and ILD combine to determine lateral binaural localisation, the spectral response found across an HRTF set forms cues that determine elevation perception, discrimination along the *cone of confusion* and sense of externalisation. In effect, both the VBAP and FOA systems synthesise new sets of HRIRs via virtual loudspeaker realisation. The response of these synthetic HRIRs can be compared directly to the original HRTF set from which they were derived and in respect of each of these localisation cues (Kearney and Doyle, 2015; Wiggins, 2017). To achieve this, unit impulse signals were fed into either system at each of the 187 LISTEN database co-ordinates and the outputs processed to derive localisation cue metrics compared to the original HRTF set, as follows:

- *ITD* – Measurement techniques advocated by Katz and Noisternig (2014) were applied to calculate ITD. VBAP-generated, FOA-generated and original HRIRs were first filtered with a tenth order Butterworth lowpass filter at 3 kHz. ITDs for each set and position were then computed in microseconds using cross-correlation functions from the MATLAB Signal Processing Toolbox[2].
- *ILD* – VBAP-generated, FOA-generated and original HRIRs were first filtered with a tenth order Butterworth highpass filter at 1.5 kHz. ILDs for each set and position were then computed as the mean-squared power difference in decibels.
- *Spectral response* – Peak normalisation was applied to both the VBAP-generated and the FOA-generated HRIRs. In each case and to ensure uniform gain increase, normalisation was referenced to the most significant unsigned value found in either the left or right channel of all 187 HRIRs viewed collectively. The VBAP-generated, FOA-generated and original HRIRs were then processed with a 40 band gammatone filter bank implemented by The Institute of Sound Recording Surrey (2016), with lower and upper centre frequencies at 0.1 kHz and 16 kHz . For the VBAP-generated, FOA-generated and original HRIRs, spectral response was calculated as the mean-squared power of each band, for either channel, at every position.

[2] uk.mathworks.com/help/signal/ref/finddelay.html [accessed 3/6/2022].

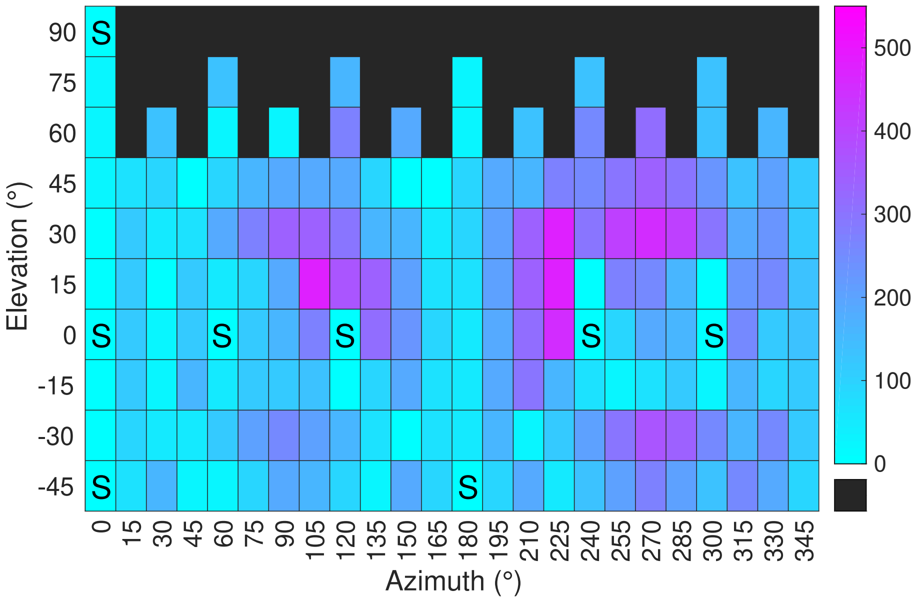


FIGURE 5.4: Heatmap of Virtual VBAP unsigned ITD error (μs).
S = virtual speaker location.
■ = location not measured in the original HRTF set.

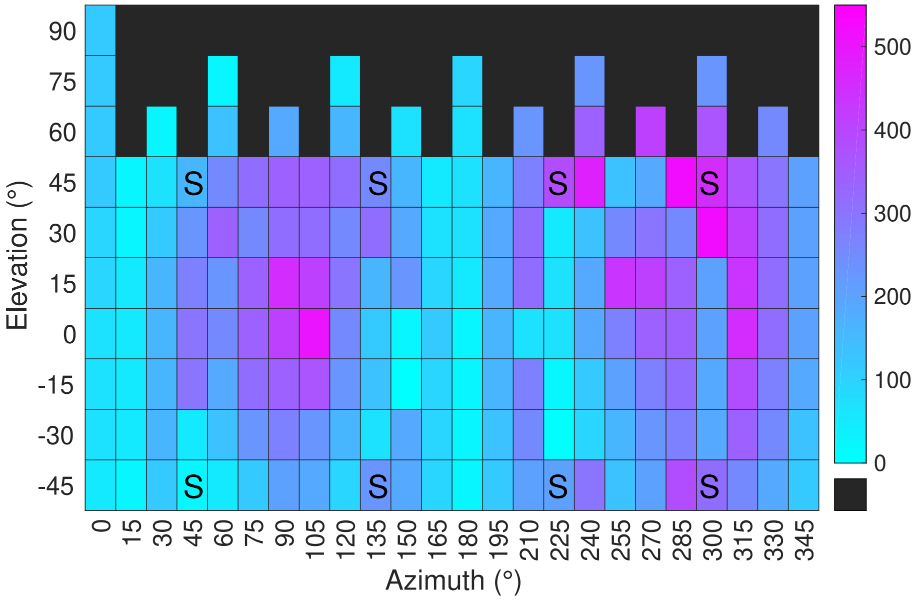


FIGURE 5.5: Heatmap of virtual FOA unsigned ITD error (μs).
S = virtual speaker location.
■ = location not measured in the original HRTF set.

5.1.2 Results

To evaluate the VBAP and FOA systems, the ITD, ILD and spectral shaping error measurements described above are compared against the original LISTEN 1013 HRTF set ground truth.

	All locations	**Front locations***	**-45° elevation**	**0° elevation**	**45° elevation**
VBAP	150 (114)	115 (98)	120 (79)	140 (120)	151 (98)
FOA	202 (123)	175 (117)	157 (97)	216 (138)	246 (140)

TABLE 5.1: Mean and standard deviation of unsigned ITD error (μs) for virtual VBAP and FOA systems, by location

** Front locations include all elevation points within 0 to +/- 60° azimuth*

	All locations	**Front locations***	**-45° elevation**	**0° elevation**	**45° elevation**
VBAP	3.18 (2.58)	1.75 (1.45)	3.43 (1.61)	1.62 (1.43)	4.83 (3.49)
FOA	4.69 (3.05)	3.90 (2.50)	5.66 (3.03)	3.85 (2.08)	4.91 (3.69)

TABLE 5.2: Mean and standard deviation of unsigned ILD error (dB) for virtual VBAP and FOA systems, by location

** Front locations include all elevation points within 0 to +/- 60° azimuth*

ITD Error

Unsigned ITD errors for each system compared against the full LISTEN 1013 HRTF set are presented in Figures 5.4 and 5.5, for all 187 locations. As expected, no error is seen at positions where virtual VBAP speakers are located. At these points the signal is reproduced solely with the LISTEN 1013 HRIR for that origin. This is not the case with FOA, where there is no direct relationship evident between ITD error and speaker location. Overall, greater deviation from the ground truth is generally apparent across wider areas in the case of FOA.

These tendencies can be seen more clearly in Table 5.1, where error rates are summarised into location groups. Inaccuracies are seen to be consistently lower and more stable in the VBAP implementation than FOA across every zone. In particular, virtual VBAP shows considerably less error than FOA at frontal locations and on the horizontal plane – areas where the VBAP speaker layout has been specifically concentrated to provide greater resolution. Even at locations where the VBAP speakers are most dispersed and FOA more concentrated (+ or - 45° elevation), virtual VBAP ITD is more closely aligned to the original HRTF set.

ILD Error

Unsigned ILD errors for each system when compared against the full LISTEN 1013 HRTF set are presented in Figures 5.6 and 5.7 for all 187 locations. In this instance, the pattern of error is similar between implementations, but its extent is greater for FOA. Virtual VBAP again benefits from points of no error at speaker locations, where FOA does not. Table 5.2 also confirms that VBAP outperforms FOA in every location grouping for ILD, with less than half the average error of FOA at frontal and horizontal plane positions.

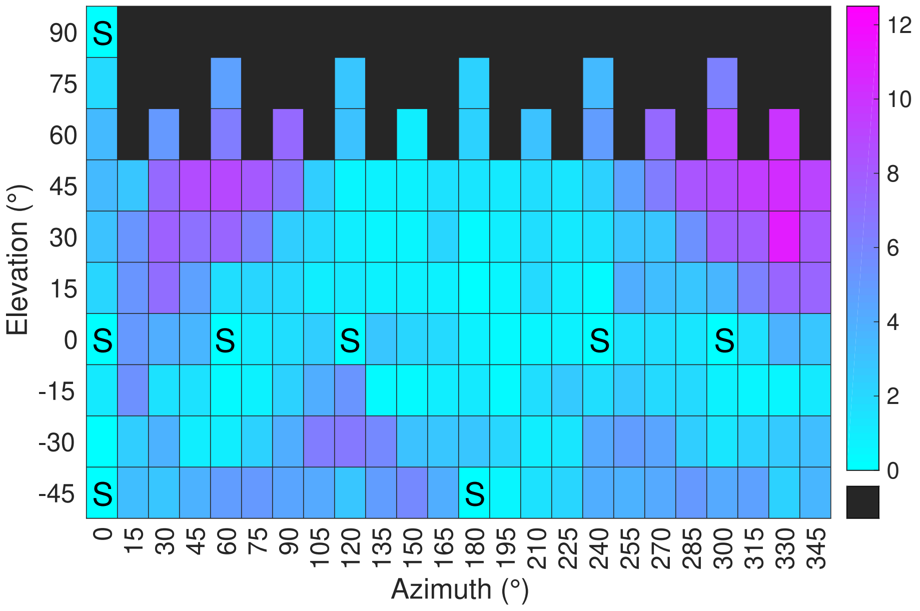


FIGURE 5.6: Heatmap of virtual VBAP unsigned ILD error (dB).
S = virtual speaker location.
■ = location not measured in the original HRTF set.

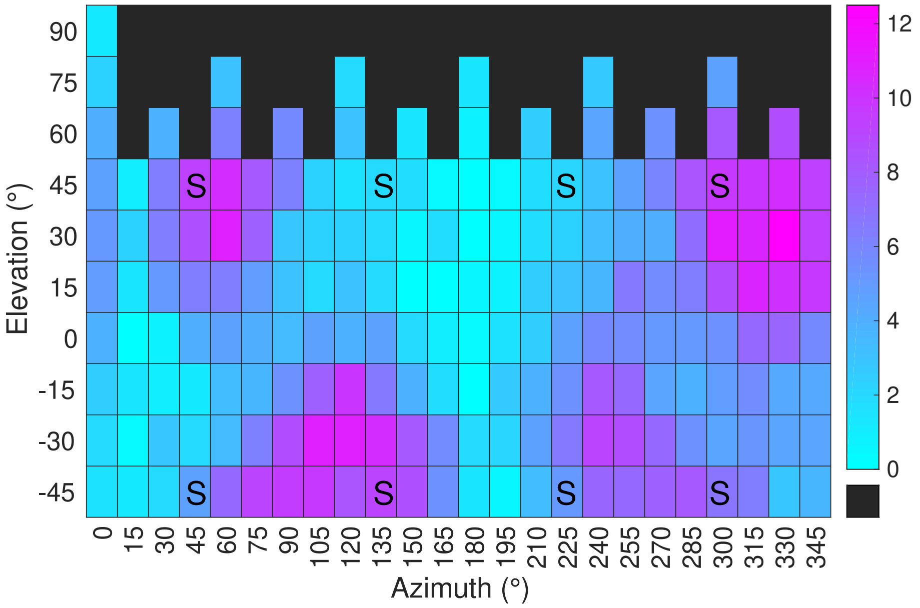


FIGURE 5.7: Heatmap of virtual FOA unsigned ILD error (dB).
S = virtual speaker location.
■ = location not measured in the original HRTF set.

Spectral Error

The mean unsigned spectral errors for each system and channel when compared against the full LISTEN 1013 HRTF set are presented in Figure 5.8. FOA has a lower mean spectral error for both left and right channels quite consistently up to about 1-2 kHz. Above 2 kHz,

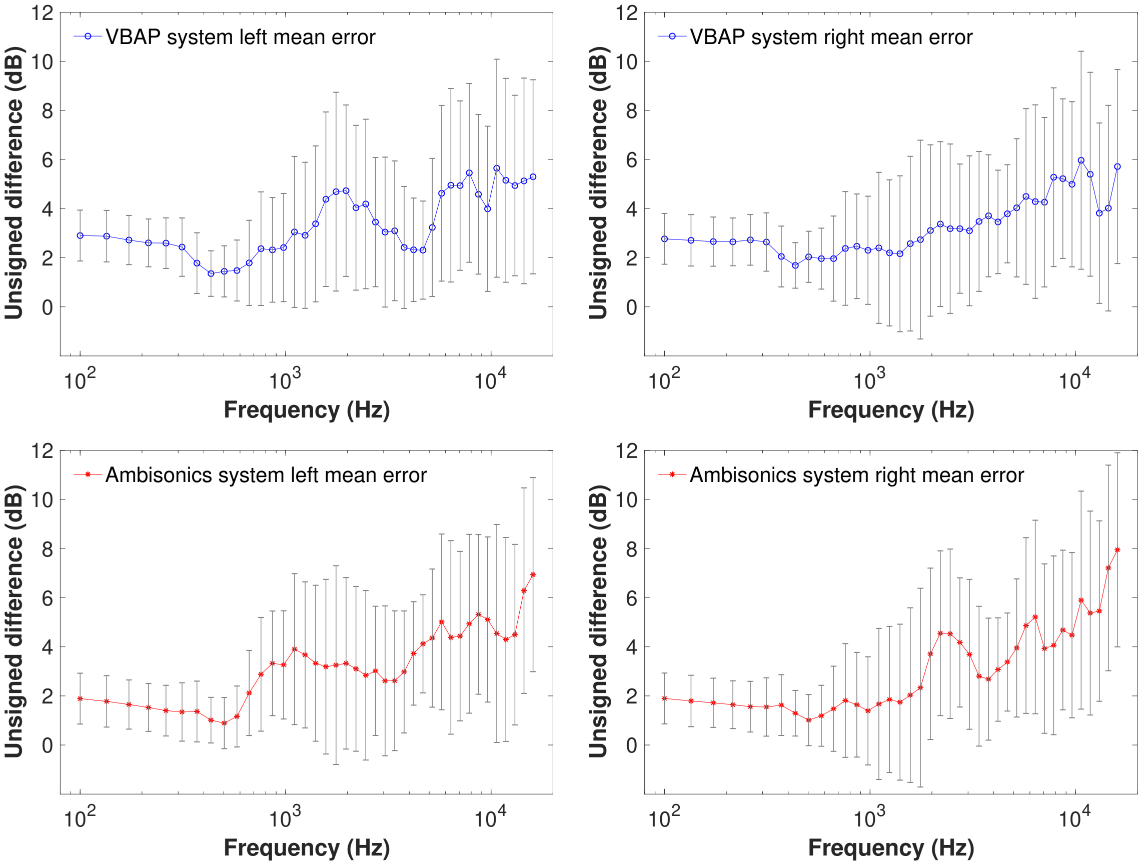


FIGURE 5.8: Unsigned spectral error for all locations, by frequency band, system and channel. A gammatone filter bank was used to compute the mean error (plots) and standard deviation (error bars) for all 187 HRIR locations at 40 frequency bands.

there is a fairly steady and relatively even increase in mean error seen for both systems and in either channel.

Figure 5.9 gives some illustration of where the error manifests at different locations. Points showing ≈ 0dB error occur at virtual VBAP speaker positions, as expected. It is also apparent that the FOA error, although generally lower, tends to show greater asymmetry between left and right channels than VBAP. This is particularly so at lower frequencies, where the performance of FOA is seemingly better in aggregate (in Figure 5.8). The VBAP system error is somewhat more symmetric between channels than FOA. There are slight indications that the VBAP virtual loudspeaker configuration has resulted in lower spectral error in some frontal locations (particularly at 0° elevation). However, unlike with ITD and ILD error, in general spectral response at frontal locations is not clearly shown to be any more in line with the original LISTEN HRIRs than FOA.

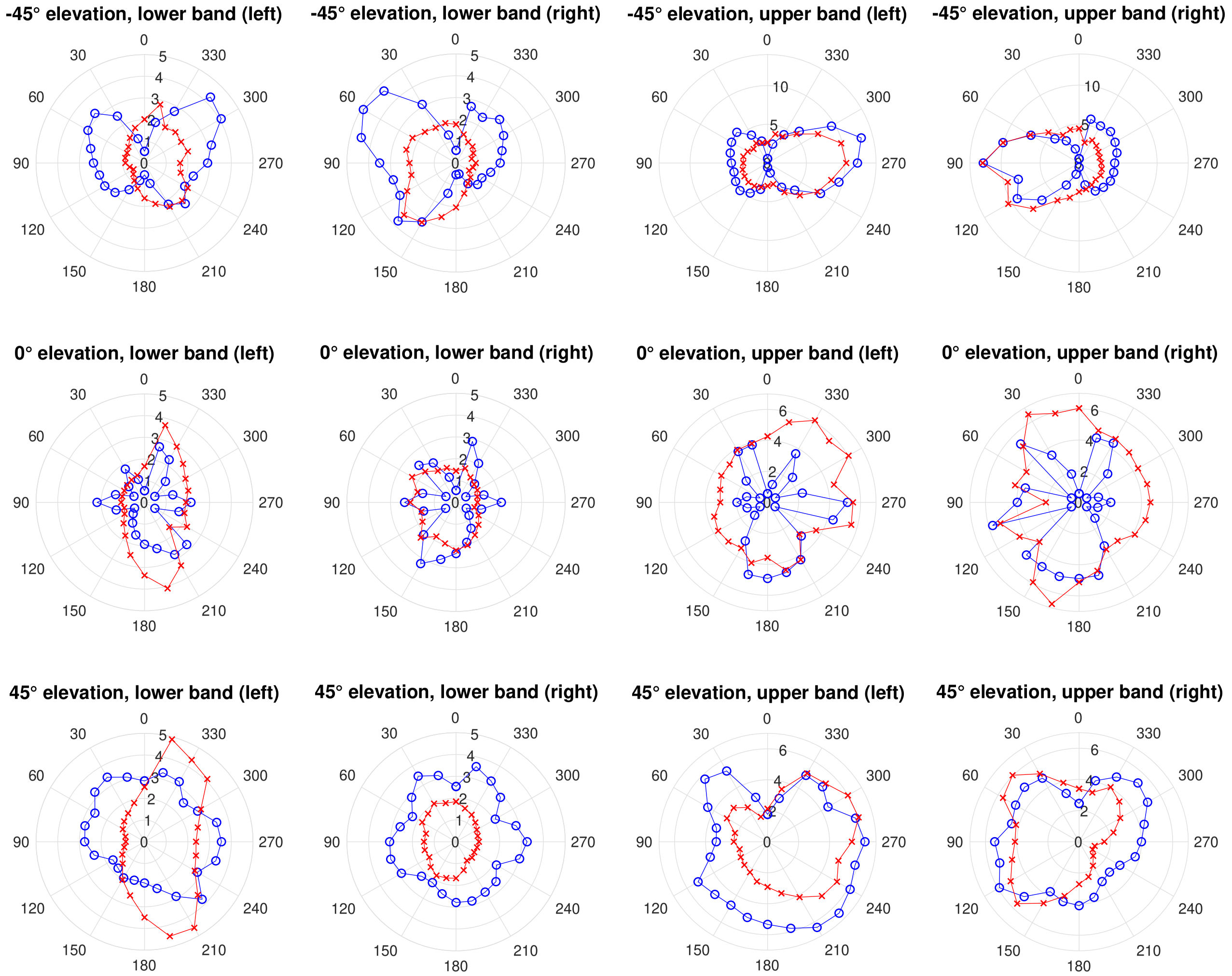


FIGURE 5.9: Mean unsigned spectral error (dB) in upper/lower frequency band and L/R channel groups, for selected locations.**

** *The lower band grouping includes the first 17 filters, with centre frequencies ranging from 0.1 to 1 kHz. The upper band grouping includes the remaining 23, from 1.5 to 16 kHz. (Note varying plot scales.)* KEY: **–o–** = VBAP system; **–x–** = FOA system.

5.1.3 Discussion

Impulse response analysis suggests the virtual VBAP system has clear reproduction benefits for horizontal localisation. This configuration seems to reconstruct original HRIR ITDs with fewer "black spots" of significant divergence than FOA and with lower error overall. ITD is most pertinent to lateral perception of frequencies below 1.6 kHz, thus it could be expected that VBAP would provide a more faithful representation of source azimuth for low-frequency dominated sounds and from a broader range of locations. Although the pattern of ILD error is loosely shared by both systems, VBAP again shows closer adherence to the original HRIRs across all positions. ILD provides the dominant horizontal cue above 1.6 kHz, so it can be anticipated that localisation would also be more acute throughout the sound field within this frequency range (Blauert, 1997).

In contrast, spectral error analysis suggests FOA has an aggregate frequency response that seems more faithful to source HRIRs. A simple interpretation might hypothesise that FOA could therefore present a more reliable representation of the original HRTF set's elevation cues, since frequency shaping inherent in HRIRs is key to vertical localisation. However, the majority of elevation cues are derived from HRIR frequency responses above 4 kHz (Kearney and Doyle, 2015), where spectral error patterns between the two implementations become similarly discontinuous. As such, there is actually little data to show that HRTF set elevation cues would be preserved any more faithfully by the FOA system.

In fact, evidence of left/right ear asymmetries in the FOA spectral error (particularly in lower frequencies) suggests that extent of colouration is more location-dependent than for VBAP. It is possible that this unevenness is a byproduct of irregularities in the human-measured HRTF set itself – whether down to subject morphology or marginal discrepancies in the measurement procedure – in which case the VBAP configuration appears to smooth discontinuity with more even error distribution. Understanding the perceptual effect of these contrasting left/right channel spectral error patterns would require further investigation.

Finally, it is clear that on the horizontal plane VBAP rendering represents the source HRTF set's ITD, ILD and spectral response more closely. Having more faithful rendering on this particular plane is a significant benefit to any VAD system for source segregation purposes. Furthermore, frontal locations are also shown to reproduce HRTF set ITDs and ILDs more accurately with the proposed VBAP loudspeaker configuration. Combined with head-tracking, this approach enables dynamic enhanced resolution rendering across the 3D sound field.

5.1.4 Summary of virtual VBAP system

This section outlined the design and performance of a portable embedded system for rendering multi-channel spatial audio scenes binaurally in real time, using virtual VBAP. The system supports head-tracking within recommended levels of latency. Technical analysis shows that it outperforms an equivalent FOA setup when assessed against key localisation metrics. These specific findings related to virtual VBAP are in line with previous investigations into FOA and VBAP spatial reproduction more generally (Pulkki, 2001b; Satongar et al., 2013). Indications from that existing research and the findings uncovered here suggest that second or even third order Ambisonics is required to meet localisation cue reproduction accuracy of virtual VBAP (Pulkki, 2001b). This would require considerably more computational resource that might not be available on embedded devices or other working contexts with similar CPU resource constraints, such as mobile or wearable technology. Thus, in scenarios utilising embedded or other portable, low-resource computing platforms, the nature and requirements of the immersive or interactive audio application at hand may determine whether virtual VBAP is a viable (or even preferable) approach compared to virtual FOA or HOA. Such use cases would include graphics-less user interfaces (as is the topic of this thesis), or more broadly for purposes encompassing multi-modal media experiences, art installations and interactive live performance.

5.2 Study 3b: Subjective localisation test

The process of development and analysis described in section 5.1 provide an objective measurement of virtual VBAP's performance. To further assess the system's performance for the intended application, it was important to conduct a subjective perceptual evaluation of the binaural rendering using an interactive context. The user study presented in this section addresses this consideration and consequently identifies known system limitations. The study design also validated the 3D HRTF selection procedure devised in Chapter 4, outcomes for which have been briefly summarised in section 4.2.5.

The test environment and participant makeup was identical to that outlined in relation to part a, in section 4.2.2. For each of the participants' three sessions, the localisation test (part b) was conducted following a break of around five minutes after the 3D HRTF selection procedure (part a).

5.2.1 Experimental design

The Sennheiser HD650 playback headphones were secured to participants with an elasticated hairband. No equalisation for the purpose of binaural synthesis was applied. The head-tracking sensor was mounted securely on top of the headphones to counter-rotate the sound scene and take head position readings at 1° angular resolution. Before starting, the head-tracker was calibrated to a personalised position measured and agreed as approximately directly ahead and level with the participant's eyeline and therefore considered as 0° azimuth, 0° elevation. Figure 5.10 shows the physical setup of the test environment. The localisation stimulus used 20 seconds of continuous music from the

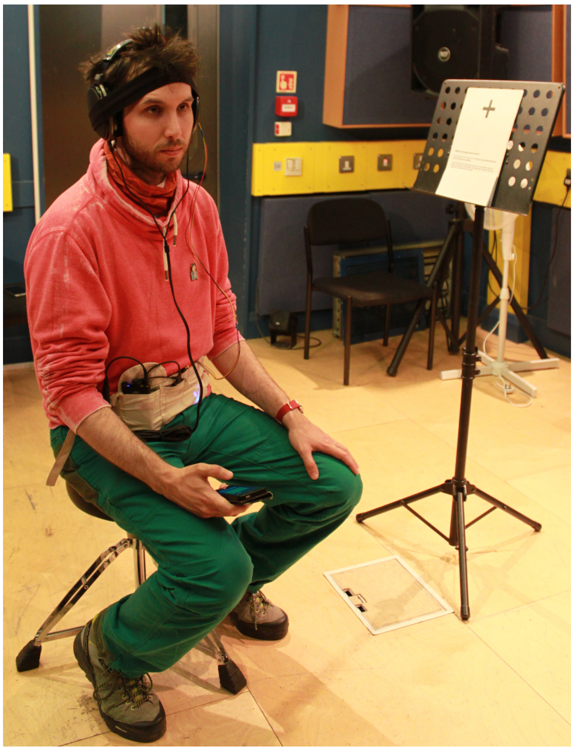


FIGURE 5.10: Localisation testing setup showing equipment worn, personalised calibration point and smartphone interaction mode

same recording used in part one (Hansen and Munch, 1991). A link to the source code repository is provided in Appendix A.2.

Target locations

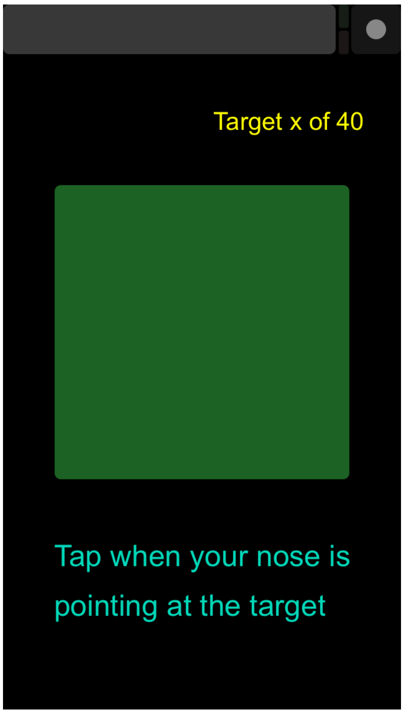
Localisation targets were divided into three strata, so that the anticipated shortcomings in upper and lower hemisphere rendering cues highlighted by the data in section 5.2.2 could be assessed independently:

- **at** 45° **elevation** – seven azimuths of -153°, -102°, -51°, 0°, 51°, 102° and 153°
- **at** 0° **elevation** – six of the azimuths stated above (0° was not used)
- **at** -45° **elevation** – the seven azimuths stated above

Localisation process

The test used egocentric head-pointing to report perceived source position, comparable to Bahu et al. (2016). Participants used the simple smartphone GUI in Figure 5.11 with the instruction:

> ***Where is the target sound source?***



Target x of 40

Tap when your nose is pointing at the target

FIGURE 5.11: Smartphone user interface for localisation test response submission

> *Find the location of the target sound. Point your nose towards what you hear to be the source position.*
>
> *The source will be from somewhere around you and sometimes above or below your ear level. In some cases, you might need to rotate in your seat and/or tilt your head to point accurately.*

Participants completed two example responses to check that they understood what was required before starting the task. For each response, both the time elapsed and variance in head position from target location was logged automatically (as azimuth and elevation co-ordinates). Participants were allowed as much time as they needed to respond for each target. The 20 second excerpt continued on a loop, if necessary, until they registered a response, after which time the next target location began automatically.

Both the sequence of 20 co-ordinates and the order of the two groupings (preferred and least favoured HRTF set) were randomised for each participant, at all three sittings. Therefore, a total of 120 data points was recorded for each participant using between a minimum of two (in the event of perfectly repeated best and worst selections) and maximum of six different HRTF sets.

5.2.2 Results

All participants completed three study sessions at least 48 hours apart. During the session and on later examination of data, it became evident that one participant had not understood the requirements of the localisation task and had provided responses that did not actively seek out the position of the sound source. This participant's data was reflected in the analysis for part a (3D HRTF selection), but is not included here for part two.

Interactive localisation accuracy tolerances

Three factors mean a reasonable degree of error was to be expected, particularly at the upper and lower strata (45° / -45° elevation). Firstly, even under optimal acoustic conditions, localisation blur of broadband sound immediately in front of a listener is established to be in the order of +/- 3.6° for azimuth and +/- 9° for elevation (Blauert, 1997). Secondly, inaccuracy in head pointing orientation was a further contributor to response error. Bahu et al. (2016) suggest that, for sources with substantial vertical displacement (57°), head pointed localisation can introduce mean unsigned error of 3° in azimuth and 12° in elevation. A further consideration related to pointing precision is the effect of laterality (i.e. left- or right-handedness) on perception of auditory space. Ocklenburg et al. (2010) found systematic bias in sound localisation that was contralateral to the dominant hand. This was evident for both left- and right-handed participants and whether hand- or head-pointing was used as the mechanism for indicating source locations. Thirdly, sparseness of the virtual speaker array in the upper and lower binaural hemisphere would have degraded spatial representation of sources originating in these areas far beyond optimal acoustic conditions (Baumgartner and Majdak, 2015).

Given these constraints, minimum standards of accuracy were established to evaluate localisation outcomes. For azimuth, a tolerance of +/-15° was used to test whether the rendering system could provide reliable interactive presentation of sound sources at a minimum lateral separation of 30°. For elevation, a +/-22.5° threshold was applied to test simply whether users could reliably distinguish between sources located above, below and on the horizontal plane.

Influence of HRTF selection

Quality of HRTF fit could have impacted both response accuracy and time in the localisation test. Figure 5.13 shows the distribution of participant outcomes for best and worst HRTFs. Plots show the distribution of participants' overall azimuth and elevation success rate and their mean response duration, for each stratum (45°, 0° and -45°). If there were objective interaction benefits to the HRTF selection procedure, higher successful identification rates and lower mean response times would be expected . This is only evidenced clearly in relation to azimuth accuracy in the upper hemisphere (upper and middle plots of column one in Figure 5.13). A Wilcoxon signed rank test confirmed significant improvement in accuracy of azimuth for sources placed on the horizontal plane, when using a preferred HRTF set compared to least preferred ($p = 0.047$). The same

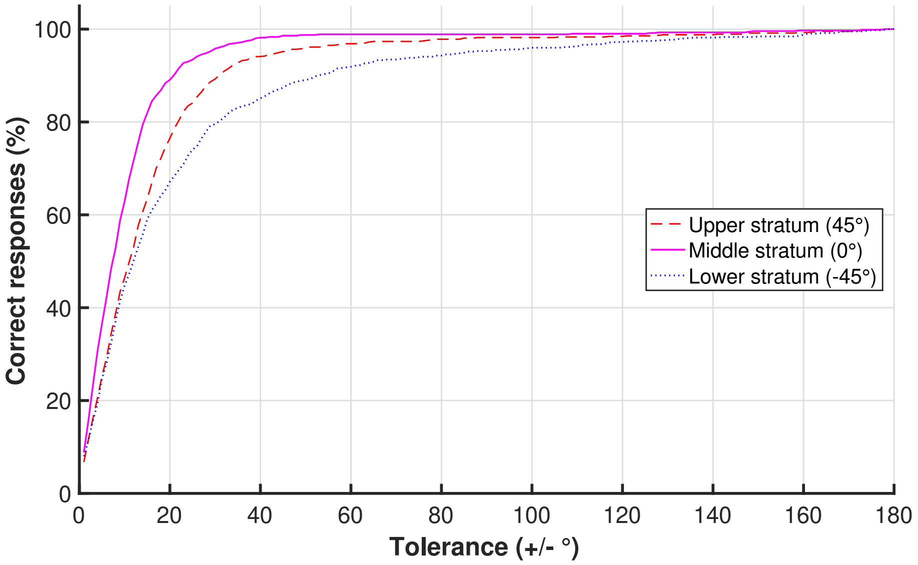


FIGURE 5.12: Proportion of correct azimuth responses from all test targets according to applied tolerance value, by elevation stratum

non-parametric test for significance did not uncover any other effects from using the best judged HRTF set, for any of the remaining eight metrics in Figure 5.13.

Further analysis was conducted to evaluate the influence of HRTF selection consistency on localisation performance. A Kruskal-Wallis test was conducted between the participant groupings shown in Table 4.4 (but without the participant identified in earlier, who was within the 'Weak Agreement' group) and the same nine metrics reflected in Figure 5.13. Significant difference was found in elevation accuracy at 0° ($p = 0.008$, $\chi^2 = 9.575$). Post-hoc Tukey-Kramer analysis identified that the 'Systematic Disagreement' group performed with significantly better accuracy against this metric than the 'Weak' group.

Influence of learning effects

Figure 5.14 represents the distribution of participant outcomes when viewed as the first and second halves of their sessions (irrespective of best/worst HRTF sequencing, which was always randomised). A Wilcoxon signed rank test confirmed significant improvement in accuracy of elevation identification with sources placed at 45°, for responses given in the second half of localisation trials ($p = 0.041$). The same non-parametric test further identified that responses at 45° ($p = 0.003$) and 0° ($p < 0.001$) were quicker in the second half, without any detrimental impact on azimuth or elevation precision.

Overall azimuth tolerances

Figure 5.12 shows the proportion of correct azimuth responses returned, according to the tolerance value applied. Overall, lower tolerance values are needed for the middle stratum to achieve a higher proportion of correct azimuth responses. The benefits of widening the tolerance start to tail off for the middle stratum around about +/-20°, then plateaux from about +/- 45°. By this point, residual errors such as front/back reversal, unfocussed or fatigued attempts, or accidental answers are likely to form a substantial portion of incorrect azimuth responses. Performance at the upper stratum is generally less accurate, but with similarly continual benefits gained from more widening tolerances up to +/-20°, then again flattening around +/- 45°. The lower stratum follows a noticeably different trajectory, with gains from wider tolerance slowing around +/-15°, then about 10% of responses still remaining incorrect even with a +/-45° margin of error for azimuth.

5.2.3 Discussion

Findings related to the two aims of the localisation study - validation of the HRTF selection approach and identification of interaction error margins - are addressed separately in detail.

Validation of HRTF selection consistency

Analysis showed apparent benefits to azimuth localisation accuracy in the upper hemisphere from preferred HRTF selection, which was significant along the horizontal plane. It is unsurprising that preferred HRTFs were of most benefit across this dimension, where five of the eight virtual speakers reside. Although this finding validates the selection approach in one respect, it is notable that positive elevation detection was increased by general exposure to the localisation task (albeit from a low starting base). This improvement and accompanying increases in response speed occurred independently of best or worst HRTF usage. The selection routine might therefore be validly applied in tandem with a structured pre-exposure phase to optimise perceptual experience.

It should also be noted that no meaningful statistical relationship was found between participants' HRTF ranking consistency and localisation performance. Significantly improved elevation accuracy was found in the 'Systematic Disagreement' group for sources on the horizontal plane. However, this apparent strength is actually a by-product of that group returning a greater overall proportion of responses that neglected vertical localisation and remained overly focussed at 0° elevation. The group was less likely to have noticed vertically displaced sources and performed particularly poorly in elevation accuracy at heights of 45° and -45°.

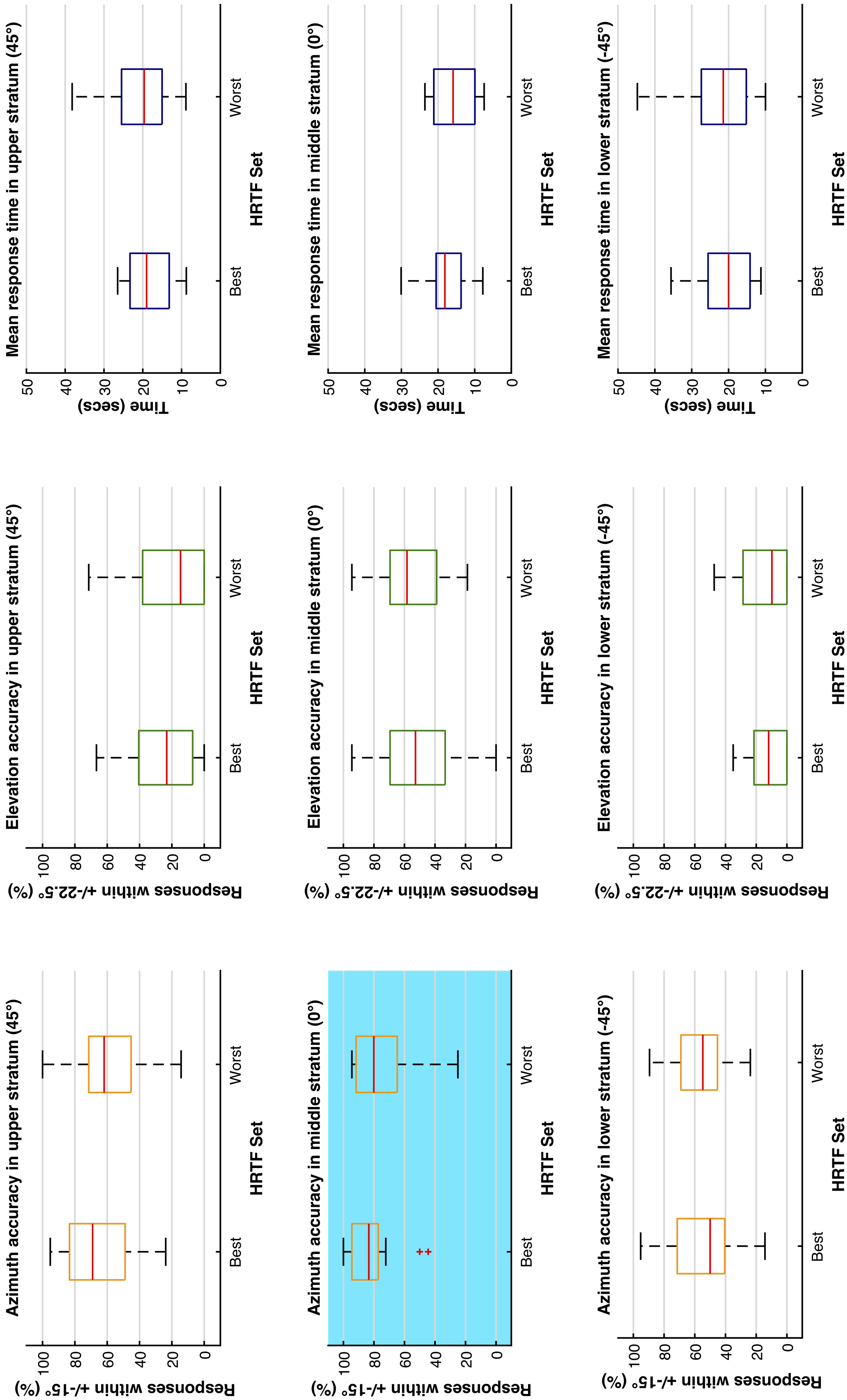


FIGURE 5.13: Distributions of participant azimuth/elevation localisation success rates and mean response times, by HRTF preference. Box plots indicate the 25th-75th percentile range and median value. Shaded plots show significant differences.

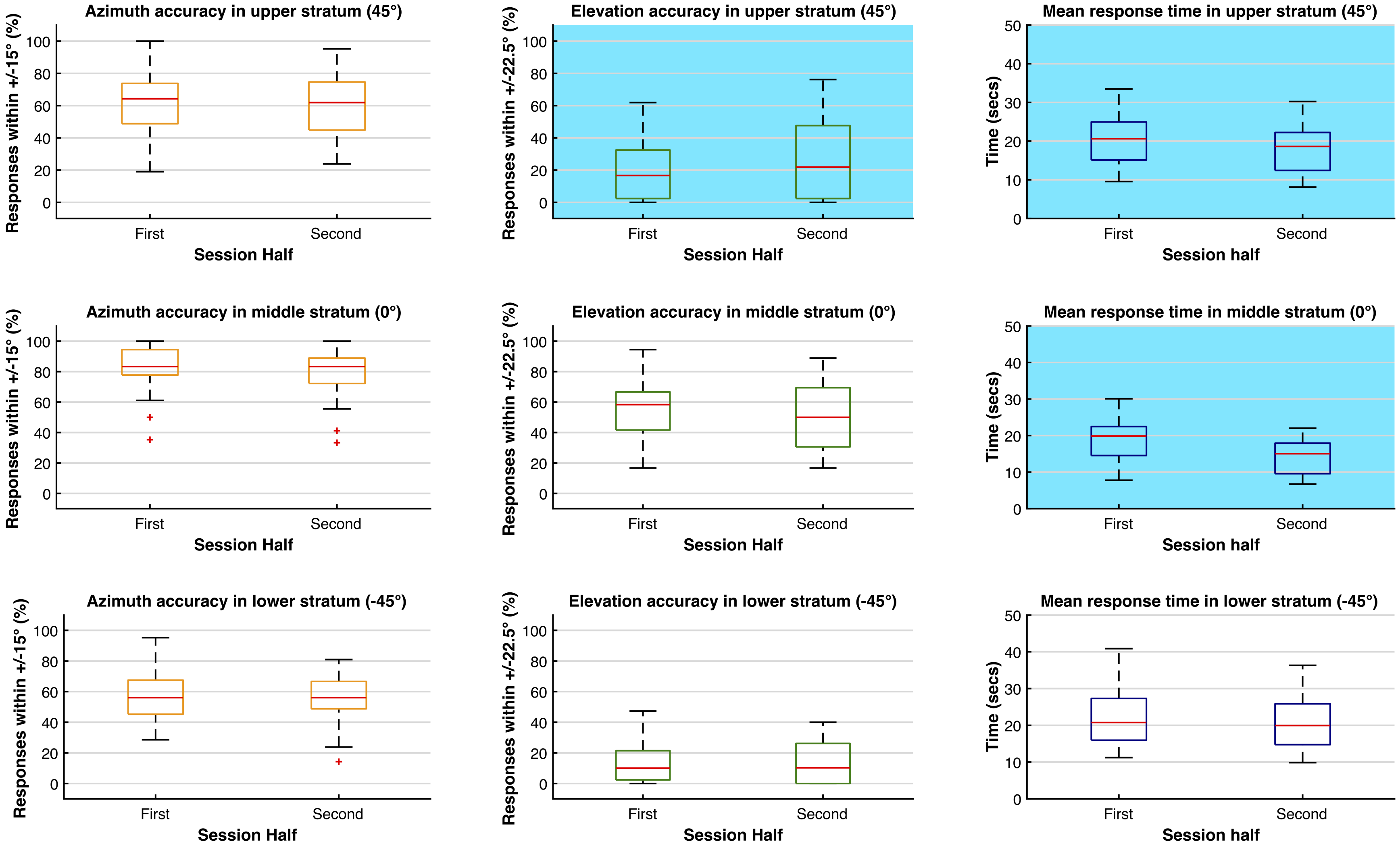


FIGURE 5.14: Distributions of participant azimuth/elevation localisation success rates and mean response times, by sequence. Box plots indicate the 25th-75th percentile range and median value. Shaded plots show significant differences.

Error margins for interactive system design

The three plots in the second column of Figure 5.13 reflect quite clearly that elevation perception was inconsistent whether or not a preferred HRTF set was being used. Even though there was significant improvement in participants' perception of positively elevated sources over time (top centre plot in Figure 5.14), for a majority of participants this still remained below chance level of 33%. Source elevation cannot, therefore, be relied upon as a parameter with specific interaction outcomes by any spatial VAD interface implemented using the virtual VBAP system as it stands. Even the most fundamental judgement of whether an audio source is above, below or at eye level is not perceived reliably.

However, elevation remains a parameter of importance that still requires consideration. When head-tracking is in operation the virtual speakers remain fixed relative to the headset, but the sound sources are counter-rotated against head movement. This means that, when a listener's head is not level, even sounds with a source position of 0° elevation will be rendered with a vertically altered location, within the context of the virtual speaker array. I.e., if a listener's head is pointed downward the source is elevated upwards within the virtual loudspeaker layout, and vice-versa. Since it is uncommon and physically uncomfortable to listen with one's head tilted backwards for any duration, the primary consideration is accuracy of azimuth perception when sources are elevated (i.e. head titled downwards).

Figure 5.12 gives a clearer picture of the tolerance that may need to be considered for a 3D interactive system founded on selection of sound source azimuth. When all participant responses are taken in aggregate, the tolerance of +/-15° used to evaluate the effects of HRTF selection is too stringent and would seem to result in interaction error rate of between 18% (on the middle stratum) to 37% (on the upper stratum). However, a tolerance of +/-20° reduces prima facie errors to between 11% (on the middle stratum) and 24% (on the upper stratum). It is also important to note that user interaction precision is likely to differ in a real world use case, compared to the laboratory test environment. In the latter case each participant was making 40 localisation responses one after another, using an identical stimulus and then repeating the same task two further times on separate occasions. Motivation and fatigue are factors that could have impacted the precision with which some responses were made in the listening test environment.

Given the results in Figure 5.12 and this associated analysis, a potential user error of +/-20° must be assumed for head-directed selection of audio targets using the virtual VBAP system. This requirement dictates that at least 41° of lateral separation be provided between sources for any VAD design prototyping, because the +/-20° tolerances for individual targets must not intersect.

5.2.4 Summary of Study 3b

Following the 3D HRTF selection study, the same participants tested its impact on their capacity to locate a continuous musical sound rendered in varying 3D positions. Patterns of interaction show a significant benefit to horizontal precision that results from the selection process. In contrast, length of system exposure (rather than HRTF preference) demonstrates a significant degree of improvement to aspects of vertical perception and overall speed of response, with no detriment to horizontal accuracy. However, despite significant improvement over time, elevation accuracy did not reach a reliable level of detectability amongst participants. These findings provided a general basis from which to consider the design of interactive audio-only immersive applications using the virtual VBAP system, which should be founded on horizontally located interactions only. More specific analysis of azimuth error margins established a recommended tolerance of +/-20° for a 3D interactive system prototype using the rendering system in its existing form.

5.3 Chapter summary

The objective of this chapter was to advance the case for using virtual VBAP for interactive binaural VAD purposes. Section 5.1 demonstrated the feasibility of implementing pseudo-spherical 3D virtual VBAP on an embedded computing platform. The target device has processing capabilities comparable to those potentially available for audio handling on a modern smartphone. Comparative technical measurement showed clear advantages against FOA in horizontal localisation across the 3D rendering environment, but with particular gains at frontal locations that would be especially important for an interactive use case. There was some evidence of reduced accuracy in representation of spectral cues compared to FOA. However, these shortcomings were more prominent at lower frequencies, where FOA tended to show greater asymmetry in error between left and right binaural channels. Therefore it is at best uncertain whether FOA would be likely to recreate any more faithfully the spectral cues that determine apparent elevation, sense of externalisation, or front/back and up/down discrimination.

By undertaking the subsequent perceptual evaluation presented in section 5.2, it was possible to obtain an understanding of the system's perceptual limitations and thresholds of localisation precision in practice. It is evident that no aspect of interaction design using the current virtual VBAP system can be contingent on even the most rudimentary distinction in vertical placement (i.e. at, 45° above and 45° below ear level). Furthermore, a tolerance of +/-20° separation has been clearly established as the minimum required for lateral interaction with sound sources. The same testing process was used to verify the validity of the 3D HRTF selection procedure, which showed the significant benefit to horizontal localisation accuracy already reported in section 4.2.

Overall, the work in Chapters 4 and 5 supports the use of virtual VBAP and an accompanying personalisation process for interactive binaural VAD design. The next chapter outlines how the system was used to create and evaluate a prototype binaural playlist editor. Up to this point, no reverberation was applied in the development or testing process. However, potential integration of virtual VBAP with scattering delay network (SDN) synthetic reverberation was mooted in section 2.3. Following the work described in this chapter, SDN reverb was incorporated into the virtual VBAP system and evaluated via research conducted in collaboration and separately from this thesis (Yeoward et al., 2021). Further work on integrating SDN reverb for the prototype binaural playlist editor, to create a fully immersive binaural experience, is described in section 6.1.1.

Chapter 6

Prototype binaural playlist editor

> Back to life, back to the present time.
> Back from a fantasy, yeah!
> Tell me now, take the initiative.
> I'll leave it in your hands until you're ready.
>
> 'Back to Life'
> *Soul II Soul*

Chapter 1 laid out the rapidly and substantially transformed landscape of music consumption practices that have emerged over the past decade. This posited that, in the age of mobile digital distribution, immersive binaural audio holds potential to augment music discovery and recommendation. It may be possible to direct innovations in personalised spatial sound towards designing augmented experiences in which listeners are active arbiters in their engagement with recorded music, rather than purely passive recipients of algorithmically aggregated offerings. Chapters 3 to 5 investigated and successfully addressed some of the engineering and design considerations that are key to implementing a viable, immersive, audio-only environment for music exploration. This included, in section 5.1, direct objective comparison of the proposed virtual auditory display (VAD) rendering system against a standard approach used widely in commercial immersive media applications. Subsequent perceptual evaluation with human participants then established design considerations for the specific use case of a mobile, audio-only music browsing interface deploying 3D VAD - a novel application that does not present comparable alternatives. To complete the research, this final chapter describes and evaluates a binaural VAD prototype for rapid, immersed exploration and customisation of music playlists against an equivalent graphical interface.

6.1 System design

Chapter 3 presented the design and evaluation of two different applications of binaural spatial audio for interacting with music content.

Auditory Archive Explorer (AAE) addressed the problem of how spatial auditory display could be applied to support or enhance voice controlled exploration of defined catalogues. No inherent practical advantages towards aiding orientation were found in spatialising the navigation structure of content in that instance; though voice interaction and the short exposure period may have acted as limiting factors on the fluency of participants' engagement. However, accompanying qualitative data suggested that the binaural prototype provoked an evidently different type of interactive auditory experience in the minds of those participants.

A second prototype system investigated the limitations of using concurrently presented sound sources in a binaural music search and browsing environment. Findings from that enquiry established three key working principles for prototype system development:

- a maximum of three concurrent sources should be presented to the user at any given time
- accompanying visual cues to aid localisation and discrimination between simultaneous sound streams are not necessary or advantageous – an audio-only approach seems just as or potentially even more effective
- musical similarity needs to be defined as comprehensively as possible, so that sonic contrast between concurrently presented streams can be maximised

The prototype described in this chapter builds directly on each of the findings from these two studies. To evaluate this design, an alternate use case is adopted that differs from the previously examined contexts of exploring a limited catalogue of content (as in the AAE study), and targeted search or browsing of a returned shortlist (used for the concurrent presentation study). Instead, this experimental scenario centres on interaction with playlists of recommended content, which are a commonly found feature in contemporary digital music distribution services. The prototype and user study described in subsequent sections were devised to address two research questions:

1. *How successfully can a binaural virtual auditory environment be used for personalising playlists of unfamiliar content?*
2. *How does use of a binaural virtual auditory environment for this purpose differ from interaction with standard graphical media playing software?*

6.1.1 Binaural virtual auditory display (BinVAD) prototype system

The mobile 3D rendering system described in chapter 5 was redeployed as the experimental platform for this inquiry. The configuration used to generate a binaural virtual auditory display (BinVAD) interface for interactive music browsing and playlist editing is described in detail below. A link to the source code repository is provided in Appendix A.2.

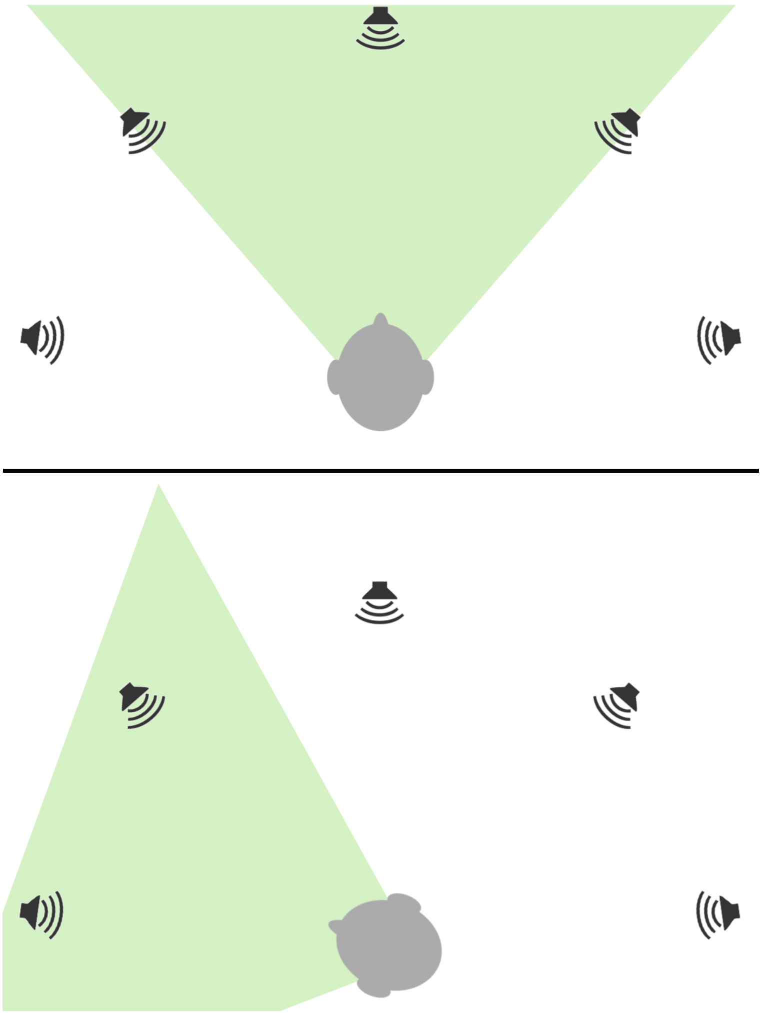


FIGURE 6.1: Example bird's eye representations of BinVAD orientations with three (top) and two (bottom) concurrent sources audible

Binaural display interface

The primary design objective behind BinVAD was to enable auditioning of multiple concurrent streams of music as rapidly and intelligibly as possible. Presentation of five sound sources, or options, within an auditory browsing scene was a working principle first established by Hiipakka and Lorho (2003) and Lorho, Hiipakka and Marila (2002), and explored further in the AAE system outlined in chapter 3. In the BinVAD prototype a maximum of three sources can be heard at any one time, but five music tracks are always present in the virtual auditory environment. Each of the five tracks runs continuously on an independent loop – irrespective of whether or not they are currently being monitored by the user. The sources are spatialised binaurally and can be imagined as five monophonic speakers in a physical space, each playing a different track, arranged in front of the listener at ear level (0° elevation) and at azimuth angles of -82°, -41°, 0°,

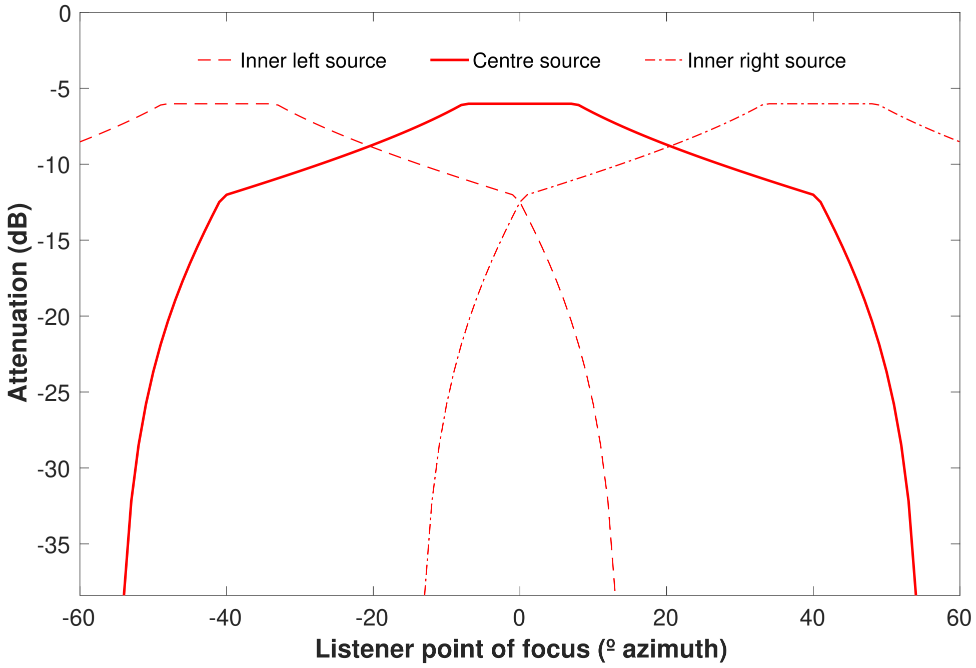
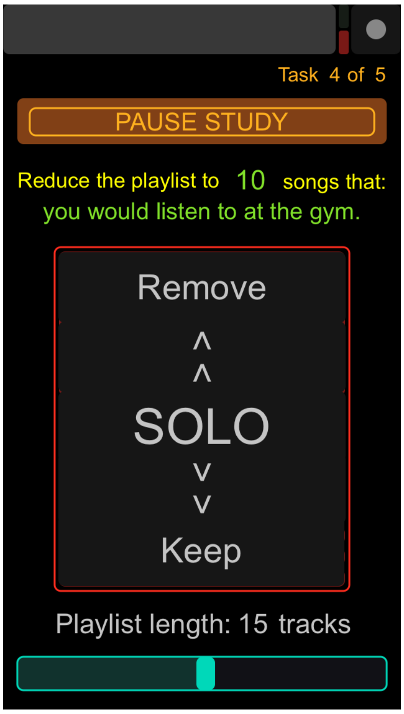


FIGURE 6.2: BinVAD attenuation curve shown for the central sound source, with illustrative crossovers for the inner left and right sources only

41° and 82°. This separation is based on the 41° lateral separation requirement identified in section 5.2.3 to achieve adequate tolerance for user selection accuracy. Horizontal head movement directs an imaginary beam of auditory focus across the available pieces, such that either three or two adjacent pieces can be previewed at any given moment and forward-facing head orientation (as illustrated in Figure 6.1).

To smooth transition and aid auditory discrimination between tracks, a specially devised attenuation curve is applied according to the listener's point of visual attention (derived from real-time tracking of head orientation) relative to the source. Figure 6.2 shows the gain reduction applied to a source (in this case, the track placed in the centre position of the five options) when listener focus is within +/- 60° azimuth of its location. Volume of the source is stable so long as it remains within +/-7° of the listener's point of focus. As the listener directs their attention further away, the amplitude of the source is reduced – in addition to the binaural spatialisation already applied by the main virtual VBAP algorithm. Volume reduction is initially applied at an even logarithmic rate between locations at +/-8° and +/-40°. The source then attenuates steeply to silence as it traverses locations between +/-41° and +/-55° from the listener's point of focal attention. The source remains silent at any locations beyond this range.

Relevant portions of the inner left and inner right attenuation curves are superimposed in Figure 6.2, to illustrate the crossover with adjacent sources. Identical and evenly spaced curves were also applied to the outer left and outer right sources, but for clarity are not represented in Figure 6.2. Because sources are spaced at intervals of 41°, each



Task 4 of 5

Reduce the playlist to 10 songs that:
you would listen to at the gym.

Playlist length: 15 tracks

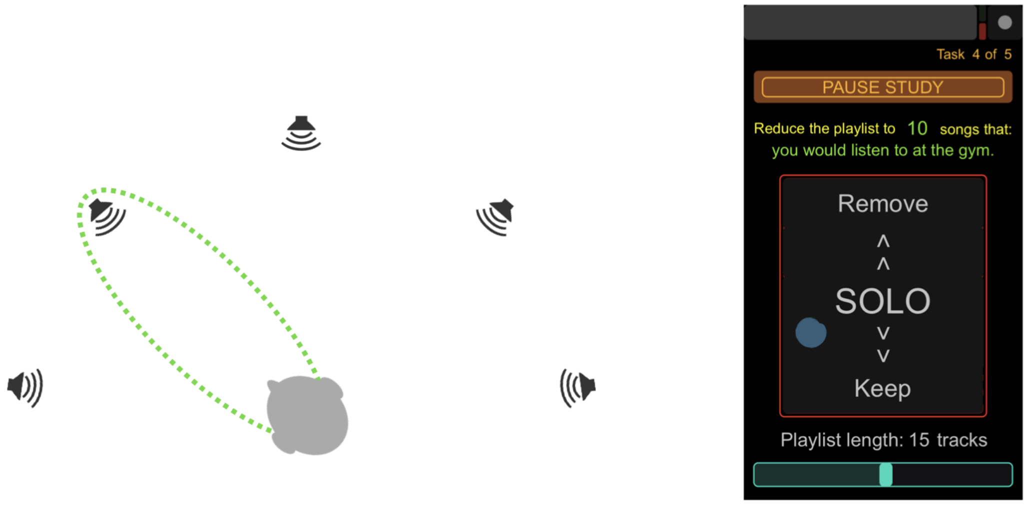
FIGURE 6.3: Eyes-free gestural interface for the BinVAD auditory environment. (Text supports the research study construct only.)

one's active range overlaps with those of its two neighbours (in this case, the inner left and inner right sources). This enables the listener to place their auditory attention between either three or two sources. When focussing precisely on the centre source (as depicted in the example of Figure 6.1 top) its playback volume is at -6dB, the level of those on the inner left/right is at -12.5dB and the outer left/right are entirely silent. When focussing directly in between two sources, (as depicted in Figure 6.1 bottom) the playback volume of both is at -8.8dB and the three other sources are silent.

To assist with targeting and orientation of focal attention, a single click sound effect is triggered whenever the user transitions into the stable level zone of a sound source (i.e. within +/-7°) .

Mobile input interface

The five active sources represent only a portion of the available tracks in a playlist. Access to and editing of the wider playlist is enabled via the simple touch screen interface shown in Figure 6.3. This was again developed on a smartphone for ease of prototyping and featured accompanying text information to support participant progress through the research study construct. In reality, the gestural commands are designed to be suitable for eyes-free use on a smartwatch or other compact, wearable, touch surface interface that may not provide any visual feedback. On this occasion, unlike for Study 3a and 3b, the



Reduce the playlist to 10 songs that: you would listen to at the gym.

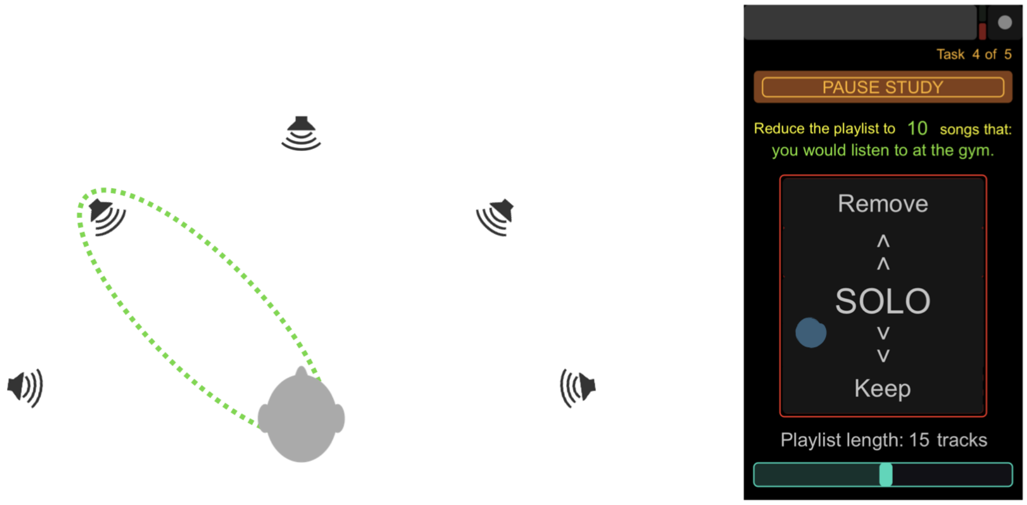
Playlist length: 15 tracks

FIGURE 6.4: Illustration of BinVAD "SOLO" mode activation. Touching and holding anywhere in the red bounded area immediately and seamlessly switches the scene to play back only the track closest to the listener's point of visual attention. This persists until touch contact is released.

interface is used by the participant to manipulate the auditory scene, as well as submit their responses and progress through the tasks.

All three gestures for interacting with playlist content are conducted within the rectangular area bordered red, which acts a single, continuous touch surface interface. The text "Remove", "SOLO" and "Keep" and the vertical chevrons do not represent specific points of interaction, but serve as a reminder to study participants of the available actions. Making continuous contact anywhere within the rectangular area enables solo mode, where the track closest to the listener's point of visual attention is played back in isolation, as illustrated in Figure 6.4.

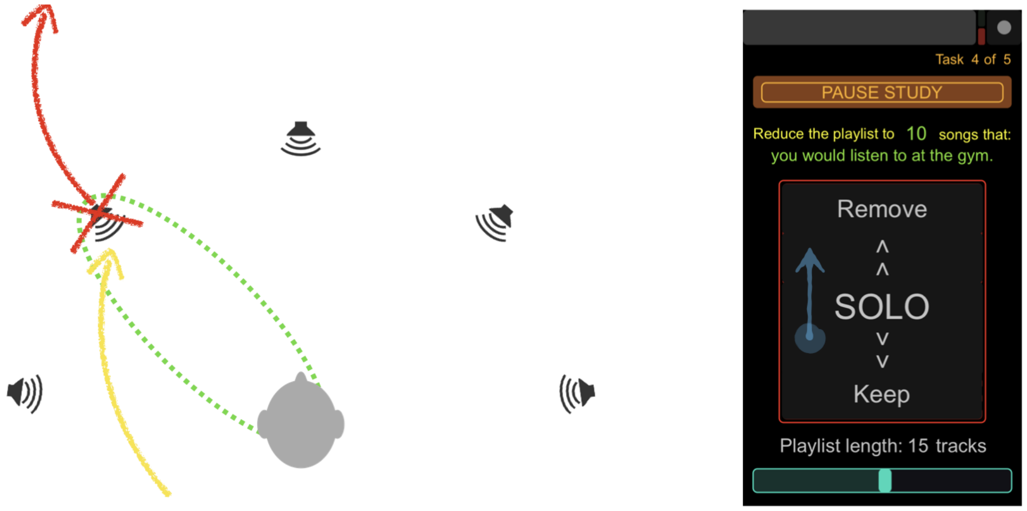
Solo mode remains active until touch contact is released from the interface. When entering solo mode, a spoken word announcement of the selected track title and artist is heard once, whilst the music plays seamlessly in parallel and without interruption from its progress position at the transition from full browsing mode. The announcer's voice occurs at the same lateral position as the corresponding track, but elevated by 30° to aid auditory segregation. This elevation angle matches that used for the same purpose in the design of AAE, in which subjective informal testing determined it sufficient to audibly separate the voiceover from associated music, but without pushing the speech source too far overhead the listener. Whilst solo mode is engaged (i.e. during continuous contact with the interface area), only the active selection is heard, fixed to its virtual location and playing at a constant volume, irrespective of head movement. This enables users to solo a track in a lateral location with required head position, but return to any preferred physical orientation for auditioning the piece in a relaxed posture, as shown in Figure 6.5



Reduce the playlist to 10 songs that: you would listen to at the gym.

Playlist length: 15 tracks

FIGURE 6.5: Illustration of BinVAD "SOLO" mode with updated head position. Playback of a soloed track persists in its virtual location and at a constant volume, irrespective of any head movement or orientation, so long as touch contact is continued on the interface.



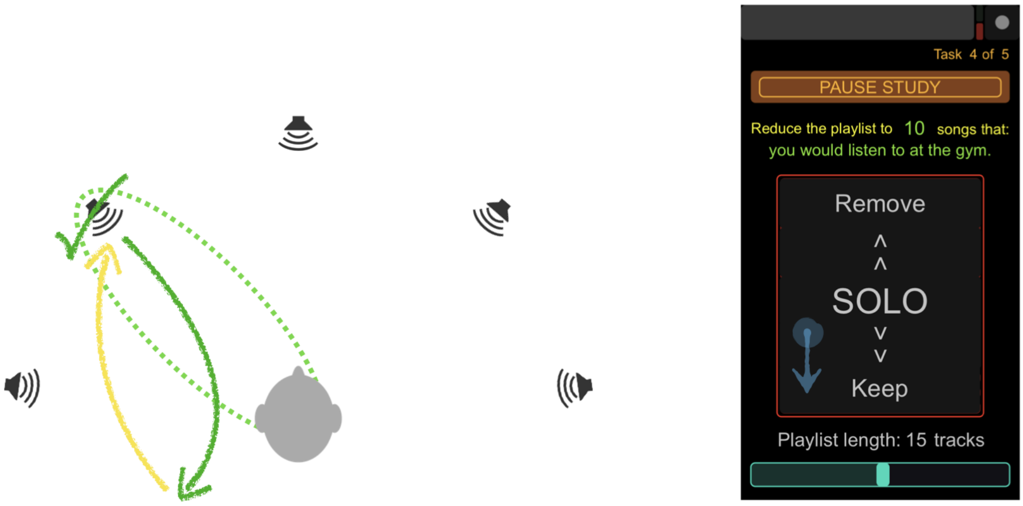
Reduce the playlist to 10 songs that: you would listen to at the gym.

Playlist length: 15 tracks

FIGURE 6.6: Illustration of the "Remove" gesture in BinVAD. Swiping up on the bounded area removes the currently selected track from the auditory scene and from the playlist. The track is discarded at this point and cannot be retrieved.

Any of the five active tracks can be removed or kept, and supplanted by the next item in the playlist, which is a longer list comprising any number of tracks. A standard touch screen up swipe (smooth upward stroke followed by immediate release) removes the currently selected track, as shown in Figure 6.6. When a track is removed it is discarded from the active scene of five sources and also deleted permanently from the playlist.

Conversely, a standard touch screen down swipe (smooth downward stroke followed by immediate release) keeps the currently selected track, as shown in figure 6.7. In this instance, the track is discarded from the active scene of five sources, but it is retained in



Reduce the playlist to 10 songs that: you would listen to at the gym.

FIGURE 6.7: Illustration of the "Keep" gesture in BinVAD. Swiping down on the bounded area removes the currently selected track the auditory scene, but retains it in the playlist. The track is is kept for later retrieval if and when the user cycles all the way through remaining content.

the playlist. Kept tracks will eventually reappear if the playlist is cycled through in its entirety. Any tracks that remain in a playlist can be revisited when the list is looped over from the beginning.

Three further confirmatory audio notifications are used to support eyes-free use of the mobile input interface and orientation with the playlist:

- a waste paper bin sound plays immediately to confirm successful removal of a track before the next item in the queue starts playing in its place
- an electronic buzzer sound operates in the same way, but to confirm that a track has been successfully kept
- a glass chime sounds three times to indicate when the user has reached the final entry in the playlist

Further text and interaction features on the interface are present for the research study itself. Text information displays the current task description, number and playlist length, which updates automatically on successful removal of a track. A 'PAUSE STUDY' / 'RESUME STUDY' button and volume control allows participants to manage their progression through the tasks.

Binaural rendering

The binaural rendering used to implement BinVAD follows the implementation described in Chapter 5, with three modifications to accommodate the far greater processing requirements of this more complicated interactive system.

1. Following completion of the work in this Chapter 5, a separate investigation was carried out through supervision of a Masters degree researcher, Christopher Yeoward, which incorporated SDN reverb into the virtual VBAP system (as suggested in section 2.3). The work examined methods for reducing the computational complexity of the SDN algorithm and consequent impact on reported 'envelopment' and 'naturalness'. One simplification explored a *distributed mono* spatialisation of SDN, which calculated all six surface reflections according to the original algorithm, but summed these streams and rendered the resulting signal equally to all eight virtual loudspeakers. This technique avoided continual vector rotation of six independent reverberation channels that would otherwise be necessary with virtual VBAP. Published results from perceptual evaluation identify the effect as a perceptually optimal simplification for achieving sense of naturalness in small virtual spaces (Yeoward et al., 2021). This is consistent with other findings indicating that use of spatially simplified reverberation in binaural synthesis may be judged more favourably than implementations with a higher degree of fidelity (Picinali et al., 2017). The *distributed mono* SDN approach was adopted and has been simplified further for the final study prototype, as follows:

    - Four surfaces are modelled in the SDN algorithm (i.e. only the signals generated by wall reflection nodes are computed and reproduced, not those for floor and ceiling nodes). The reverb algorithm is therefore implemented in 2D (without height), but the sound sources themselves are positioned in 3D.
    - These four reflection signals are summed to mono and sent equally to all eight virtual loudspeaker positions, rather than being directed and counter-rotated as independent streams in the virtual VBAP algorithm. Reflections are therefore calculated according to true physically modelled room dimensions, but are rendered in a 'immersive mono' form rather than being discretely spatialised.
    - A single scattering network generates reverb for all sound sources. This means that, although each dry source signal was spatialised and counter-rotated faithfully to its designated position, the acoustic reflections simulated for all sources are computed using a single virtual point of origin (which is at ear level and 0° azimuth).

2. The length of the HRTF set has been shortened from 512 to 256 samples, with a preserved sampling rate of 44.1 kHz, to reduce the computational effort and time required for achieving binaural synthesis via multiplication in the frequency domain. Although HRTFs from the LISTEN database are available by default at a length of 512 samples and rate of 44.1 kHz, use of shorter HRTFs to conserve computational effort are common practice. For example, HRTFs in the CIPIC database are available only at a length of 200 samples and rate of 44.1 kHz (Algazi et al., 2001).

The SADIE database offers HRTF sets at both 256 samples (either at 44.1kHz/16 bit resolution, or 48 kHz/24 bit resolution) and 512 samples (but at 96 kHz/24 bit resolution) Armstrong et al., 2018. The revised duration is therefore in line with the SADIE formats and exceeds CIPIC. Nevertheless, the reduction truncates later reflections from the subject's head, upper body and ear that are embedded in the HRIRs and contribute to perception of elevation, externalisation and hemispherical direction (as discussed in section 2.1.

3. Further system buffering of 512 samples has been introduced to enable all synthetic reverb and interactive processing deadlines to be met reliably. This results in a total end-to-end latency of 70 ms (3093 samples at 44.1 kHz sample rate), which remains within the target 75 ms response time advocated by Suzuki, Yairi and Iwaya (2007) and comprises:

    - 512 samples of system buffering
    - 2048 samples of input signal buffering to handle Fourier transformation and convolution processes
    - 512 samples head-tracking refresh rate
    - 21 samples of known latency at the Bela digital-to-analogue converter

A likely consequence of these modifications is an unknown degree of reduction in spatial fidelity. Better understanding of their impact would require direct measurement of source locations' perceived accuracy, or assessment of the virtual scene's acoustic plausibility. Instead, for the needs of this enquiry, the resulting binaural effect has been assessed and tailored through subjective informal testing during development. This verified that the required level of spatial clarity, responsiveness and immersion resulted. With the above limitations consciously built-in, the final BinVAD system is configured to simulate physical acoustics for a listener positioned in the centre of a small rectangular room, as detailed in Figure 6.8.

All four system sound effects (i.e. the auditory focus transition click, the 'remove' and 'keep' action confirmation sounds, and the playlist cycle notification) are rendered without any spatial effect. This segregates them from the BinVAD scene to create a subset of sonic information about the user's own interactions that is abstracted from the virtual environment itself. This approach is similar to the method deployed for AAE and described in section 3.1.3.

HRTF personalisation

A further compromise in the rendering chain was included for the purpose of this study, which faced necessary restrictions governing its completion during the COVID-19 pandemic. These factors are discussed further in section 6.2.4. One outcome was that it became logistically unfeasible to incorporate personalised binaural rendering via the HRTF

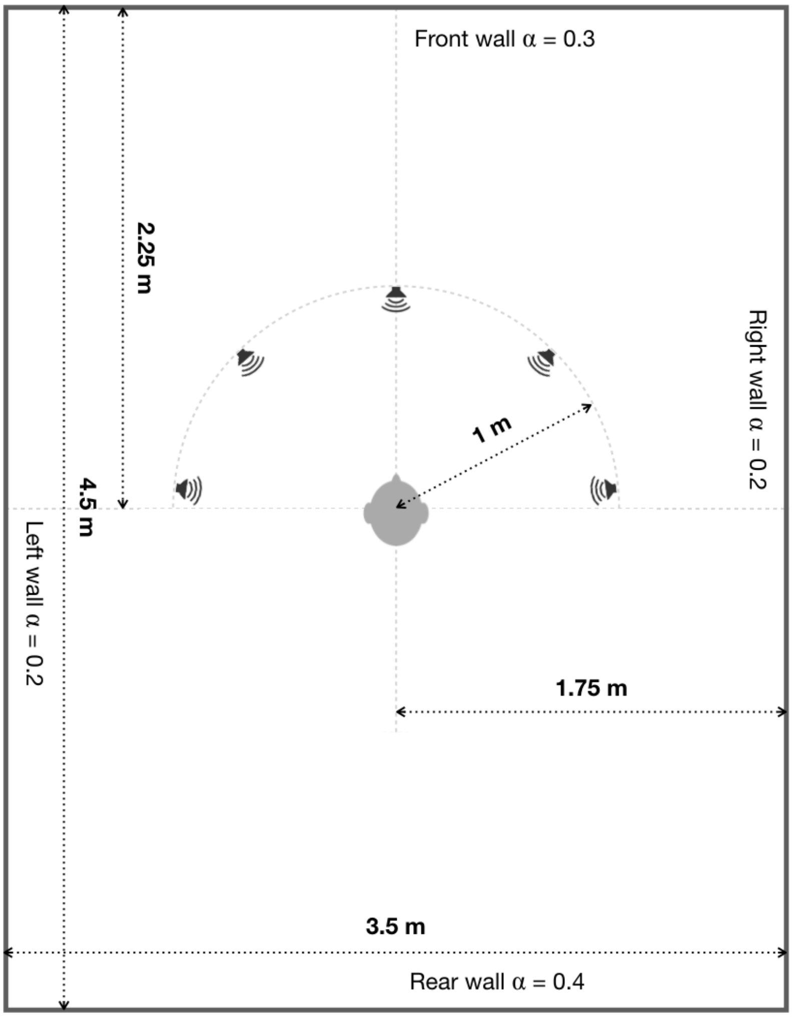
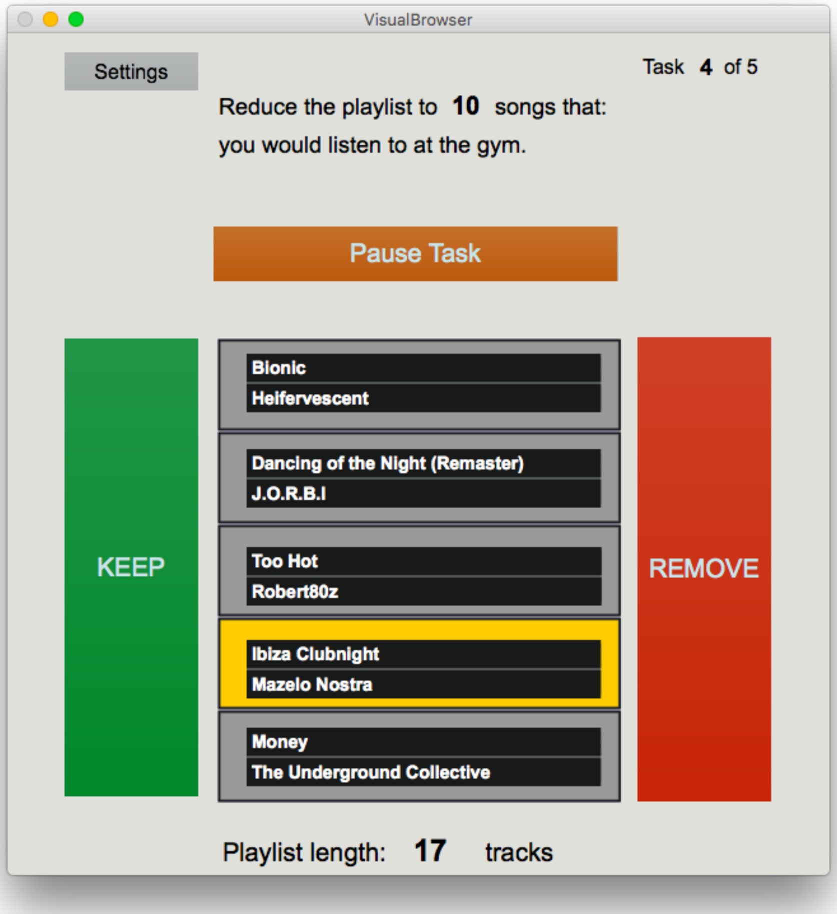


FIGURE 6.8: Virtual acoustic design of the BinVAD environment. (α = frequency independent absorption coefficient (Sabine, 1922)
.

selection routine developed and evaluated in section 4.2. Instead, the most frequently selected HRTF from that research study, *LISTEN 1031*, was used for all participants.

Use of a generic HRTF means that the quality of the binaural image experienced by participants is likely to have varied more between participants. Indeed, the data in section 4 strongly indicates that lateral localisation was improved when participants used a preferred HRTF set. Although *LISTEN 1031* was the set most often judged as preferable in pairwise comparisons by 21 participants in the conditions of that study, its performance in this context with a different group was not evaluated. For this study, any reduction in lateral fidelity from a generic HRTF might consequently impact perceived spatial segregation between sources. It is therefore possible that some participants would have found the concurrent browsing scene more challenging to process than they might with a preferred HRTF set. In the event, two participants reported a particular degree of confusion and even discomfort in navigating the binaural scene – participants L and Q,



Task 4 of 5

Reduce the playlist to 10 songs that:
you would listen to at the gym.

Bionic
Helfervescent

Dancing of the Night (Remaster)
J.O.R.B.I

Too Hot
Robert80z

Ibiza Clubnight
Mazelo Nostra

Money
The Underground Collective

Playlist length: 17 tracks

FIGURE 6.9: GUI developed as a reference system for evaluation of BinVAD

as per observation 3 in Table 6.3.

6.1.2 Graphical user interface (GUI) reference system

The graphical user interface (GUI) illustrated in Figure 6.9 was developed as a secondary system against which to evaluate aspects of participants' interaction with BinVAD. It was developed using Max/MSP and created as a standalone application for Windows and macOS. The design principle behind its development was to mimic the features of a typical visual interface for music playlists, but to mirror the constraints of the BinVAD system for comparison purposes.

Like with BinVAD, the GUI offers five tracks that play back simultaneously, each on a continuous loop. However, only one track is heard at any time as a standard stereo signal. There is no binaural spatial effect applied anywhere in the system's audio playback. A standard mouse click on any of the five grey track/artist blocks selects that option (in yellow highlight) and playback switches immediately to the current progress position of the chosen track. There are no voiceover announcements, but the equivalent information is provided in the track title and artist text. The 'remove' and 'keep' functions are implemented with large clickable buttons that edit the current playlist with the same effect as for the BinVAD system (as shown in Figures 6.6 and 6.7).

6.2 Study Design

To evaluate BinVAD, participants were recruited to complete a series of playlist editing tasks using both systems. The definition of the tasks, curation of content, participant recruitment, and execution of study sessions during the COVID-19 pandemic are each explained in the following subsections.

6.2.1 Study session format

The study took a total of 1h 45m, split across two separate sessions. Each participant's sessions were arranged at least a week apart to reduce the likelihood of order effects in their responses. To mitigate further, half of the participants completed part A first and the other half started with part B. Both parts followed an identical format that was tailored to the given system, other than a short verbal discussion that was added at the end of the BinVAD session only:

- **Part A:** music selection tasks using the GUI system (approximately 45 minutes)

    1. software setup
    2. overview and interface instruction
    3. task explanation
    4. example task completion and follow-up queries
    5. study tasks 1-4

- **Part B:** music selection tasks using the BinVAD system (approximately 1 hour)

    1. hardware setup
    2. overview and interface instruction
    3. task explanation
    4. example task completion and follow-up queries
    5. study tasks 1-4
    6. short qualitative interview

Details of the software and hardware setup stages are provided in section 6.2.4. The overview and instruction stage presented a summary of the information provided here

Task	Length	Task selection criteria	Target	Qualifying tracks
Example	10	*... have a fast tempo.*	5	*7*
1	10	*... you like.*	5	–
2	15	*.. have a strong beat.*	8	*9*
3	20	*... you would listen to at the gym.*	10	*13*
4	30	*... you would listen to on a commute.*	15	–

TABLE 6.1: Task definitions for parts A and B of the BinVAD evaluation study sessions

in relation to Figures 6.1, 6.3–6.7 and 6.9. An identical script was used to explain the format and aim of tasks for both part A and B, as follows:

> ***The Tasks***
>
> *There are five tasks. The first one is an example that we will use to get you used to working in the environment. Each task describes a specific listening scenario. For each task you are asked to reduce a long list of songs to half its original length. The starting length of the list increases with each task. There are 10 tracks in the first task and 30 tracks in the final task.*
>
> ***The Aim***
>
> *With each task, you are asked to provide the best response you can to satisfy the request. How you make your decisions on which tracks to keep and which to remove is entirely up to you. You can work at whatever speed you need to provide your response. For each task, there is no one correct answer.*

The criteria for the example and four subsequent study tasks of both part A and B are specified in Table 6.1. Criteria from the exploratory concurrent browsing study were reused, since a constructive pattern of engagement had been observed in that instance (section 3.2.4). On this occasion, participant responses to *task 2* and *task 3* were to be assessed for accuracy, since these selection criteria (and that of the example) could be regarded as somewhat objective, whereas *task 1* and *4* were wholly subjective. Recruitment prerequisites for the study are described later in section 6.2.3. It was considered a reasonable expectation that this profile of participants would be able to identify tracks with a more prominent pulse (in the case of *task 2*), or which have a suitable degree of energy and impetus to encourage physical activity (for more *task 3*). The number of tracks qualifying as 'correct' responses included in the playlists are shown in the final column. The specific means by which each playlist and any qualifying tracks were curated is explained further in section 6.2.2.

The two open questions posed at the end of the BinVAD session were:

> *Can you describe your experience of interacting with the auditory environment?*

Can you describe any strategies you used to make decisions when reducing the playlist?

These questions were designed to gather additional qualitative information to enlighten patterns of interaction that might emerge through quantitative analysis.

6.2.2 Content curation

Careful selection of music content for each task was an important aspect of the study design.

Playlist criteria

The makeup of task playlists needed to satisfy the following four criteria:

1. minimise likelihood of familiarity bias whilst presenting an experience akin to editing playlists of commercial music;
2. populate the BinVAD and GUI tasks with playlists that were totally different, but which were demonstrably comparable in their makeup;
3. establish a degree of ground truth against which to evaluate participant responses for the *example task*, *task 2* and *task 3*;
4. facilitate the maximum possible musical contrast between concurrently presented pieces in the BinVAD system, as discussed in section 3.?.

Choice of content library

Jamendo[1] is an online music distribution service for unsigned artists to promote and/or monetise their work. It offers free streaming and download for personal consumption, or a licensing model for creators who wish to seek financial return. Although contributors range from hobbyists to full-time professional musicians, the service's promotion algorithms tend to surface material that is of higher, near-commercial standards of production. A proprietary API (application programming interface)[2] allows querying of the *Jamendo* catalogue to retrieve content according to popularity or contributor tagged metadata, such as genre, vocalist gender, language, instrumentation, etc. As such, it offers a library of music that satisfies criteria 1 and 2 by providing unfamiliar but convincing substitutes for mainstream releases, whose surface features can be filtered directly.

[1] www.jamendo.com [accessed 30/11/2021].

[2] developer.jamendo.com/v3.0/docs [accessed 30/11/2021].

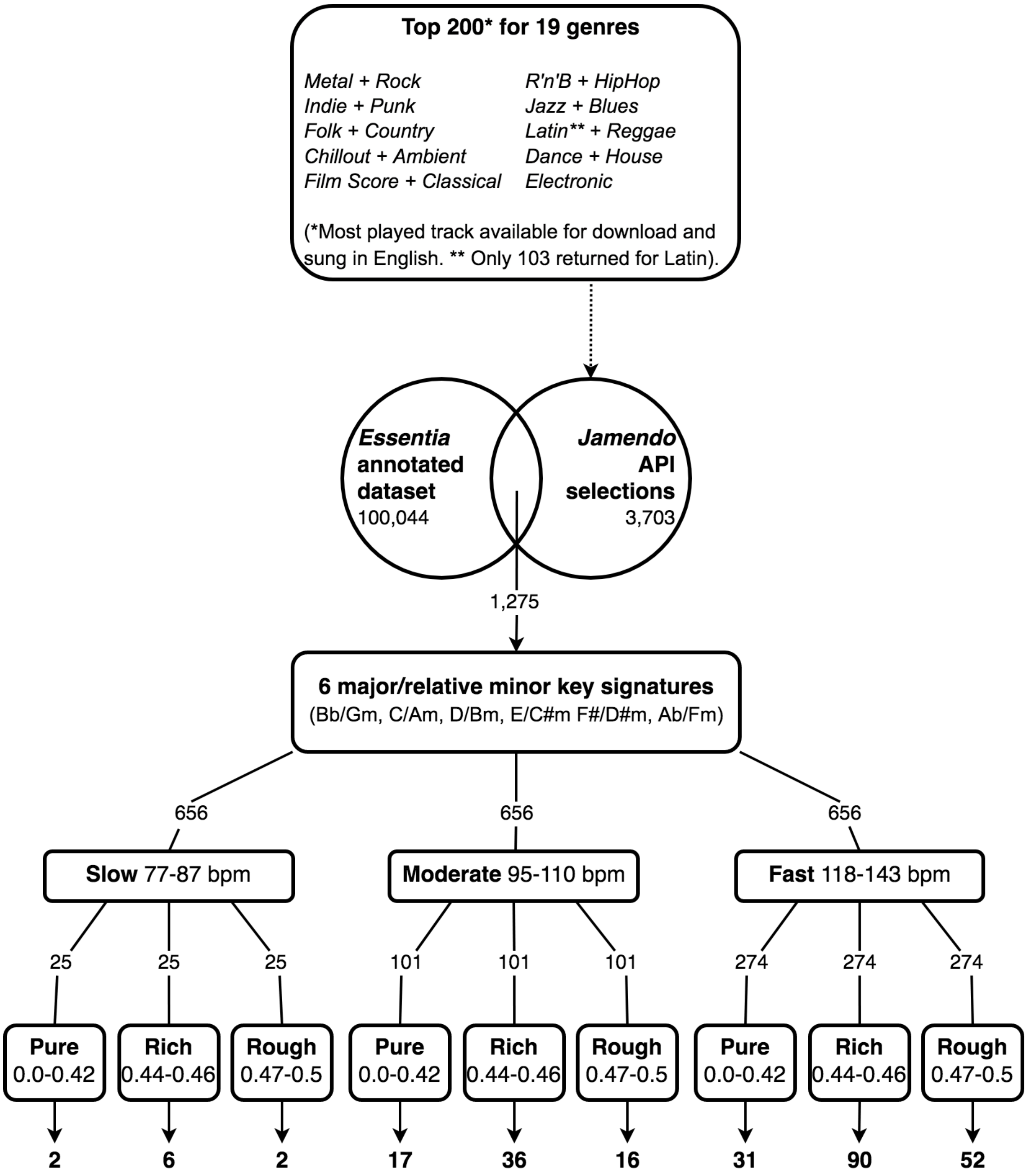


FIGURE 6.10: Flow diagram of the *Jamendo* content curation. Flow path values show the number of tracks returned after each filtering stage.

The *Jamendo* dataset has also been enhanced by other researchers in ways that make it adaptable towards fulfilling criteria 2, 3 and 4. Xambó et al. (2018) previously conducted a harmonic analysis of a large cross-section of the *Jamendo* catalogue using digital signal processing techniques. This work formed the ‘query-by-chord’ engine of an interactive tool for learning and practising new pieces of music. When creating that enhanced dataset, the researchers also performed broader content analysis using the *Essentia* open source, audio based music information retrieval library (Bogdanov et al., 2013). The result is a subset of 100,044 *Jamendo* tracks with computationally derived metrics related to a wide range of musical features such as key, tempo, timbre and beat strength.

Figure 6.10 illustrates how the *Essentia* annotated dataset and main *Jamendo* API were integrated to curate the required content. Nineteen of the twenty genres listed on *Jamendo* were queried to return the 200 most played tracks in each style available for download and sung in English (for the purpose of consistency). 'Pop', which was considered stylistically too generic, was excluded. Due to this language stipulation the result for 'Latin' only returned 103 tracks. These 3,703 tracks were cross-referenced against the *Essentia* annotated dataset, resulting in a list of 1,275. From here, a number of data points were used to filter down the selection pool to 252. This pool and its accompanying data was used to determine the makeup of the 10 study playlists, which incorporated a total of 170 tracks.

Content categorisation and filtering

Two contributor-defined parameters were used to assist in curation decisions:

- **Genre:** Due to sonic and musical similarities, and frequent co-tagging of certain stylistic categories, the list of 19 *Jamendo* genres was consolidated down to 10 groups. These consisted of nine pairs and one single style ('Electronic'), as shown in Figure 6.10. This broader set of genre groupings provided a better degree of contrast between stylistic categories.

- **Vocalist gender:** Information about the gender identity of the singer on a track was used to balance the makeup of part A and part B across each task playlist.

Three further stages of filtering (shown in Figure 6.10) were applied to segregate tracks, as far as possible, into contrasting classifications related to different musical features defined by *Essentia* algorithms:

- **Key[3]:** Only tracks identified as having one of six key signatures were retained. The permitted keys included C major/A minor and then every other key signature occurring in the "circle of fifths" that defines the principles of harmony underpinning music written in the Western tradition (Campbell and Greated, 1987; Taylor, 1989). This resulted, for example, in the exclusion of keys G major/E minor and F major/D minor, which both share six of the seven pitches within C major/A minor. This approach ensured that no two included keys had any more than five out of seven pitches in common.

- **Tempo[4]:** Three tempo bands of slow (77-87bpm), moderate (95-110bpm) and fast (118-143bpm) were defined. These values precluded any tempos corresponding to

[3] essentia.upf.edu/reference/std_Key.html [accessed 17/6/2022]
[4] essentia.upf.edu/reference/std_RhythmExtractor.html [accessed 17/6/2022]

integer multiples of another. This ensured that tracks in different groups would always have markedly different rhythmic rates, even if perceived in double or half time by the listener. Additionally, the seven qualifying (valid answer) tracks in the *example* playlist for both part A and B had a tempo of 135bpm or higher. All other distractor (invalid answer) tracks in these playlists had a tempo of 105bpm or lower.

- **Spectral Dissonance[5]:** *Essentia* calculates spectral dissonance as a representation of the overall timbre of a track, formed by the cumulative effect of the type and tone of instrumentation used within a recording. Pure (0.0-0.42), rich (0.44-0.46) and rough (0.47-0.5) classifications were defined to ensure that tracks from different groups had a notably contrasting timbral character.

Two further parameters were used to identify qualifying tracks for participants' answers to *task 2* and *task 3*:

- **Beat Loudness[6]:** This metric indicates the mean ratio of spectral energy found in beat segments across an entire track (as a value from 0 to 1). It was used to assist in identifying tracks more likely to be songs that "have a strong beat". The nine qualifying (valid answer) tracks in the *task 2* playlist for both part A and B had a value of 0.06 or greater. The six distractor (invalid answer) tracks in these playlists had a value of 0.04 or less.

- **Danceability[7]:** This metric ranges from 0 to 3 (where higher values are more danceable) and was used to assist in identifying tracks more likely to be songs that "you would listen to at the gym". The 13 qualifying (valid answer) tracks in the *task 3* playlist for both part A and B had a value of 1.3 or greater. The seven distractor (invalid answer) tracks in these playlists had a value of 1.1 or less.

Track selection protocol

The list of 252 tracks returned after the filtering in Figure 6.10 was manually sifted and organised into the 10 required playlists (five each for part A and B), to satisfactorily address criteria 2, 3 and 4. To address criterion 2 (comparable content for part A and B), the ratio of different genres, female vocalists and tracks in a minor key between each task's two playlists was balanced as evenly as possible, as shown in Table 6.2. Because *task 1* and *task 4* were entirely subjective, particular effort was made to match the characteristics of their part A and B playlists to maximise comparability between the two. To address criterion 3 (ground truth for assessed tasks), tempo, beat loudness and danceability values were applied as described above. To address criterion 4 (contrast between concurrently presented pieces), all tracks within a playlist had a unique combination of genre, key,

[5] essentia.upf.edu/reference/std_Dissonance.html [accessed 17/6/2022]
[6] essentia.upf.edu/reference/std_BeatsLoudness.html [accessed 17/6/2022]
[7] essentia.upf.edu/reference/std_Danceability.html [accessed 17/6/2022]

Task	Part	Genre Makeup	Vocalist Ratio	Tonality Ratio
		*no. tracks in each style**	*male : female*	*major : minor*
Example	***A***	[1 0 1 0 1 ; 1 2 1 2 1]	7:3	6:4
	B	[1 2 2 1 0 ; 1 1 0 1 1]	6:4	7:3
1	***A***	[1 2 2 1 1 ; 0 1 0 1 1]	7:3	7:3
	B	[1 2 2 1 1 ; 0 1 0 1 1]	7:3	7:3
2	***A***	[2 1 0 0 1 ; 1 2 2 3 3]	10:5	9:6
	B	[2 3 0 1 1 ; 1 2 0 3 2]	10:5	9:6
3	***A***	[3 2 1 1 1 ; 1 1 2 7 1]	13:17	9:11
	B	[2 2 1 2 0 ; 1 2 4 3 3]	13:17	8:12
4	***A***	[6 3 4 2 1 ; 0 4 2 6 2]	21:9	18:12
	B	[6 3 4 2 1 ; 1 4 1 6 2]	21:9	19:11

(* See Figure 6.10 for the 10 genre category definitions.)

TABLE 6.2: Balance of playlist characteristics for parts A and B of the BinVAD evaluation study sessions

tempo or spectral dissonance categories, as defined above, which differed from that of every other item in the playlist. This manual curation of contrasting tracks was in lieu of a hypothetical automated system for real-time selection of tracks for optimal auditory distinction between concurrent sources based on the similar sets of musical data.

Finally, track and artist voice announcements for the BinVAD system were automatically generated from the macOS onboard text-to-speech converter, using the 'Serena' synthetic voice. All music tracks used in the BinVAD and GUI systems were edited to two minutes duration with a five minute fade-out, before being normalised to equal program loudness of -23 LUFS.

6.2.3 Participant recruitment and makeup

Nineteen participants (aged 23-47; eight identifying as female, ten as male, one as non-binary) were recruited on to the study, which was approved by the Queen Mary University of London (QMUL) Ethics Committee (reference 2488a). The study was conducted during the COVID-19 pandemic and was pursued entirely remotely under the protocol explained in section 6.2.4. Three of those recruited formed a pilot study to test and refine the remote format, whilst sixteen participated in the final study format for which data is presented.

It was judged that, to engage in tasks with the required degree of purpose and focus, participants had to have a demonstrable and discerning interest in a variety of musical styles. This prerequisite and the social distancing restrictions in force at the time meant that an open call for participation was neither desirable nor possible. Instead, study recruitment was first achieved by targeted invitations distributed amongst professional networks of individuals working in the fields of music and audio. For the final study this resulted in eight music/audio professionals, including: an orchestral composer, a broadcast music editor, a violin teacher/performer, a contemporary music producer, two audio software developers, and two digital music researcher/practitioners. Eight people with an amateur or personal interest in music were also recruited by targeted invitation. The suitability of amateur participants was confirmed on the basis of their response to a profiling questionnaire, which invited them them to state their level of musical engagement against three indicators. All participants declared levels of engagement that met the accepted minimum for at least two of the following three criteria:

- *musical training* - minimum of informally developed/self-taught musical skills
- *average weekly music listening* – minimum of 10 hours per week
- *styles of music actively listened to* - minimum of 25% (6 from the list of 23 stylistic categories in (Rentfrow and Gosling, 2003))

Only one participant declared a known hearing impairment that was self-defined as 'light tinnitus'. Each participant received a £15 gift voucher in compensation for their time at the end of their second session.

6.2.4 COVID protocol

The BinVAD system was based on an almost identical set of equipment to that described in section 5.2. The only difference for this study was that the headphone set was substituted for a closed-back model (KRK KNS 8400) with good isolation from external noise, to control for participants' varied listening environments as far as possible. The following procedure was approved and used for preparation, delivery, handling and return of the bespoke equipment used for part B between each participant:

1. All research hardware was sanitised and placed in a small box.
2. The small box was placed in a larger one, sealed with a lid and left in quarantine for a minimum of 72 hours.
3. A socially distanced package drop was made for the next participant by placing the large box at least 2 meters away, removing the lid and retreating 2 meters for them to collect the smaller box from inside.
4. The research activity was conducted over remote video link.
5. The participant repacked the equipment into the small box.

6. The package was collected by using the same method as in point 3, this time with the participant replacing rather than removing the smaller box.

The pilot study was used to test and validate the remote format. For part B it became apparent that participants' own working environments influenced their physical interaction with the BinVAD system, due to the configuration of their seat and desk. To reduce this potential influence, all participants in the final study were asked to complete part B whilst standing in a clear space with the smartphone interface strapped to the wrist of their non-dominant arm.

It also became apparent during the pilot that gradual drift in the head tracking's reference position (i.e. the precise sensor orientation that provided a neutral reading of 0° azimuth and 0° elevation), varied depending on the idiosyncrasies of individual participants' movement and ambient electromagnetic interference in their surroundings. The sensor had been attached permanently and securely to the headphones' headband using three tightly adjusted cable ties. Participants had also been guided to adjust the headphones so that they were always worn as securely as possible. These precautions – plus the fact that the noticeable drift occurred predominantly in azimuth readings and amounted to tens of degrees on occasion – meant that the inaccuracy could be clearly attributed to drift in sensor calibration, rather than physical slippage of the headphones themselves. If headphone movement had been the problem, it would have occurred more prominently via forward/backward shifts of position, as opposed to lateral drift. This would have affected elevation readings rather than azimuth. Whilst it was possible, during the final study, to check via video link for relative stability of the headphones' position, it was not possible to monitor magnetic drift remotely in real time. To minimise variance between participants, the head-tracking reference point was therefore checked and reset with the participant before they started each task, using a similar calibration procedure to that described in section 5.

For part A, participants used their own headphones and personal computer to use the GUI version of the standalone software application, but the session was still conducted in real time over a video call. The pilot study was used to debug software issues on specific operating system iterations, or to expose and resolve issues that prevented data being written and saved successfully.

Both systems automatically logged the participants' interactions (i.e. which tracks they listened to, kept and removed, for how many times and for how long in total). A written record was made of each participant's responses to the two open questions in 6.2.1, which were asked verbally at the end of the video call for part A.

6.3 Results

Results displayed in this section do not include the pilot study (participants A-C) and do not examine the outcome of the *example task*.

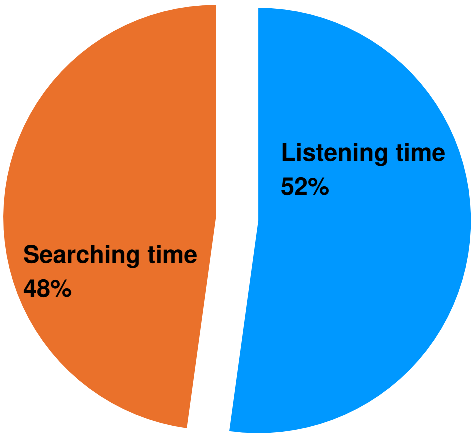


FIGURE 6.11: Aggregated time spent by participants between listening (solo mode) and searching (browsing mode) using BinVAD

6.3.1 BinVAD system interaction

Three sets of data related solely to participant use of the BinVAD system are presented to illustrate patterns of use amongst participants.

Figure 6.11 shows the aggregated division of time spent by participants between searching for tracks in standard mode (browsing) and listening to tracks in isolation (solo), across all tasks. It is notable that just over half of participants' interaction with BinVAD was spent listening closely to individual pieces. This is in marked contrast to the AAE system (see Figure 3.5), for example, where participants spent 36% of their engagement listening and 64% searching.

Examination of individual participants' interaction across each task in Figure 6.12 shows this division of activity differed between individuals and over time. Overall, 38 of 64 tasks were completed with a majority of time spent listening using solo mode. In some instances, participants completed a task almost exclusively in standard browsing mode. This was possible and actively enabled by the design of the swipe to 'keep' / 'remove' functionality, which did not necessitate use of the solo function. For example, participant F appears to have made virtually all of their playlist decisions within browsing mode only, as did participant K for *task 1*. Participant H, on the other hand, spent between 76% and 90% of their time for each task listening in solo mode. Similarly, by *task 4* participant K worked by auditioning tracks in isolation for 87% of their time.

It seems likely that, despite the interface instruction phase, participant F had either misunderstood or forgotten the solo listening function. Indeed, their comments during the qualitative interview made no reference to the solo feature. In contrast, Participant H's focus on solo listening would have required a deliberate decision. Their comments also confirmed that they preferred to work through the content methodically using close listening. Participant K's patterns of interaction were also in line with their subsequent

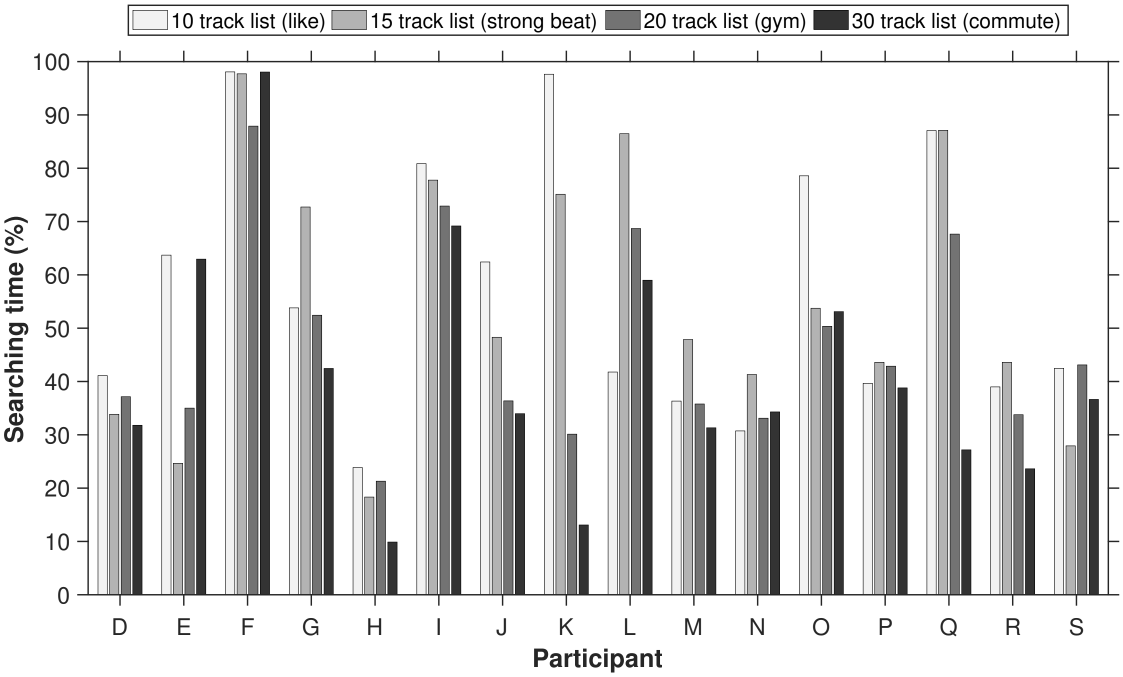


FIGURE 6.12: Proportion of time spent by each participant searching (in standard browsing mode) using BinVAD.

feedback. They acknowledged not fully understanding the solo listening function during the first task and then making greater use of this in the later tasks.

One indication of the BinVAD browsing environment's usability and usage is spread of activity across sound source locations. A similar principle was evaluated for the AAE prototype in Figure 3.6. This illustrated that content was explored relatively evenly, irrespective of category or track position in the laterally-organised auditory menu structures. Figure 6.13 shows the distribution of each participant's 'keep' and 'remove' actions within BinVAD, aggregated for all study tasks. Some participants appear to have favoured either the central (participants D and E), leftmost (H and L) or right (J, K, M and N) locations by an additional 10% or more of overall activity – i.e. 30% or higher. Five participants (G, J, K, M and P) only used a location for 10% of their interaction or lower, each in a single case at either the leftmost or rightmost position. The lowest proportion for any case was participant P's use of the leftmost position, which attracted only 8.04% of their overall activity.

6.3.2 BinVAD and GUI system comparison

128 tasks (16 participants completing 4 tasks on 2 systems) were completed in total. Three comparisons between the participant responses provided using the BinVAD and GUI system are presented to account for differences in engagement.

An Anderson-Darling test for normality confirmed log-normal distribution of all task completion times. A t-test analysis showed no significant difference between the time

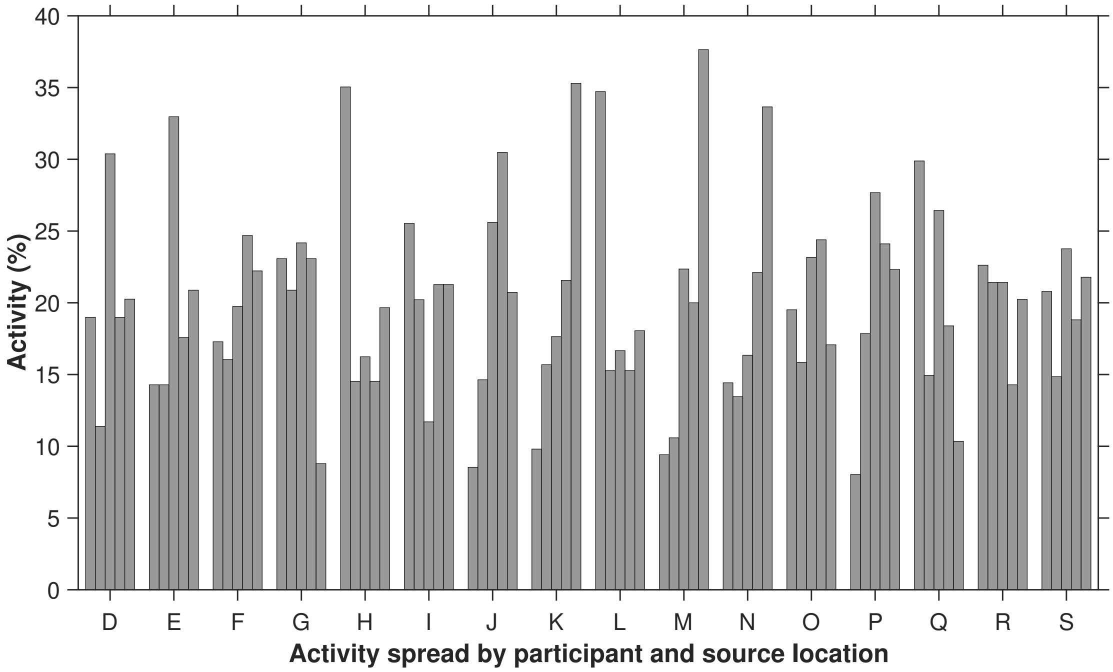


FIGURE 6.13: Proportion of activity using BinVAD ('keep' and 'remove') across source locations (-82°, -41°, 0°, 41°, 82°) for each participant

taken to provide task responses on either system (p = 0.341). This indicates that, as a cohort, participants did not take any more or less time to complete their study tasks on the BinVAD prototype compared to the GUI system. Figure 6.14 shows the distribution of task completion times for either system. In both cases, a clear majority of the shorter and more straightforward responses are found at durations of five minutes or less. In the case of BinVAD, 15 *task 1* and 13 *task 3* responses were submitted within this timeframe, compared to 13 and and 11, respectively, for GUI. There was a more even division of *task 2* and *task 4* submissions either side of five minutes, with both systems resulting in eight and nine responses exceeding this duration (respectively for either task). The main difference seen between the two systems is slightly greater consistency in the compact log-normal distribution of response times for BinVAD, compared to some unevenness and outlier responses in the GUI.

An indication of how carefully participants considered their responses can be derived from the average number of track appearances per task. An average track appearance value of 1 indicates that every track in the playlist was encountered just once by the participant before they submitted their response to that task. Values greater than one indicate that the participant continued to cycle through the playlist from the beginning to revisit tracks already heard and kept – i.e. they looped back over content before making final decisions on which tracks to remove. Values less than 1 occur in instances where the participant completed the task before hearing every track available in the playlist – i.e. they removed the requisite number of tracks to fulfil the task without experiencing every item in the list. Figure 6.15 illustrates the distribution of track appearances for all tasks

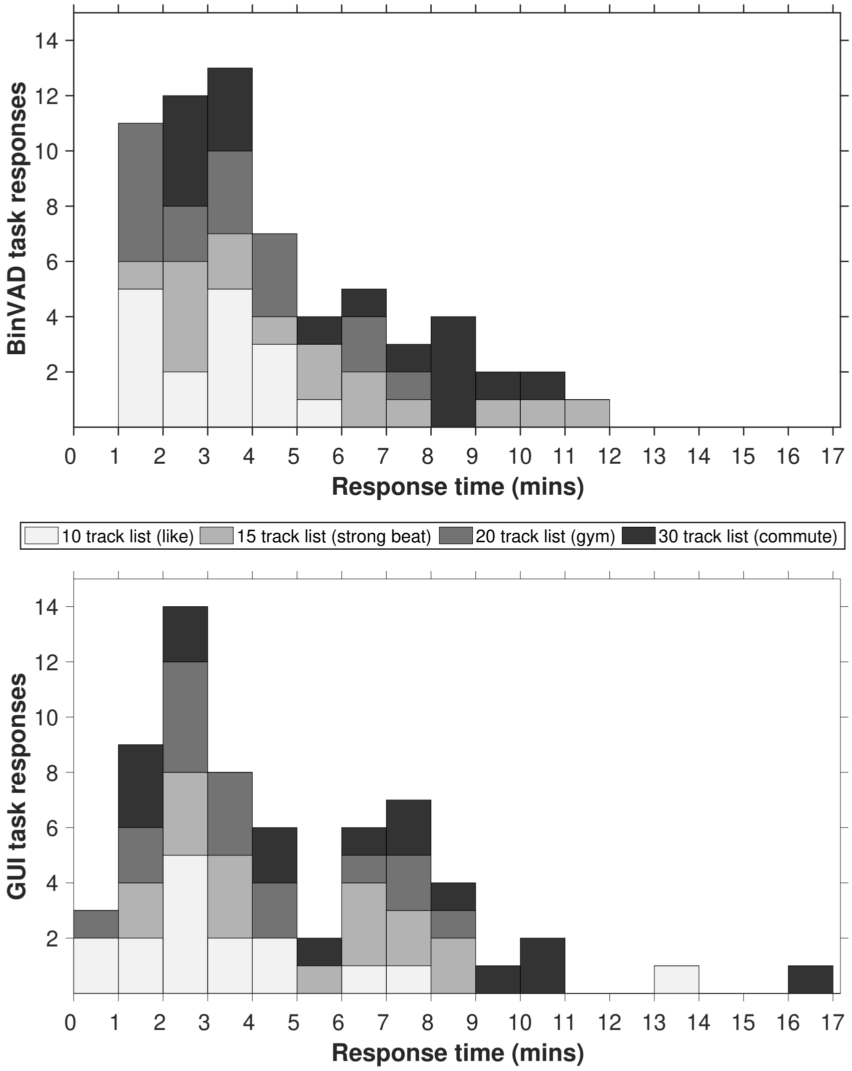


FIGURE 6.14: Comparison of all completion times by system, broken down by task type

completed by participants on either platform. A Wilcoxon signed rank test indicates that average track appearances were higher for BinVAD than for GUI to a statistically significant degree (p = 0.045), with median values of 1.3 and 1.2 respectively.

As described in section 6.2.2, *task 2* and *task 3* were conceived to enable objective assessment of participant responses. Any item submitted for either of these tasks that had not been included as a 'qualifying track' was judged to be an error (see Table 6.1 and section 6.2.2 for explanations of this definition). Figure 6.16 illustrates the distribution of response errors for assessed tasks returned by participants on either platform. A Wilcoxon signed rank test indicates that errors in participant submissions for *task 2* and *task 3* were more frequent for BinVAD than for GUI to a statistically significant degree (p = 0.002), with median values of 3 and 2 respectively.

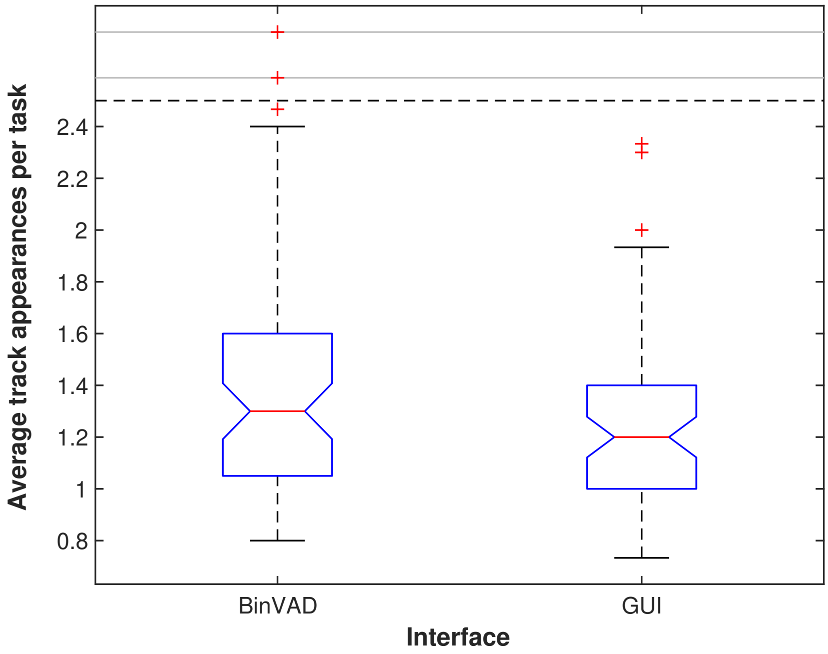


FIGURE 6.15: Distribution of average playlist track appearances per task, by system. Box plots indicate the 25th-75th percentile range and median value, which are significantly different between groups.

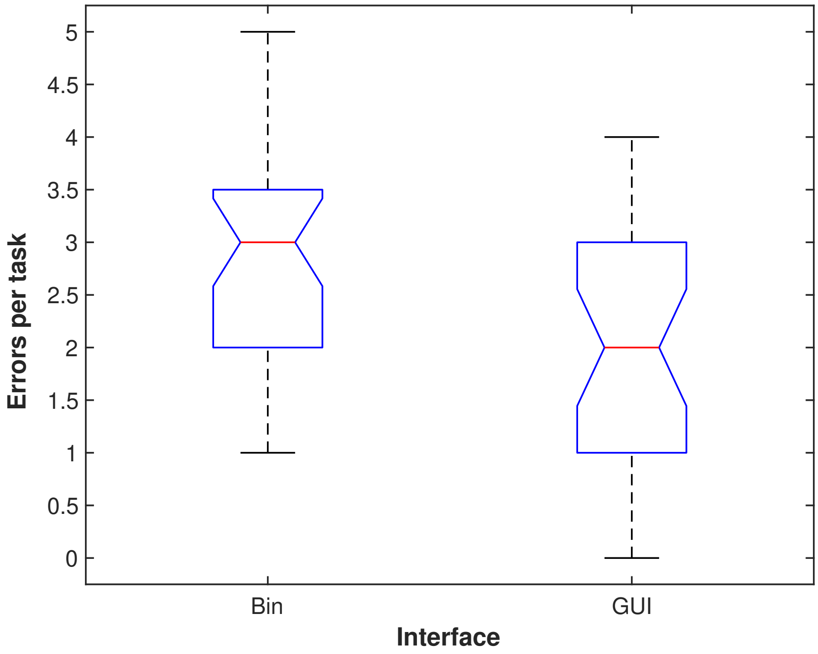


FIGURE 6.16: Distribution of response errors per assessed task, by system. Box plots indicate the 25th-75th percentile range and median value, which are significantly different between groups.

ID	Observation category	Participants
1	*Found the interactive binaural experience interesting/enjoyable*	D, F, H, M, N, O, P, S
2	*Noted a degree of challenge or learning curve to navigating the binaural scene*	D, E, G, I, J, K, M, N, S
3	*Found the binaural scene very confusing or uncomfortable to operate*	L, Q
4	*Felt that they spent most of their time in solo mode*	D, N, Q
5	*Felt inclined to focus their listening in one location*	E, G, K, M, Q
6	*Felt that they preferred browsing at the sides (where there were less streams audible)*	D, E, I, K, L
7	*Noted that the interface enabled them to make quick judgements*	F, M, N, O, P, S
8	*Noted that the touch surface interaction was intuitive*	E, I, M, P
9	*Noted difficulties with aspects of the touch surface interaction*	K, L, R
10	*Found the track voiceover announcements unclear*	E, L
11	*Found the track voiceover announcements distracting*	G, M
12	*Described a personal approach to completing tasks that exploited the concurrent auditory browsing environment*	E, H, I, J, K, M, N, O, P, R, S

TABLE 6.3: Qualitative observations on BinVAD performance by two or more participants

6.4 Discussion

Qualitative information gained from the two open questions posed at the end of Part B are presented in Table 6.3, to assist with interpretation of the quantitative data. Participants' answers to both open questions were treated as a single response and comments were categorised through manual analysis after all sessions were completed. Many participants chose to focus part of their responses on addressing the particular features or challenges of the playlist tasks or content. These factors were not relevant to this enquiry so they were ignored during the categorisation of comments. Otherwise, all categories of comment expressed by more than one participant were deemed noteworthy and included in Table 6.3.

The two research objectives identified in section 6.1 are now addressed in turn, to analyse results from the study in detail.

6.4.1 Using BinVAD for personalising playlists

The overall division of time spent between ensemble browsing (48%) and solo track auditioning (52%) suggests that the two modes of the auditory interface were evenly explored. Though 9 of 16 participants noted a degree of challenge or learning curve to using the auditory environment (category 2 in Table 6.3), the breakdown of usage for each participant and task in Figure 6.12 generally shows confident interaction with the interface features.

Only 14 of 64 tasks were completed with less than a third of that time spent in ensemble browsing mode. Four of these were submitted by participant H and just two other participants, K and M, spent under a third of their time on browsing in more than one task. Though some participants reported feeling that they favoured solo mode (category 4), none of these was H, K or M.

There is evidence of an inclination towards reduced use of ensemble mode as the tasks progressed (specifically the data in Figure 6.12 for D, H, I, J, K, L, M, O, Q and R all reflected this trajectory to some extent). This trend is understandable given the length and repeat demands of the four study tasks. A degree of fatigue with the concurrent browsing display would have been likely to set in towards the end of the session. This pattern would be less likely in a real-world use case of customising a single list at one moment in time, before listening to the edited playlist in its entirety. In this context, it would typically be some hours, or even days, before a user edited another playlist using the proposed system. Conversely, participant E shows a potentially more advanced degree of strategic engagement with the two modes. They had markedly different searching:listening time ratios, depending on whether they were engaging in subjective or objective tasks. For *task 1* and *task 4*, where decision making was entirely subjective, they spent almost two thirds of their time in browsing mode (64% and 63%, respectively). For *task 2* and *task 3*, where objective features of individual tracks had to be evaluated, they spent closer to a third of their time in browsing mode (25% and 35%, in either case).

Figure 6.13 further shows that the full width of the auditory environment was explored relatively well by most individuals in the cohort. Some evident preferences towards specific locations were already noted above for eight participants (including H, K and M) and for five who displayed under-use of a single location (including K and M). There is also degree of correlation between the patterns of individual activity shown in Figure 6.13 and most participants who either reported their tendency to remain focussed on a single location (category 5; E, K and M), or to browse in lateral locations for reduced auditory complexity (category 6; I, K and L). These patterns show an apparent element of self-awareness and strategising in participants' interaction decisions.

In most cases, therefore, each participant engaged openly with what were unfamiliar, immersive, audio-only design features to construct their responses. H, K and M were three participants who displayed a more narrow exploration of the auditory scene, as well as a tendency to rely on solo mode more than their peers. These three participants could be characterised as having approached the interface somewhat more conservatively and in line with traditional media playing interface usage, though both H and M were still within the 50% of participants who voluntarily expressed a positive response to the experience as a whole (category 1).

Of the two participants who noted particular discomfort with the environment (category 3), one had identified a slight hearing impairment prior to undertaking the task. The potentially disproportionate impact of using a generic HRTF on the degree of sound source segregation experienced between participants has already been discussed in section 6.1.1. It is also necessary to consider that asymmetric hearing ability is known to occur in a small but not insignificant proportion of the general population. One UK study found that 1% of adults who had neither been exposed to noise, nor been affected by conductive hearing loss from congenital or acquired illness, experienced asymmetric hearing. This definition was quantified as an average sensitivity difference between the left and right ears of 15 dB, measured with pure tones at 0.5, 1, 2 and 4 kHz (Lutman and Coles, 2009). It has been suggested that the plastic nature of human binaural perception can potentially adapt to altered spatial cues experienced from asymmetric hearing loss under normal acoustic conditions (Keating and King, 2013). However, auditory asymmetry would potentially exert a limiting effect on the spatial coherence experienced with binaural synthesis via non-individualised HRTFs, during a short listening session. It is therefore possible that the shortcomings of non-personalised binaural rendering used in the prototype would have affected the experience of one or more participants particularly acutely.

Overall, these findings provide substantial evidence that the binaural virtual auditory environment was used very successfully to pursue the four playlist editing tasks.

6.4.2 Differences in interaction between BinVAD and a GUI

That there was no significant difference in task completion time between either system suggests that, despite being presented with a wholly unfamiliar audio-only interface, this cohort of participants adapted to and utilised BinVAD as efficiently as an equivalent GUI. Further, the comparison of Figure 6.14 indicates a slightly more consistent distribution of response times in the former case. This statistical pattern is echoed by the third of participants who commented that BinVAD enabled them to make quick judgements (category 7). Together, these data indicate an initial suggestion that an improved flow state may have resulted for participants using this interface.

The possibility that BinVAD enabled improved focus is further supported by the significantly higher average track appearances with this interface. The average appearance value of 1.3 for BinVAD (compared to 1.2 for GUI) meant that participants were more likely to reach the end of the full playlist and/or revisit tracks more frequently. The auditory browsing environment therefore encouraged participants towards more widespread exploration and consideration of their responses before submission, but all within a comparable timeframe. Of the 16 participants, 11 described using a technique that specifically relied on the auditory nature of the BinVAD interface to pursue the task objectives (category 12). Such approaches broadly comprised three different methods:

- *'scanning'* - slowly surveying the whole of the current auditory scene to identify pieces of potential interest (e.g. for subjective *task 1* and *task 4*)

- *'homing-in'* - more rapidly evaluating the whole of the current auditory scene to evaluate pieces for required features (e.g. for objective *task 2* and *task 3*)
- *'parking'* - purposefully leaving a track in the auditory scene for it to progress further and be revisited (e.g. if it is of uncertain usefulness or interest in the case of any task)

It is difficult to evaluate how much added value these techniques specific to BinVAD brought in relation to the more clearly subjective *task 1* and *task 4*. However, the significantly higher error rate for *task 2* and *task 3* shows that aspects of BinVAD actually resulted in reduced accuracy of responses. Qualitative feedback reveals three potential causes of this increased rate of error. First, three participants found the touch surface swipe to 'remove' or 'keep' behaviour contrary to their expectations (category 9), compared with four who commented that they found this design intuitive (category 8). During informal discussion to resolve follow-up queries after the example task, more than one participant noted that they felt the swipe direction was inverted from the gestures they would instinctively complete for these purposes. Occurrences of accidental swipe motions when simply releasing the interface from solo mode were also informally flagged by participants, in between tasks. Inadvertent 'remove' or 'keep' gestures may therefore have contributed to increased error in BinVAD responses. Second, a total of four participants reported finding the track voiceover announcements unclear (category 10) or distracting (category 11). The implementation of this feature could also have affected some participants' execution of 'remove' or 'keep' gestures, since it occurred simultaneously with a touch interaction. Third, the degree of challenge that many participants noted in navigating the binaural scene itself has already been noted (category 2), as has the varying impact on spatial segregation experienced by individual participants due to the non-personalised binaural rendering implementation. Misdirected actions may well have been resulted on occasions where participants decided to make 'keep' or 'remove' actions without first verifying their target track in solo mode.

It should be noted that the the difference in errors, although statistically significant, is relatively small in the absolute terms, which amount to an average of one additional incorrect track per task completed using BinVAD, compared to the GUI. The potential causes discussed here could be explored further to be identified more precisely and potentially corrected through design refinements, such as: use of HRTF selection (as outlined in Chapter 4), a more dispersed auditory scene using just four or three tracks in the active scene (as explored in Chapter 3) and more simple refinements to the swipe interaction implementation.

6.5 Summary

The design and evaluation of a binaural auditory music browsing system for playlist editing formed the concluding part of this research enquiry. Pre-curated playlists are a mainstay of music streaming services and so an increasingly common feature in contemporary listening practices. The prototype was conceived with eyes-free operation in mind and careful curation of content enabled direct comparison of its usage with a standard GUI media player paradigm. Two research questions were answered. Firstly, patterns of interaction and qualitative feedback found that most participants were able to successfully adapt to the unfamiliar, immersive, audio-only design of the interface confidently and proactively, including good evidence of conscious engagement with the entirety of its features. Secondly, there are some signs that the prototype encouraged an enhanced flow state for playlist editing compared to an equivalent GUI media player system, in which more extensive activity was achieved over a comparable duration. Three specific techniques enabled by spatial auditory browsing were voluntarily identified by users in open questioning. Although tasks completed with BinVAD were more error-prone, potential causes of this pattern have been analysed and two likely areas for improvement related to earlier chapters directly identified (i.e. use of HRTF selection and increased spatial separation between sources).

Chapter 7

Conclusion

Let the music in tonight, just turn on the music.
Let the music of your life give life back to music.

'Give Life Back to Music'
Daft Punk

You say that everything sounds the same,
Then you go buy them.
There's no excuses, my friend,
Let's push things forward.

'Let's Push Things Forward'
The Streets

This thesis set out to revisit the potential application of binaural spatial audio in music discovery and recommendation. It was a topic that warranted renewed examination given substantially altered mechanisms for music distribution (the growth of digital streaming), consumption trends (the increasing proliferation of smart devices) and 3D binaural rendering capabilities on mobile computing platforms. The findings comprise outcomes regarding binaural sound interaction design for music content, considerations for 3D HRTF personalisation, and virtual auditory display implementation. Each of these research objectives (stated in section 1.2) is core to delivering immersive audio-only experiences that could be directed to augment music discovery and recommendation. Reflections on those three research questions and some considerations for future work are offered here in conclusion.

7.1 Review of research questions

1. What interaction design approaches might be employed to enable exploration of large collections of music using only 3D binaural sound as the display interface?

The Auditory Archive Explorer highlighted the potential of binaural spatial audio for transforming the engagement that users experience with existing audio-only interfaces, particularly for the purpose of music discovery. There were, however, no benefits found from simple spatialisation of sequentially presented content in terms of navigation fluency, or recall of the system's main design features. Yet, the slow responsiveness of voice control meant the interaction mode itself could have overridden latent usability benefits from binaural display. All subsequent research phases therefore relied on gestural input interaction.

A separate experiment in concurrent display of music tracks provided strong evidence that up to three concurrent music streams could be usefully applied in a search or browsing system – an outcome that has not been systematically established in research to date. Moreover, the same investigation demonstrated that providing graphical representation of sound source positions offered no apparent advantages to speed or accuracy of track identification.

The prototype binaural playlist editor provided more substantial evidence that spatial presentation of music sources can provide novel design opportunities for more personal and active engagement with recommended content. On this occasion there was some indication that the binaural implementation enabled an enhanced flow state when reducing playlists to a required length, compared against like-for-like tasks on a comparable graphical interface. More content tended to be covered over the same period of time with the binaural interface. Although error rates were slightly higher, likely causes of these have been identified and potential remedies relate to the subsequent research questions.

2. How can personalisation through HRTF selection be achieved to benefit interaction with music content in a 3D auditory environment using a mobile device?

Although there is a significant and growing body of research related to HRTF selection, very little of this focusses on end user solutions, even less so the particular constraints and considerations for mobile devices. The work in this thesis has advanced a set of four criteria for assessing end user HRTF selection procedures (*reliability*, *validity*, *usability* and *efficiency*). It has further presented a novel design for HRTF selection in 3D on mobile devices. Evaluation of that approach through repetition of the selection process (reliability), follow-up comparative localisation testing (validity), analysis of user listening habits (usability) and modelling of a shortened procedure (efficiency) showed demonstrable success against these criteria.

3. How can a 3D virtual auditory display system for presenting music content be implemented on a portable computing platform and within significant processing constraints?

A virtual vector base amplitude panning (VBAP) method for low-cost, head-tracked, 3D binaural rendering was developed as a core element of the thesis and assessed through digital measurement of localisation cue errors. Comparison against the baseline approach used currently for immersive media applications – first order virtual Ambisonics (FOA) – showed that the method proposed is demonstrably more suitable for interactive auditory display. Representation of lateral cues is substantially improved in all directions, but particularly so in the frontal region where, owing to the flexibility of the approach, a greater concentration of rendering resource can be deployed. This flexibility further benefits the interactive virtual auditory display use case.

Localisation testing with the system identified a minimum lateral separation between sound sources of 41° to support reliable user interactions. Although vertical localisation was confirmed to be insufficiently reliable, measured errors in spectral cues did not clearly indicate whether elevation representation was likely to be any worse than that for FOA. Finally, the virtual VBAP system was developed into a complete, real-time, 3D auralisation system by integrating a substantially simplified model for 3D scattering delay network synthetic reverberation, with customisable room parameters. This framework was adopted for the prototype binaural playlist editor used in the final study.

7.2 Future development

Potential areas for future work can be categorised into improvement of binaural rendering quality (questions 2 and 3) and additional sonic interaction design considerations for spatial auditory display of music content (question 1).

Rendering quality

A clear limitation in the design of the of virtual VBAP rendering system is that only one speaker array layout has been considered and tested. The goal, for the purposes of this thesis, was to achieve pseudo-spherical 3D rendering, for full and fair comparison with FOA. However, modifying the virtual speaker layout to, for instance, an array focussed solely on the upper hemisphere would enable improvements to rendering of elevated sources by discounting de-elevated source representation altogether. The outcome of such trade-offs could now be tested and analysed carefully to determine the preferred balance of available virtual sonic space against rendering fidelity.

Prioritisation could also be used to refine the 3D HRTF selection method, which in its current iteration asks users to evaluate simultaneously accuracy in azimuth, elevation and sense of externalisation. The aim in the context of this thesis was to test the viability of holistic 3D HRTF comparison. However, for an interactive, audio-only system, azimuth precision is less crucial than in other contexts that may involve spatial alignment of audio and visuals. Furthermore, a working approximation of the virtual VBAP system's lateral tolerances for interaction purposes is now established, based on all collected data (i.e. most and least preferred HRTF sets). It is possible that both the reliability and validity of HRTF selection outcomes could improve if a simplified judgement wording and / or comparison trajectory is presented for the selection procedure.

It is worth noting that this thesis has focussed solely on audio-only system design for immersive music discovery and recommendation through binaural auditory display. Yet, the growth of virtual reality and augmented reality (VR and AR) technologies – albeit somewhat uncertain, as discussed in Chapter 1 – potentially present other new modes and opportunities for accessing and exploring music content. Any future VR or AR technologies that might emerge for accessing music content would require sufficient levels of consistency between spatial audio cues and visual information to maintain sense of immersion and plausibility (Brinkmann and Weinzierl, 2023). To this end, virtual VBAP would need to be tested further, in conjunction with visual stimuli, to evaluate its suitability in VR- or AR-based music exploration systems.

Binaural interaction design for music content

Whilst this thesis has been underway, developments in music consumption and smart technology have been continuing to evolve. Smart headphones with integrated head-tracking, such as the more recent Apple AirPod models [1] and Yamaha YH-L700A[2], are gaining prominence. Apple Music's Spatial Audio delivery format has been delivering binaurally spatialised recordings since June 2021[3]. Familiarity with the quality and sensation of binaural audio will therefore be growing amongst music consumers. Playlist creation is also an expanding feature of music consumption amongst the 78% of 16-64 year olds who use music streaming services in 19 leading music consuming countries. Of this group, 62% listen to a playlist that they created more than once a week (IFPI, 2021). In the context of these developments, there is a clear case for focussing further research on users who have particular investment in either spatial audio or generating their own custom playlists.

The nature of the playlisting activity could also be broadened in future research from that used in this thesis. Great lengths were taken in the final study to use unfamiliar content that could be methodically categorised to make like-for-like comparison against an equivalent graphical user interface. That content was mostly produced by semi-professional or amateur contributors. Using the binaural playlist editor prototype therefore did not adequately represent the excitement of encountering a dynamically curated

[1] www.apple.com/uk/airpods/compare/ [accessed 20/6/2022].

[2] uk.yamaha.com/en/products/audio_visual/headphones/yh-l700a/index.html [accessed 20/6/2022].

[3] www.apple.com/newsroom/2021/05/apple-music-announces-spatial-audio-and-lossless-audio/ [accessed 20/6/2022].

list of content, which would typically be tailored to the listener and filled with a variety of surprises – familiar or otherwise. The true value and potential of immersive music browsing is to connect listeners with music that they might otherwise pass over. In this sense, *direct sonification*, via binaural virtual auditory display, might be used for more than just swiftly sifting the week's latest releases. It could also be a means of transporting someone instantaneously to a moment in their life one, two, five or ten years ago, via the tracks they then had on rotation. It could provide an immediate snapshot of what music an old friend they've lost contact with has been exploring this month. It could fluidly showcase the most frequently played tracks this week by their loved ones. These emotional touch points are significantly diminished if playlists are only seen in written form. They become vital and visceral when heard.

This thesis has taken several first steps towards understanding how 3D immersive audio presentation might bring life to music discovery in the age of streaming. By promoting more informed and responsive user approval or rejection of algorithmic recommendations, binaural virtual auditory display could present a means to push forward current models of engagement between music consumer and content provider.

Appendix A

Source code for original research software written as part of this thesis

A.1 Auditory Archive Explorer

https://github.com/rishi-s/Auditory-Archive-Explorer

The codebase used for the Auditory Archive Explorer research outlined in Chapter 3 , including voice command recognition (via SpeakOSC[1]), interaction and navigation design (using the Max[2] visual programming language) and audio content management (handled by a Reaper[3] digital audio workstation project in real-time).

A.2 Immersive Audio Headset

https://github.com/rishi-s/Immersive-Audio-Headset

The main binaural signal processing code (using C++) and graphical user interfaces (using the TouchOSC[4] GUI designer for iOS) devised for the research outlined in Chapters 4 to 6, including:

- Binaural virtual auditory display system and measurement
- Study 3a: User HRTF selection for 3D interactive audio
- Study 3b: Subjective localisation test
- Prototype binaural playlist editor

[1] github.com/dlublin/SpeakOSC [accessed 17/2/2022].
[2] cycling74.com/ [accessed 17/2/2022].
[3] www.reaper.fm/ [accessed 17/2/2022].
[4] hexler.net/touchosc [accessed 13/6/2022].

Appendix B

Study 1: Auditory Archive Explorer mental model drawings

B.1 Models of the binaural smart headphone experience

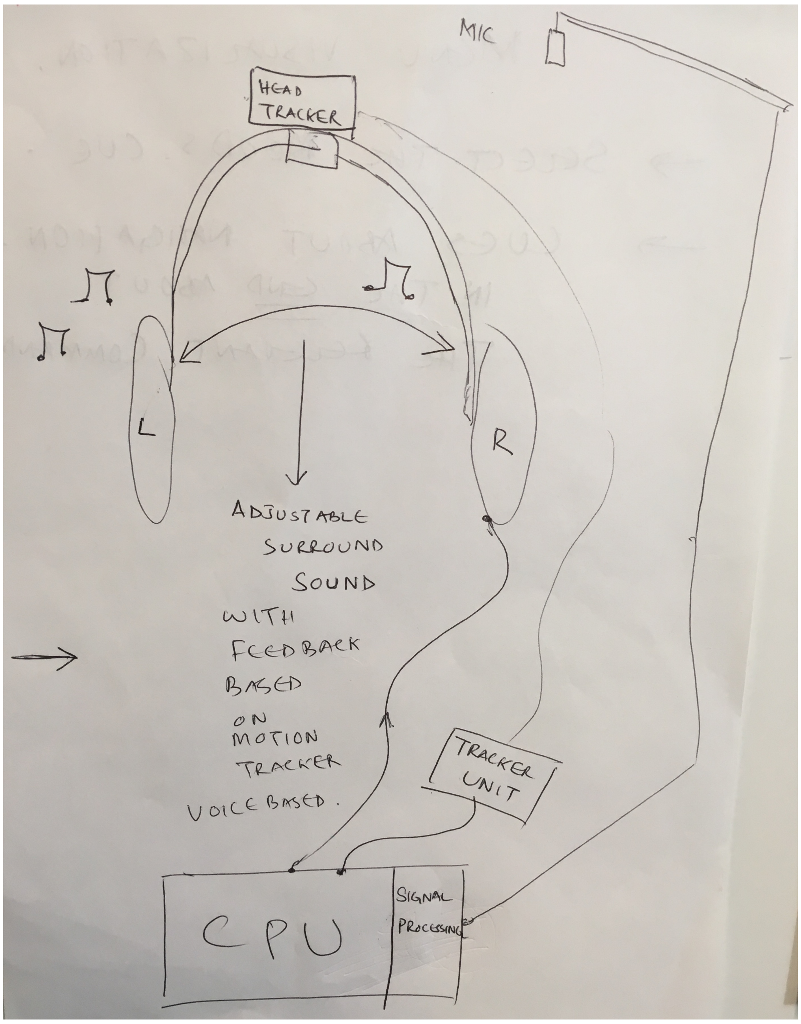

MIC
HEAD TRACKER
L
R
ADJUSTABLE SURROUND SOUND WITH FEEDBACK BASED ON MOTION TRACKER
VOICE BASED.
TRACKER UNIT
CPU
SIGNAL PROCESSING

FIGURE B.1: Binaural smart headphone experience participant A

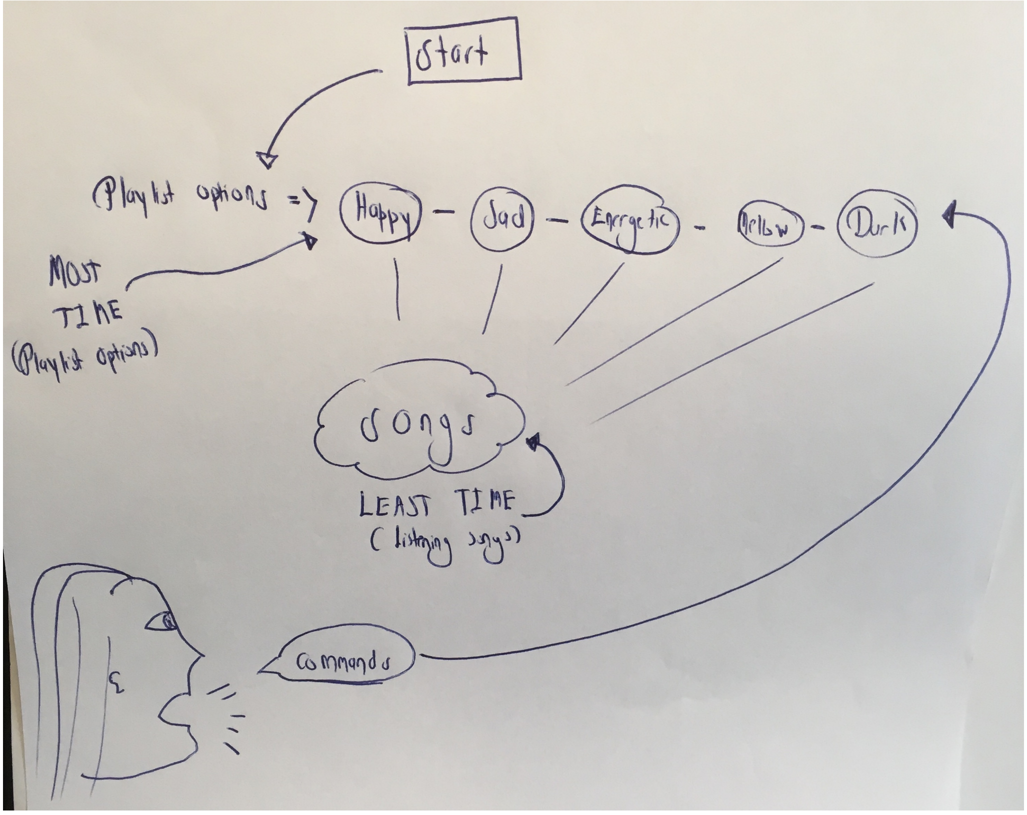


FIGURE B.2: Binaural smart headphone experience participant B

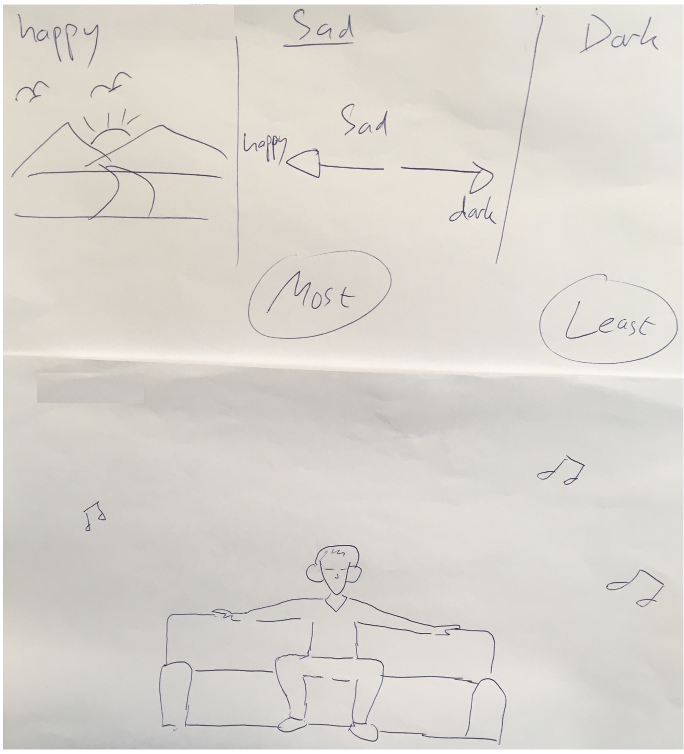


FIGURE B.3: Binaural smart headphone experience participant C

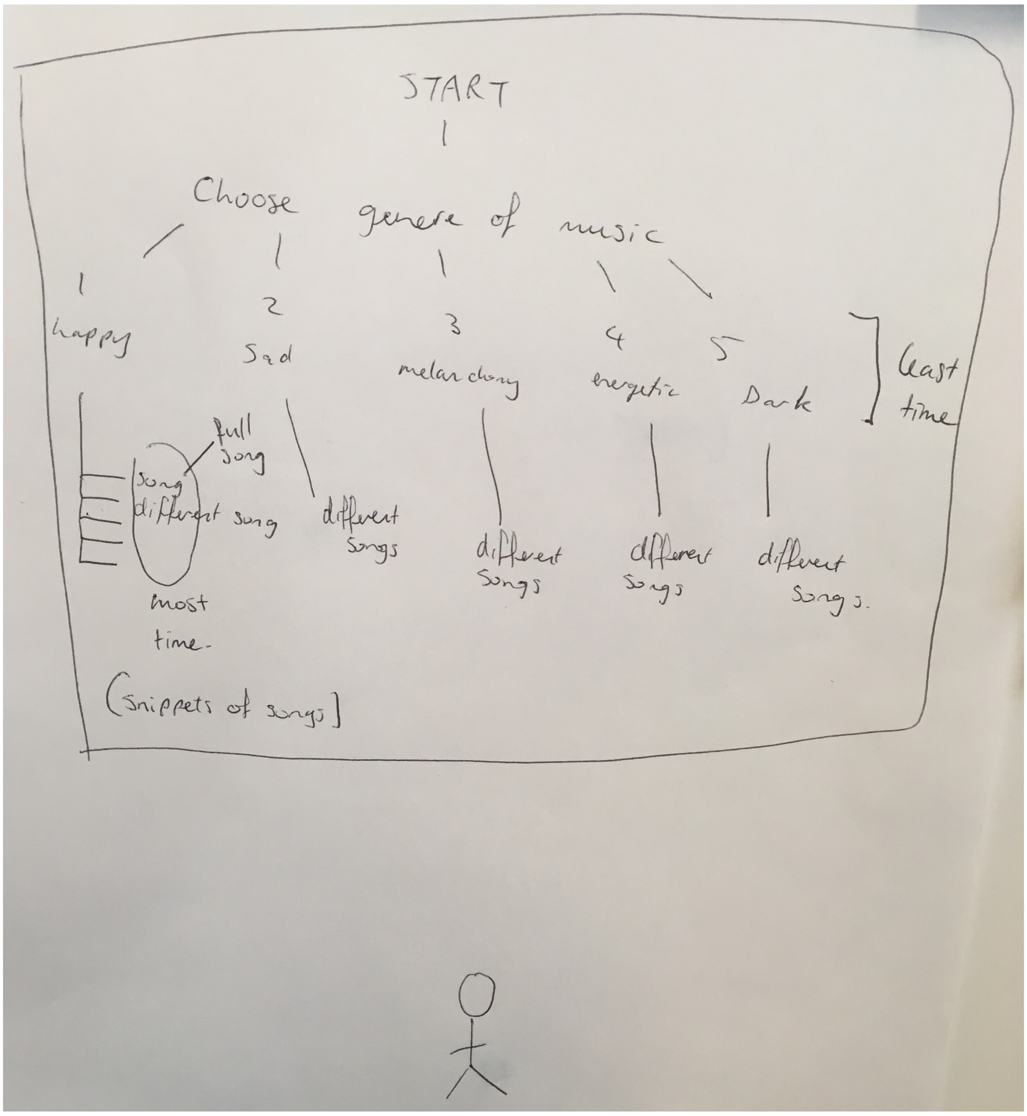


FIGURE B.4: Binaural smart headphone experience participant D

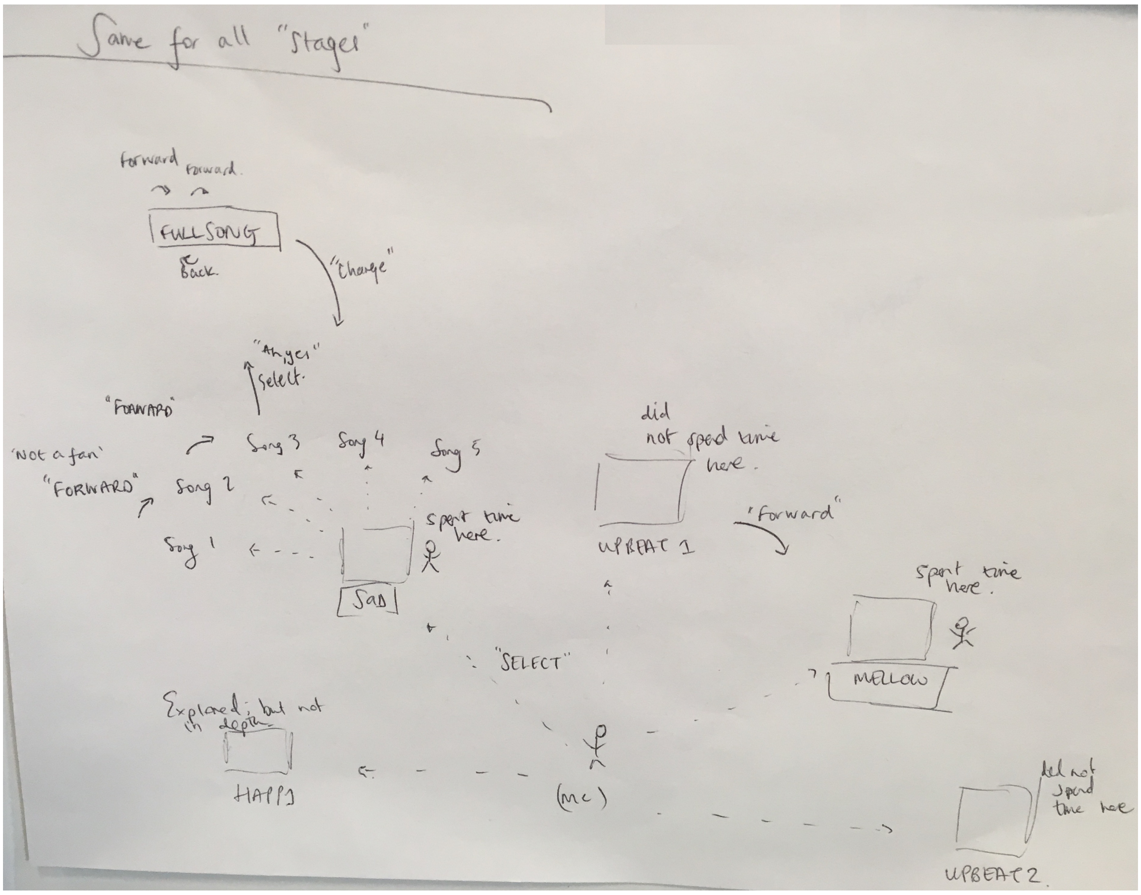


FIGURE B.5: Binaural smart headphone experience participant E

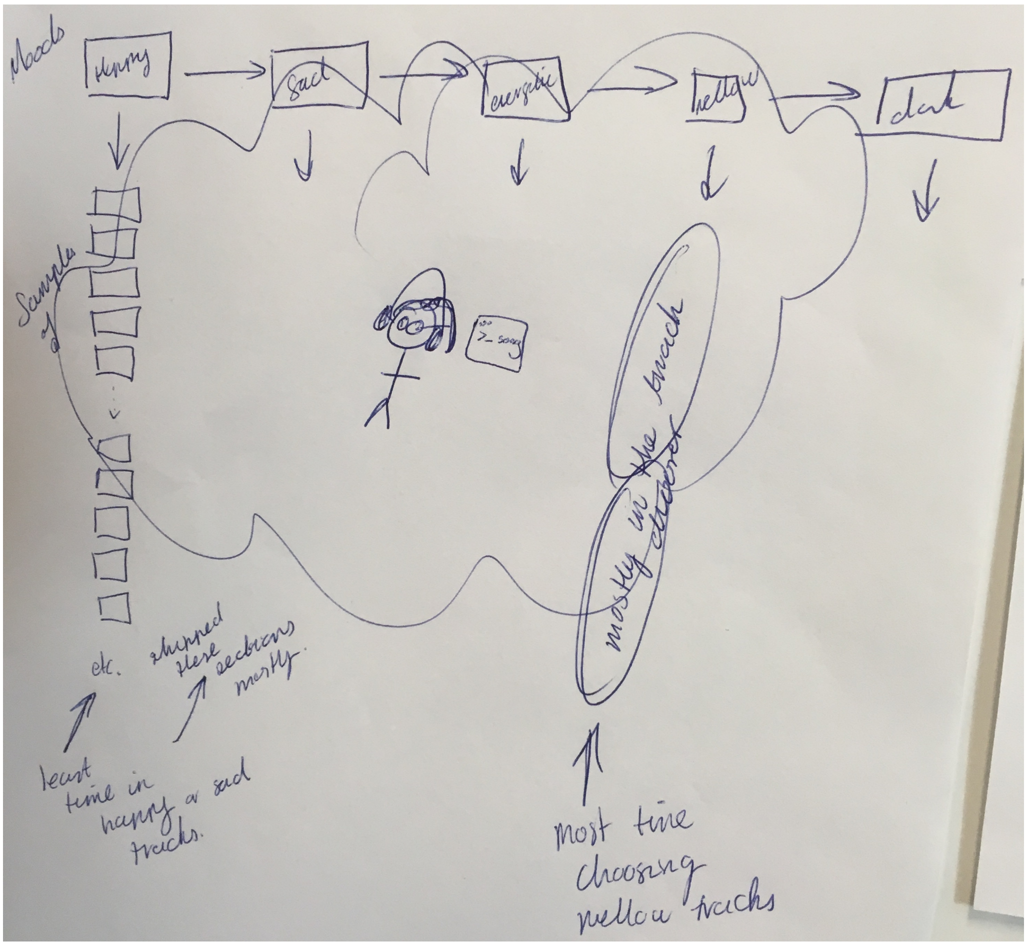

Moods
happy
sad
energetic
mellow
dark
Samples of
>_ song
mostly in the track cluster
etc.
skipped these sections mostly.
least time in happy or sad tracks.
most time choosing mellow tracks

FIGURE B.6: Binaural smart headphone experience participant F

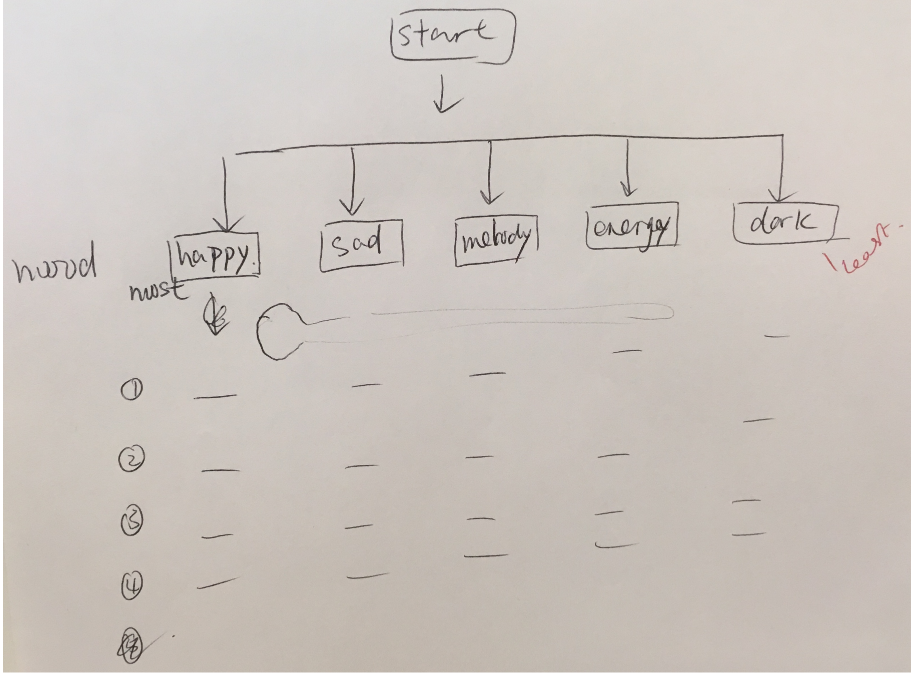


FIGURE B.7: Binaural smart headphone experience participant G

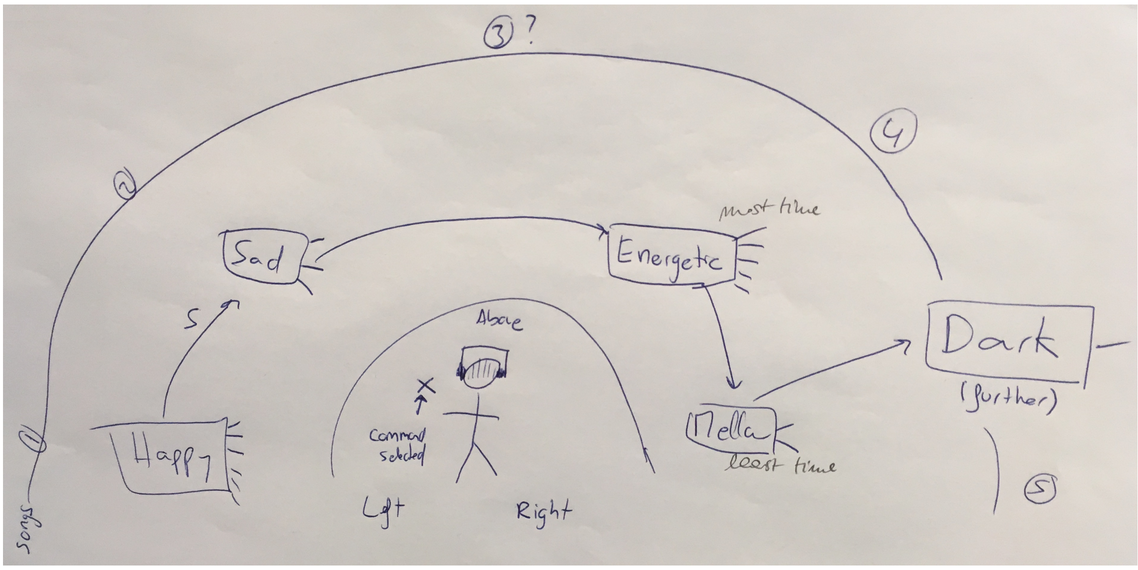


FIGURE B.8: Binaural smart headphone experience participant H

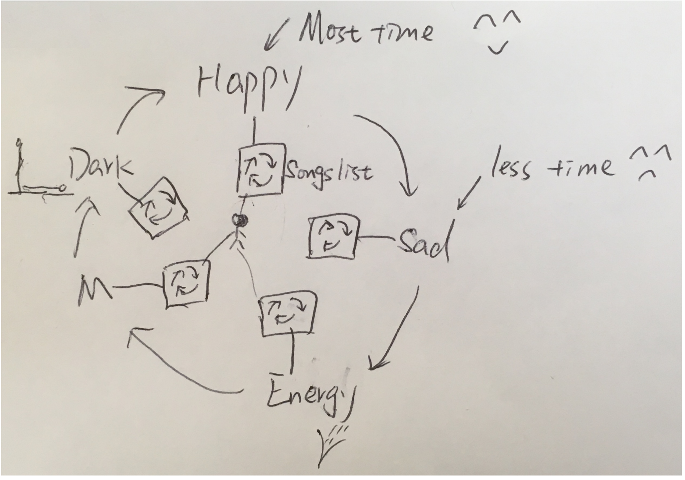


FIGURE B.9: Binaural smart headphone experience participant I

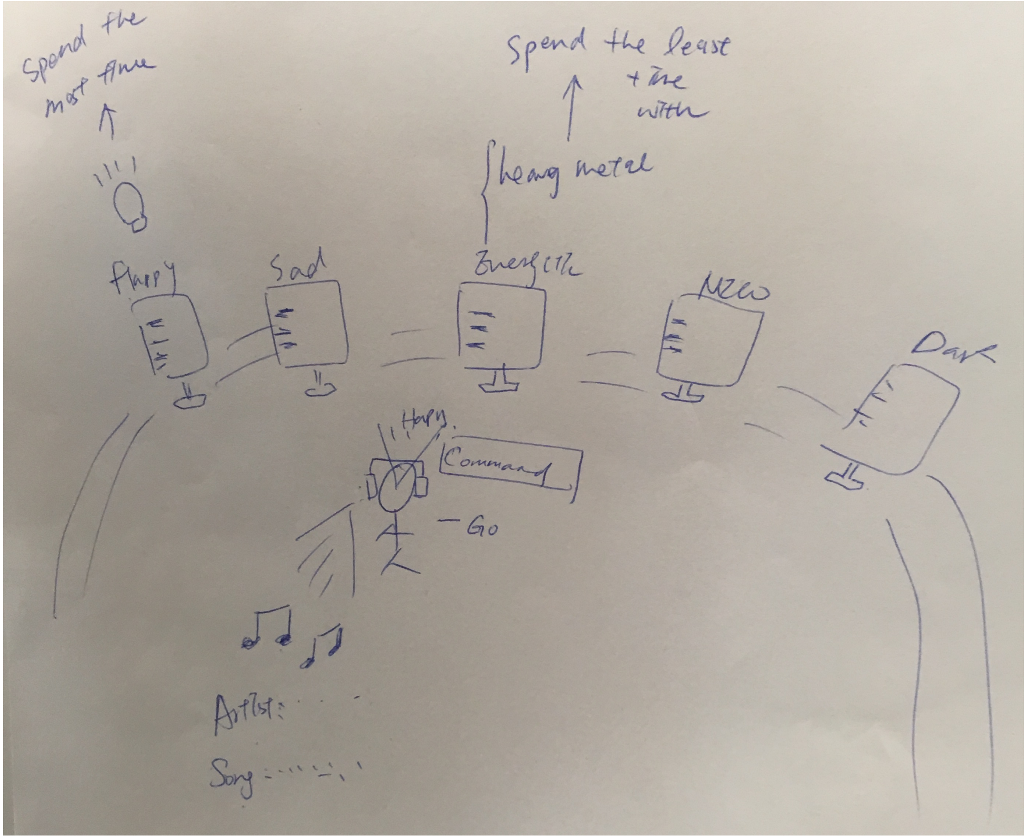


FIGURE B.10: Binaural smart headphone experience participant J

B.2 Models of the speaker experience

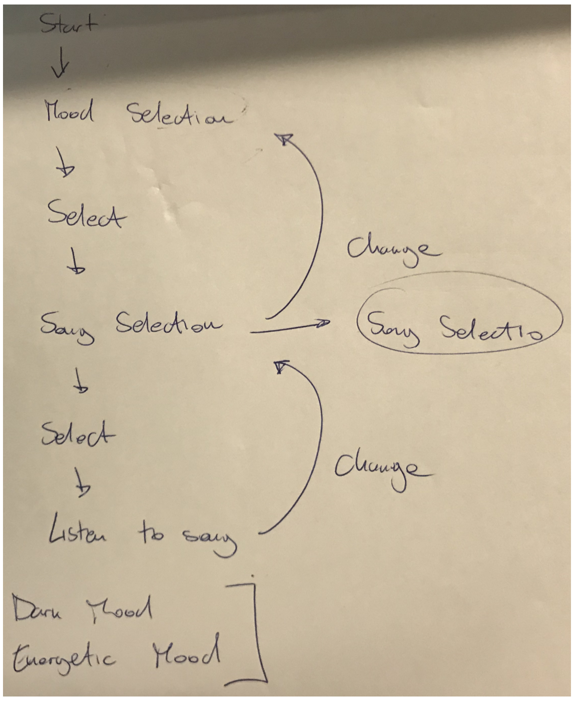


FIGURE B.11: Smart speaker experience participant K

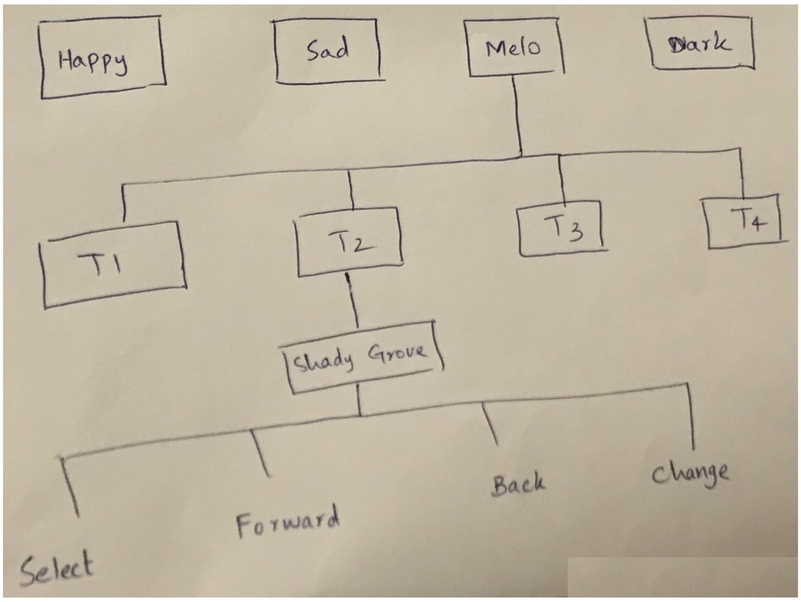
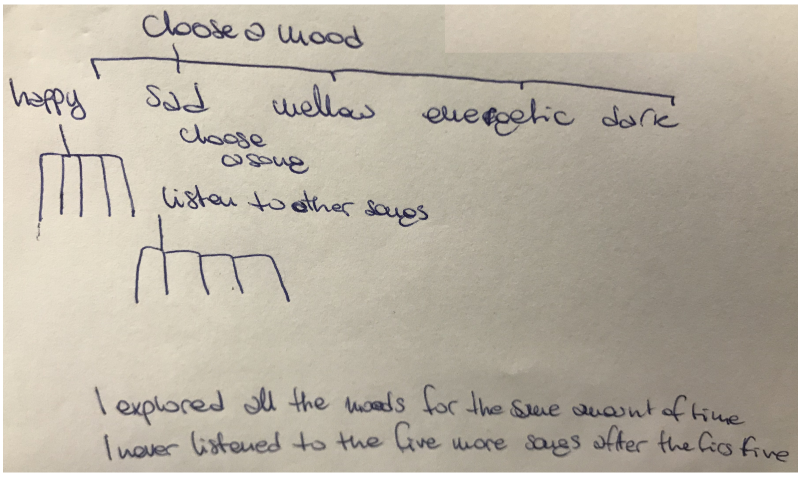
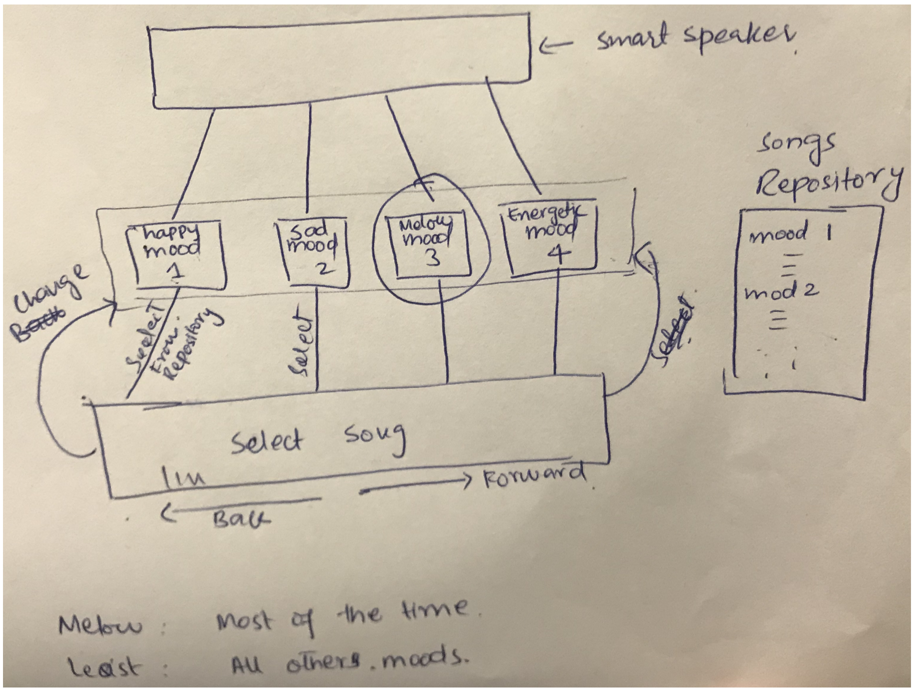


FIGURE B.12: Smart speaker experience participant L



I explored all the moods for the same amount of time
I never listened to the five more songs after the fios five

FIGURE B.13: Smart speaker experience participant M



Melow : most of the time.
Least : All others. moods.

FIGURE B.14: Smart speaker experience participant N

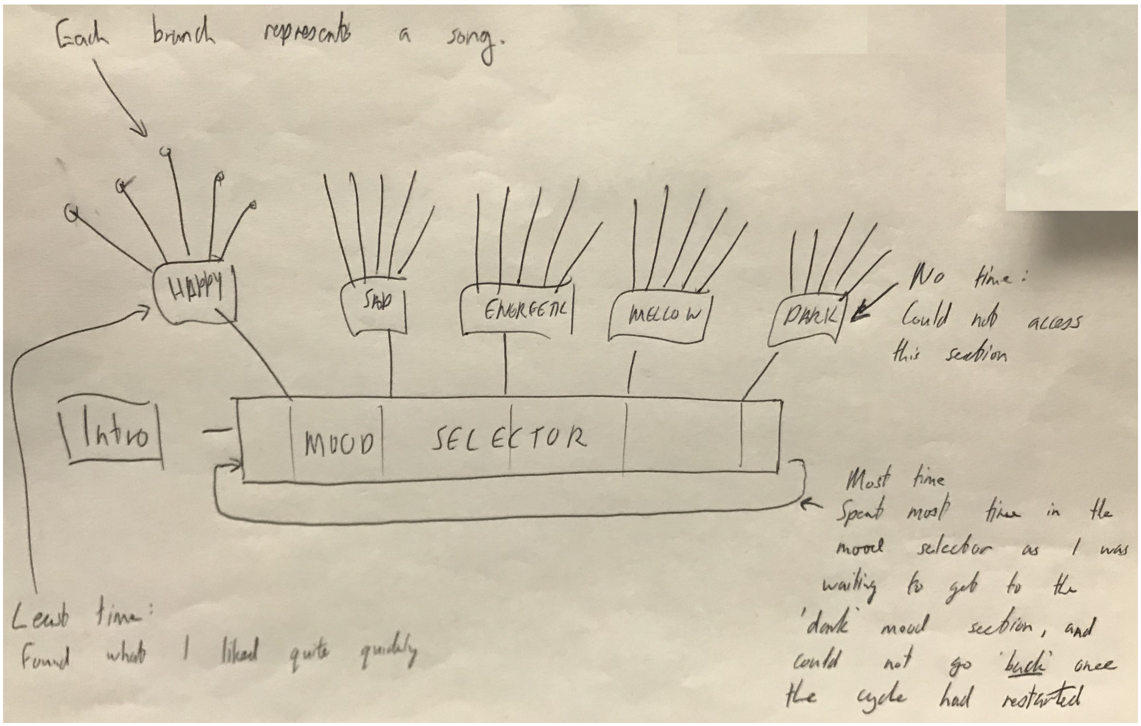

Each branch represents a song.
HAPPY
SAD
ENERGETIC
MELLOW
DARK
No time: Could not access this section
Intro
MOOD SELECTOR
Most time Spent most time in the mood selector as I was waiting to get to the 'dark' mood section, and could not go 'back' once the cycle had restarted
Least time: Found what I liked quite quickly

FIGURE B.15: Smart speaker experience participant O

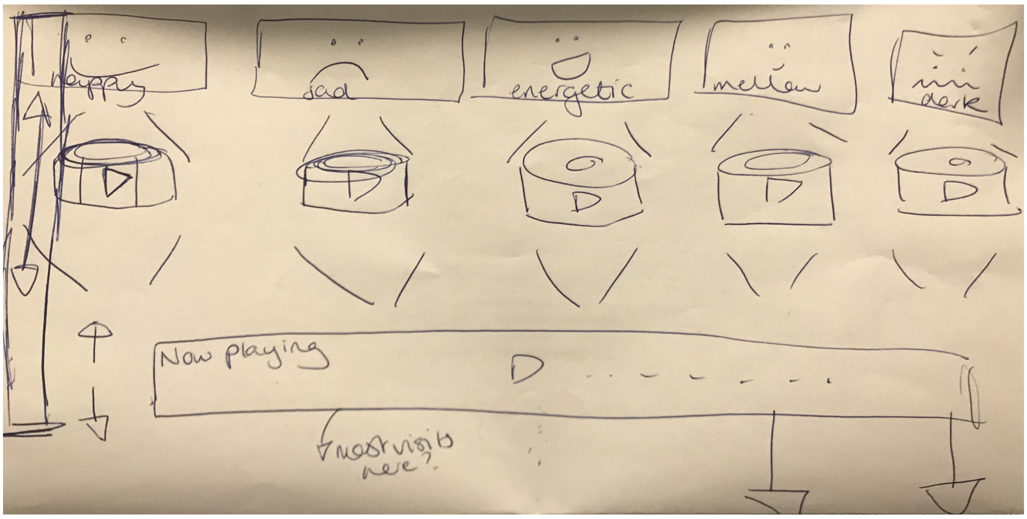


FIGURE B.16: Smart speaker experience participant P

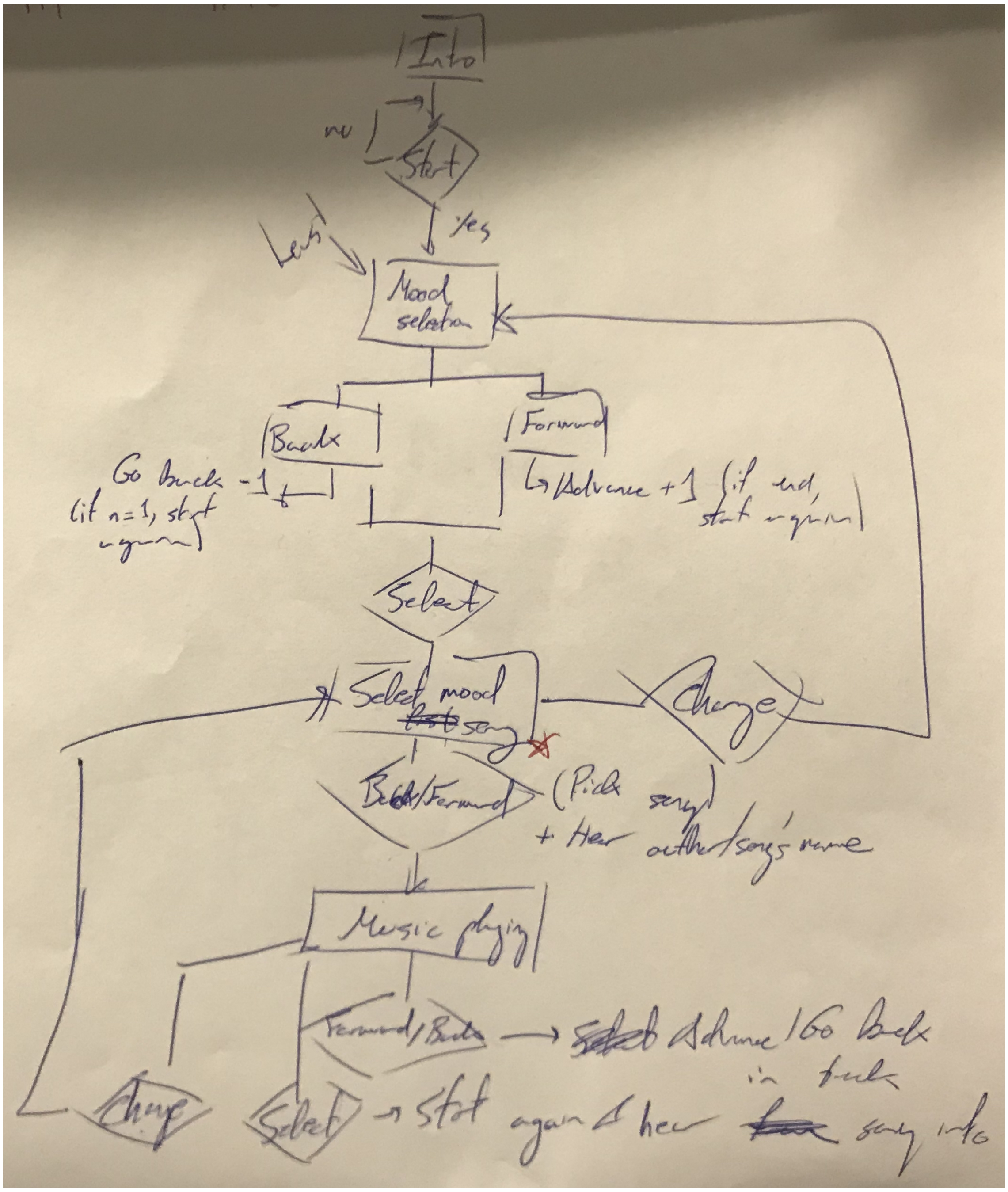

Intro
no
Start
yes
Leave
Mood selection
Back
Forward
Go back -1 (if n=1, start again)
Advance +1 (if end, start again)
Select
Select mood ~~list~~ song
Change
Back/Forward
(Pick song) + Hear author/song's name
Music playing
Forward/Back
→ ~~Select~~ Advance / Go back in track
Change
Select
→ Start again & hear ~~[illegible]~~ song info

FIGURE B.17: Smart speaker experience participant Q

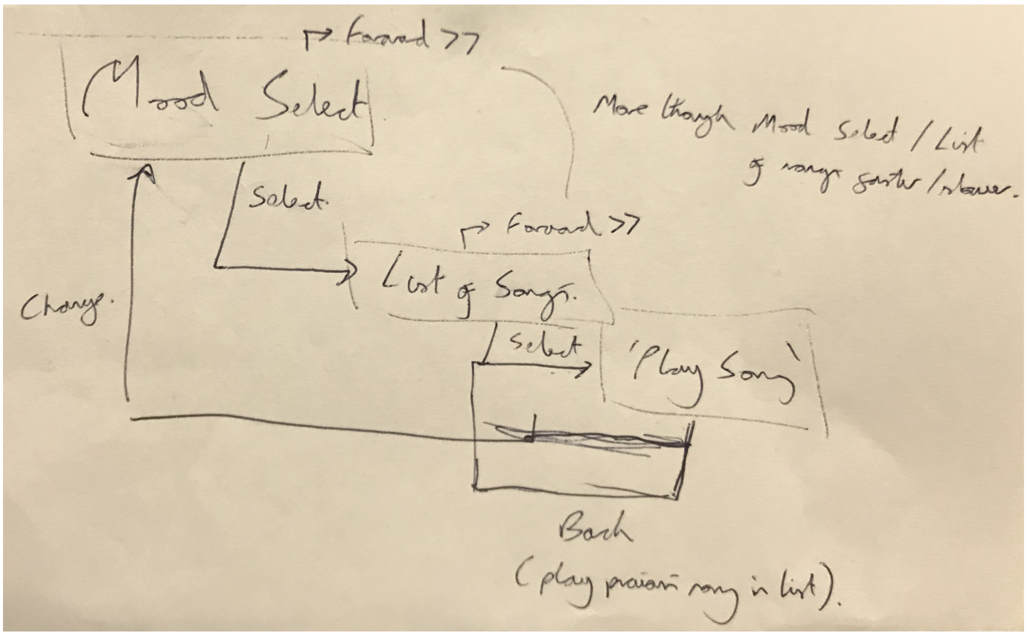

Forward >>
Mood Select
More though Mood Select / List of songs faster / slower.
Select.
Forward >>
List of Songs.
Change.
Select
'Play Song'
Back
(play previous song in list).

FIGURE B.18: Smart speaker experience participant R

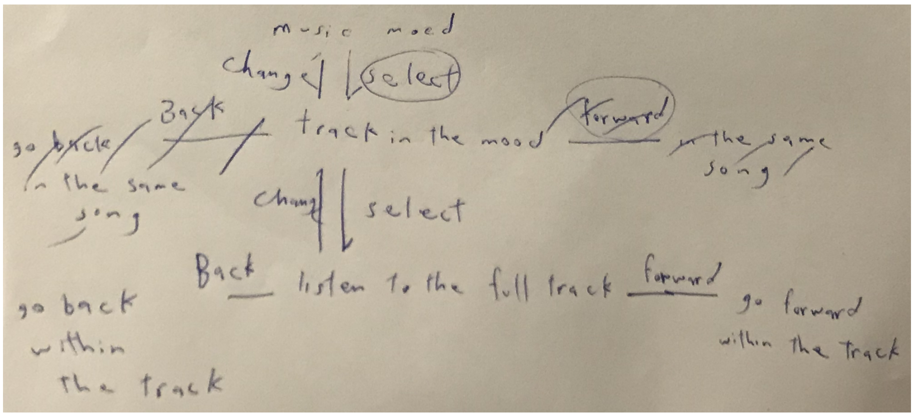


FIGURE B.19: Smart speaker experience participant S

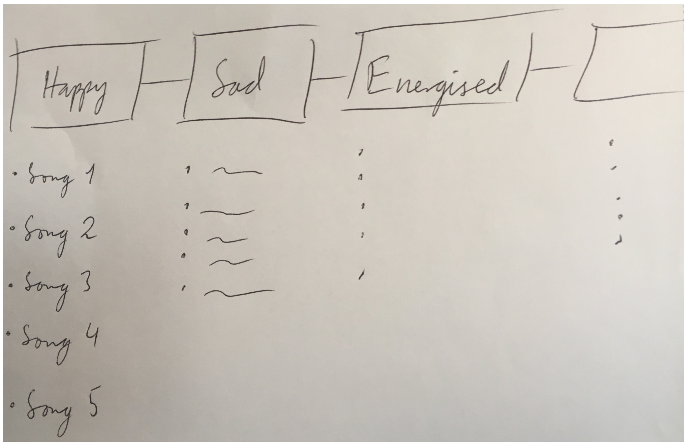


FIGURE B.20: Smart speaker experience participant T

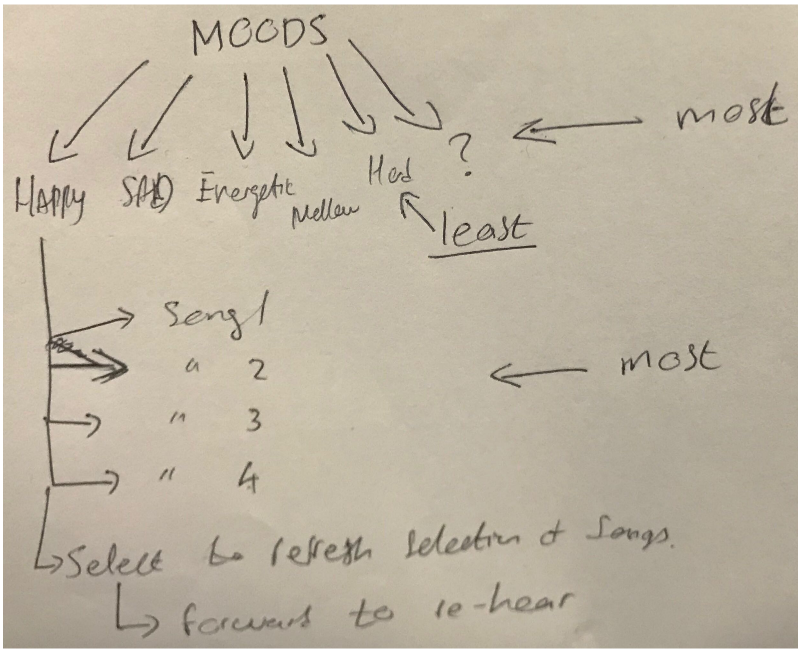


FIGURE B.21: Smart speaker experience participant U

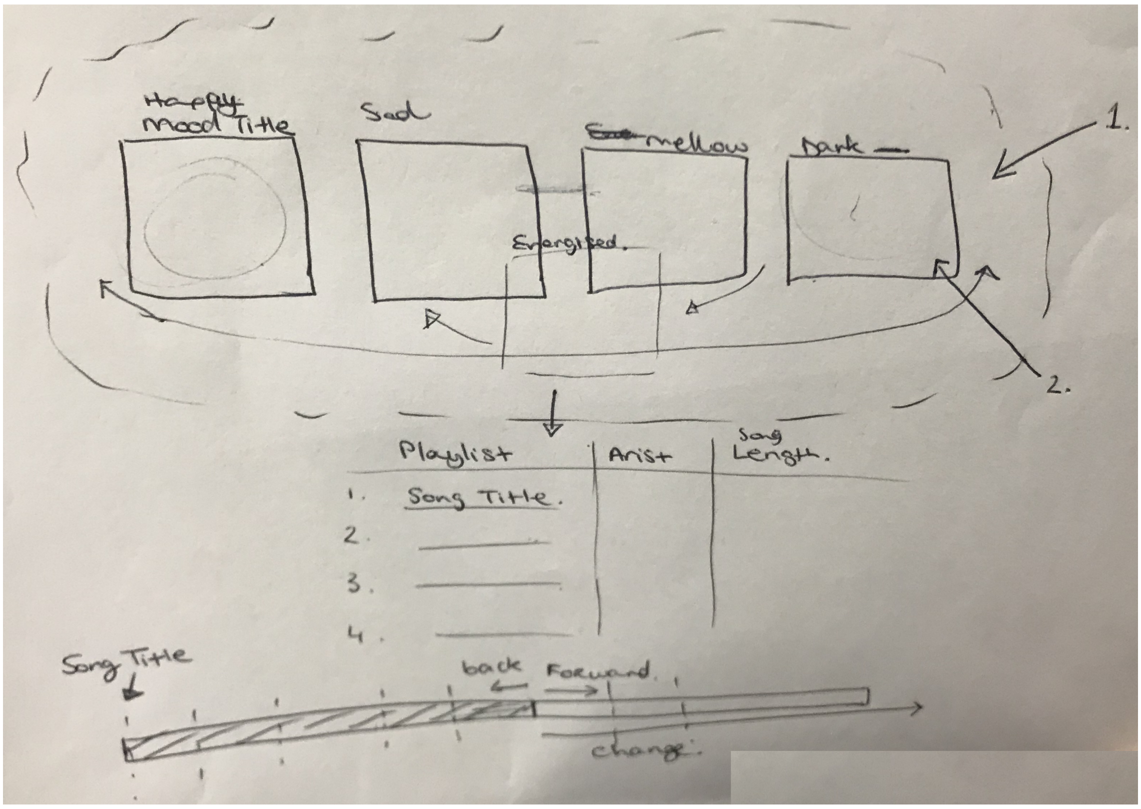


FIGURE B.22: Smart speaker experience participant V

Appendix C

Study 2b: Music dataset selection

C.1 Source and selection process

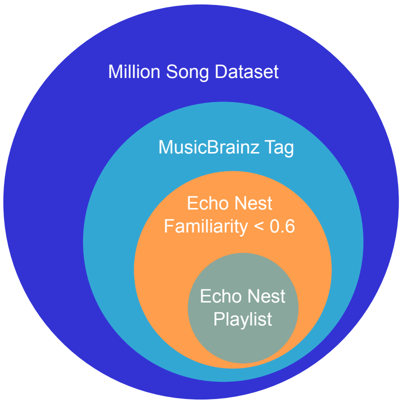


FIGURE C.1: Illustration of the song selection process from the *Million Song Dataset*. (Note that the scale is not representative of the subsets of the collection.)

The music used for the user experiment was drawn from the publicly available *Million Song Dataset* (Bertin-Mahieux et al., 2011). The dataset contains both metadata and features extracted from the audio signals of one million songs. At the time of preparing the content used in the experiment, further features could be queried through the Echo Nest API [1]. Short audio clips of about 30 seconds length could also be retrieved for songs through the 7digital API [2]. Since a large amount of music was needed for experiment, existing music information retrieval tools were used to filter and select the songs used, rather than hand curate the collection.

Figure C.1 illustrates the selection process. Five genres or styles of music were required — three of the styles are used in single-genre trials and all five are used in compiling mixed genre trials. The aim was to enable comparison of participants' ability to

[1] github.com/mukul13/REchoNest [accessed 4 April 2022]. An active Echo Nest API is no longer publicly available, but an archive of the library exists at this location.

[2] docs.7digital.com/ [accessed 4 April 2022]. The 7digital API no longer provides open access to track previews

discriminate similar (single-genre) and dissimilar (mixed genre) content presented concurrently. To find artists within the same genre, the *Million Song Dataset* was queried to retrieve all artists with a MusicBrainz tag of: "hiphop", "hip hop" or "hip-hop"; "rock"; "jazz"; "latin"; or "electronic". The three different spellings of 'hip-hop' were used as no single spelling alone retrieved a sufficient number of tracks. Popular songs that are easily recognisable are difficult to control for within a music listening experiment, as different users may not be familiar with the same songs. The selection process aimed to control for this factor by selecting songs that were from relatively unknown artists.

Firstly, The Echo Nest artist familiarity measure was used to reject artists that were too well known by retrieving only songs of artists that had a familiarity measure less than 0.6 (where 1.0 is very familiar and 0.0 relatively unknown). Secondly, the Echo Nest playlist API was used to ensure that songs in each genre pool sounded similar. To achieve this, the songs identified for each genre were grouped into five separate Echo Nest catalogues. A playlist was then requested from each catalogue based on the corresponding seed MusicBrainz tag (i.e. the original artist style descriptor), with the additional stipulation that returned songs had a maximum 'hotttnesss' of 0.6. Hotttness is the Echo Nest measure of the current popularity a song or artist, so this threshold further controlled for likely familiarity.

The Echo Nest playlist and catalogue APIs did not support query of the whole *Million Song Dataset* without a commercial agreement. So, selecting smaller sets of songs that were first filtered by the defined artist MusicBrainz tags and familiarity metric was a necessary step before querying the playlist API to ensure similarly. The outcome of this process returned the following number of tracks for each genre:

- 123 electronic
- 83 hip-hop
- 83 jazz
- 75 latin
- 76 rock

Bibliography

Aguilera, Emanuel, Jose J. Lopez and Jeremy R. Cooperstock (2016). "Spatial audio for audioconferencing in mobile devices: Investigating the importance of virtual mobility and private communication and optimizations". In: *Journal of the Audio Engineering Society* 64.5, pp. 332–341.

Algazi, V. R. et al. (2001). "The CIPIC HRTF database". In: *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*. New Paltz, NY, USA, pp. 99–102.

Andreopoulou, Areti, Durand R. Begault and Brian F. G. Katz (Aug. 2015). "Inter-laboratory round robin HRTF measurement comparison". In: *IEEE Journal of Selected Topics in Signal Processing* 9.5, pp. 895–906.

Andreopoulou, Areti and Brian F. G. Katz (June 2016). "Investigation on subjective HRTF rating repeatability". In: *Audio Engineering Society Convention 140*. Paris, France, pp. 1–10.

Andreopoulou, Areti and Agnieszka Roginska (Sept. 2014). "Evaluating HRTF similarity through subjective assessments: factors that can affect judgment". In: *Joint 40th International Computer Music Conference & 11th Sound and Music Computing Conference*. Athènes, Greece: Michigan Publishing, pp. 1375–1381.

Armstrong, Cal et al. (2018). "A perceptual evaluation of individual and non-individual HRTFs: A case study of the SADIE II database". In: *Applied Sciences* 8.11, pp. 1–21.

Bahu, Hélène et al. (2016). "Comparison of different egocentric pointing methods for 3D sound localization experiments". In: *Acta Acustica united with Acustica* 102.1, pp. 107–118.

Baldis, Jessica J. (2001). "Effects of spatial audio on memory, comprehension, and preference during desktop conferences". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 3, pp. 166–173.

Barrington, Luke, Reid Oda and Gert Lanckriet (2009). "Smarter than genius? Human evaluation of music recommender systems". In: *Proceedings of the 10th International Society for Music Information Retrieval Conference*. Kobe, Japan, pp. 357–362.

Baumgartner, Robert and Piotr Majdak (2015). "Modeling localization of amplitude-panned virtual sources in sagittal planes". In: *Journal of the Audio Engineering Society* 63.7-8, pp. 562–569.

Bederson, Benjamin B. (May 1995). "Audio augmented reality: a prototype automated tour guide". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Denver, Colorado, USA, pp. 210–211.

Begault, Durand R. (1994). *3D sound for virtual reality and multimedia*. 1st ed. London, UK: Academic Press Limited.

Begault, Durand R., Elizabeth M. Wenzel and Mark 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." In: *Journal of the Audio Engineering Society* 49.10, pp. 904–916.

Bela (2018). *Introducing the new Bela Mini*. URL: `https://blog.bela.io/2018/02/22/bela-mini-launch/`.

Berger, Christopher C. et al. (2018). "Generic HRTFs may be good enough in virtual reality. Improving source localization through cross-modal plasticity". In: *Frontiers in Neuroscience* 12.February, pp. 1–9.

Bertet, Stéphanie et al. (2013). "Investigation on localisation accuracy for first and higher order Ambisonics reproduced sound sources". In: *Acta Acustica united with Acustica* 99.4, pp. 642–657.

Bertin-Mahieux, Thierry et al. (Oct. 2011). "The Million Song Dataset". In: *Proceedings of the 12th International Society for Music Information Retrieval Conference*. Miami, Florida, USA, pp. 591–596.

Blattner, Meera M., Denise A. Sumikawa and Robert M. Greenberg (1989). "Earcons and icons: Their structure and common design principles". In: *Human-Computer Interaction* 4.1, pp. 11–44.

Blauert, Jens (1997). *Spatial hearing: the psychophysics of human sound localization*. Revised. Cambridge, Massachusetts: MIT Press.

Bogdanov, Dmitry et al. (Nov. 2013). "Essentia: an audio analysis library for music information retrieval". In: *Proceedings of the 14th International Society for Music Information Retrieval Conference*. Curitiba, Brazil, pp. 493–498.

BPI & ERA (2018). *Everybody's talkin': smart speakers & their impact on music consumption*. Tech. rep. London, UK: Music Ally. URL: `https://www.bpi.co.uk/media/1645/everybodys-talkin-report.pdf`.

Brazil, Eoin and Mikael Fernström (June 2006). "Investigating concurrent auditory icon recognition". In: *Proceedings of the 12th International Conference on Auditory Display*. London, UK, pp. 51–58.

Brazil, Eoin, Mikael Fernström and John Bowers (May 2009). "Exploring concurrent auditory icon recognition". In: *Proceedings of the 15th International Conference on Auditory Display*. Copenhagen, Denmark, pp. 1–4.

Bregman, Albert S (1999). *Auditory scene analysis: the perceptual organization of sound*. 2nd. Cambridge, MA, USA: MIT Press.

Brewster, Stephen et al. (Apr. 2003). “Multimodal ’eyes-free’ interaction techniques for wearable devices”. In: *Proceedings of the Conference on Human Factors in Computing Systems*. Ft. Lauderdale, Florida, USA: ACM Press, pp. 473–480.

Brimijoin, W. Owen, Alan W. Boyd and Michael A. Akeroyd (2013). “The contribution of head movement to the externalization and internalization of sounds”. In: *PLOS ONE* 8.12, pp. 1–12.

Brinkmann, Fabian and Stefan Weinzierl (2023). “Audio quality assessment for virtual reality”. In: *Sonic Interactions in Virtual Environments*. Ed. by Michele Geronazzo and Stefania Serafin. Cham: Springer International Publishing, pp. 145–178.

Brungart, Douglas S. and Brian D. Simpson (Oct. 2005). “Optimizing the spatial configuration of a seven-talker speech display”. In: *ACM Transactions on Applied Perception* 2.4, pp. 430–436.

Campbell, Murray and Clive Greated (1987). *The musician’s guide to acoustics*. Oxford, UK: Oxford University Press.

Cherry, E. Colin (1953). “Some experiments on the recognition of speech, with one and with two ears”. In: *Journal of the Acoustical Society of America* 25.5, pp. 975–979.

Cohen, Michael, Shigeaki Aoki and Nobuo Koizumi (Nov. 1993). “Augmented audio reality : telepresence / VR hybrid acoustic environments”. In: *IEEE International Workshop on Robot and Human Communication*. Tokyo, Japan, pp. 361–364.

De Sena, Enzo, Hüseyin Hacıhabiboğlu and Zoran Cvetkovic (Feb. 2011). “Scattering delay network: An interactive reverberator for computer games”. In: *AES 41st International Conference: Audio for Games*. London, UK: AES, pp. 1–11.

De Sena, Enzo et al. (2015). “Efficient synthesis of room acoustics via scattering delay networks”. In: *IEEE Transactions on Audio, Speech and Language Processing* 23.9, pp. 1478–1492.

Digenis, Aristotel (2017). *Ambisonic encoding / decoding and binauralization library in C++*. URL: https://github.com/videolabs/libspatialaudio.

Djordjević, Stojan et al. (June 2020). “Evaluation of the perceived naturalness of artificial reverberation algorithms”. In: *Audio Engineering Society Convention 148*. Online: AES, pp. 1–10.

Facebook (2017). *Powered by Audio360: Spatial Workstation user guide*. URL: https://facebook360.fb.com/spatial-workstation/.

Fazal, Muhammad Abu ul, Sam Ferguson and Andrew Johnston (Sept. 2018). “Investigating concurrent speech-based designs for information communication”. In: *Proceedings of the Audio Mostly Conference on Sound in Immersion and Emotion*. 1. Wrexham, UK: ACM Press, pp. 1–8.

Fellgett, Peter (1975). “Ambisonics. Part one: General system description”. In: *Studio Sound* 17.40, pp. 20–22.

Fernström, Mikael (2005). “Reflections on sonic browsing: Comments on Fernström and McNamara, ICAD 1998”. In: *ACM Transactions on Applied Perception* 2.4, pp. 500–504.

Fernström, Mikael and Eoin Brazil (July 2001). "Sonic browsing: An auditory tool for multimedia asset management". In: *Proceedings of the 2001 International Conference on Auditory Display*. Espoo, Finland: ICAD, pp. 132–135.

Fernström, Mikael, Eoin Brazil and Liam Bannon (2005). "HCI design and interactive sonification for fingers and ears". In: *IEEE Multimedia* 12.2, pp. 36–44.

Fernström, Mikael and Caolan McNamara (2005). "After direct manipulation – direct sonification". In: *ACM Transactions on Applied Perception* 2.4, pp. 495–499.

Freed, Adrian and Andy Schmeder (June 2009). "Features and future of Open Sound Control version 1.1 for NIME". In: *Proceedings of the International Conference on New Interfaces for Musical Expression*. Pittsburgh, PA, United States, pp. 116–120.

Gaver, Wiliam W. (1986). "Auditory icons: using sound in computer interfaces". In: *Human-Computer Interaction* 2.1, pp. 167–177.

Geronazzo, Michele, Simone Spagnol and Federico Avanzini (Sept. 2010). "Estimation and modeling of pinna-related transfer functions". In: *Proceedings of the 13th International Conference on Digital Audio Effects*. Graz, Austria, pp. 1–8.

Geronazzo, Michele et al. (2019). "Creating an audio story with interactive binaural rendering in virtual reality". In: *Wireless Communications and Mobile Computing* 2019, pp. 1–14.

Gerzon, Michael (1975). "Ambisonics. Part two: Studio techniques". In: *Studio Sound* 17.40, pp. 24–28.

Gorzel, Marcin et al. (Mar. 2019). "Efficient encoding and decoding of binaural sound with Resonance Audio". In: *AES International Conference on Immersive and Interactive Audio*. York, UK: AES, pp. 1–12.

Hallgren, Kevin A. (2012). "Computing inter-rater reliability for observational data: an overview and tutorial". In: *Tutorials in Quantitative Methods for Psychology* 8.1, pp. 23–34.

Hansen, Villy and Gert Munch (1991). "Making recordings for simulation tests in the Archimedes project". In: *Journal of the Audio Engineering Society* 39.10, pp. 768–774.

Hiipakka, Jarmo and Gaëtan Lorho (July 2003). "A spatial audio user interface for generating music playlists". In: *Proceedings of the 2003 International Conference on Auditory Display*. Boston, MA, USA: ICAD, pp. 267–270.

IFPI (2017). *Global music report: Annual state of the industry*. Tech. rep. London, UK: International Federation of the Phonographic Industry.

— (2018a). *Connecting with music: Music consumer insight report*. Tech. rep. London, UK: International Federation of the Phonographic Industry.

— (2018b). *Global music report: Annual state of the industry*. Tech. rep. London, UK, pp. 1–48.

— (2019a). *Global music report: State of the industry*. Tech. rep. London, UK: International Federation of the Phonographic Industry.

IFPI (2019b). *Music listening: A look at how recorded music is enjoyed around the world*. Tech. rep. London, UK: International Federation of the Phonographic Industry.

— (2021). *Engaging with music*. Tech. rep. London, UK: International Federation of the Phonographic Industry.

— (2022). *Global music report*. Tech. rep. London, UK: International Federation of the Phonographic Industry.

Iwaya, Yukio (2006). "Individualization of head-related transfer functions with tournament-style listening test: Listening with other's ears". In: *Acoustical Science and Technology* 27.6, pp. 340–343.

Iwaya, Yukio, Yôiti Suzuki and Daisuke Kimura (2003). "Effects of head movement on front-back error in sound localization". In: *Acoustical Science and Technology* 24.5, pp. 322–324.

Katz, Brian F. G. and Markus Noisternig (2014). "A comparative study of interaural time delay estimation methods". In: *The Journal of the Acoustical Society of America* 135.6, pp. 3530–3540.

Katz, Brian F. G. and Gaëtan Parseihian (2012). "Perceptually based head-related transfer function database optimization". In: *The Journal of the Acoustical Society of America* 131.2, EL99–EL105.

Kearney, Gavin and Tony Doyle (Oct. 2015). "Height perception in Ambisonic based binaural decoding". In: *Audio Engineering Society Convention 139*. New York, NY, USA, pp. 1–10.

Keating, Peter and Andrew J. King (2013). "Developmental plasticity of spatial hearing following asymmetric hearing loss: Context-dependent cue integration and its clinical implications". In: *Frontiers in Systems Neuroscience* 7.Ded, pp. 1–20.

Kelly, Kevin (2016). *The untold story of Magic Leap, the world's most secretive startup*. URL: https://www.wired.com/2016/04/magic-leap-vr/.

Kilgore, Ryan, Mark Chignell and Paul Smith (Oct. 2003). "Spatialized audioconferencing: What are the benefits?" In: *Proceedings of the 2003 Conference of the Centre for Advanced Studies on Collaborative research*. Toronto, ON, Canada, 135–144.

Kim, Chungeun, Veranika Lim and Lorenzo Picinali (2020). "Investigation into consistency of subjective and objective perceptual selection of non-individual head-related transfer functions". In: *Journal of the Audio Engineering Society* 68.11, pp. 819–831.

Knees, Peter, Markus Schedl and Masataka Goto (2019). "Intelligent user interfaces for music discovery: The past 20 years and what's to come". In: *Proceedings of the 20th International Society for Music Information Retrieval Conference*. Delft, Netherlands, pp. 44–53.

Kronlachner, Matthias (2014). *Plug-in suite for mastering the production and playback in surround sound and Ambisonics*. Tech. rep. Berlin, Germany: Gold-Awarded Contribution to AES Students Design Competition, 136th AES Convention, pp. 1–5. URL: http:

//www.matthiaskronlachner.com/wp-content/uploads/2013/01/kronlachner_aes_studentdesigncompetition_2014.pdf.

Lachaud, Christian Michel and Olivier Renaud (2011). "A tutorial for analyzing human reaction times: How to filter data, manage missing values, and choose a statistical model". In: *Applied Psycholinguistics* 32.2, pp. 389–416.

Laitinen, Mikko Ville and Ville Pulkki (Oct. 2009). "Binaural reproduction for directional audio coding". In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. New Paltz, NY, USA, pp. 337–340.

Lorho, Gaëtan, Jarmo Hiipakka and Juha Marila (Sept. 2002). "Structured menu presentation using spatial sound separation". In: *Human Computer Interaction with Mobile Devices 4th International Symposium*. Ed. by Fabio Paternò. Pisa, Italy, pp. 419–424.

Lorho, Gaëtan, Juha Marila and Jarmo Hiipakka (July 2001). "Feasibility of multiple non-speech sounds presentation using headphones". In: *Proceedings of the 2001 International Conference on Auditory Display*. Espoo, Finland: ICAD, pp. 32–37.

Lutman, M E and R R A Coles (2009). "Asymmetric sensorineural hearing thresholds in the non-noise-exposed UK population: a retrospective analysis". In: *Clinical Otolaryngology* 34.4, pp. 316–321.

McGill, Mark et al. (Apr. 2020). "Acoustic transparency and the changing soundscape of auditory mixed reality". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Honolulu, Hawaii,USA, pp. 1–16.

McGookin, David K. and Stephen A. Brewster (Oct. 2004). "Understanding concurrent earcons: applying auditory scene analysis principles to concurrent earcon recognition". In: *ACM Transactions on Applied Perception* 1.2, pp. 130–155.

McKeag, Adam and David S McGrath (Sept. 1996). "Sound field format to binaural decoder with head tracking". In: *AES 6th Australian Regional Convention*. Melbourne, Australia, pp. 1–9.

McKenzie, Thomas, Damian T. Murphy and Gavin Kearney (2018). "Diffuse-field equalisation of binaural Ambisonic rendering". In: *Applied Sciences* 8.10, pp. 1–17.

McPherson, Andrew, Robert Jack and Giulio Moro (2016). "Action-sound latency: Are our tools fast enough?" In: *Proceedings of the International Conference on New Interfaces for Musical Expression*. Vol. 16. Brisbane, Australia: Queensland Conservatorium, Griffith University, pp. 20–25.

Medonca, Catarina et al. (2010). "On the improvement of auditory accuracy with non-individualized HRTF-based sounds". In: *Audio Engineering Society Convention 129*, pp. 1–8.

Meshram, Alok et al. (Sept. 2014). "P-HRTF: Efficient personalized HRTF computation for high-fidelity spatial sound". In: *2014 IEEE International Symposium on Mixed and Augmented Reality*. Munich, Germany, pp. 53–61.

Møller, Henrik et al. (1996). "Binaural technique: Do we need individual recordings?" In: *Journal of the Audio Engineering Society* 44.6, pp. 451–469.

Morris, Chris (2015). *Is 2016 The Year of Virtual Reality?* URL: https://fortune.com/2015/12/04/2016-the-year-of-virtual-reality/.

Nagele, Anna N. et al. (2021). “Interactive audio augmented reality in participatory performance”. In: *Frontiers in Virtual Reality* 1.February, pp. 1–14.

Nicol, Rozenn (2018). “Sound field”. In: *Immersive sound: The art and science of binaural and multi-channel audio*. Ed. by Agnieszka Roginska and Paul Geluso. 1st ed. New York, NY, USA: Routledge. Chap. 9, pp. 276–310.

Nicol, Rozenn et al. (Apr. 2014). “A roadmap for assessing the quality of experience of 3D audio binaural rendering”. In: *EAA Joint Symposium on Auralization and Ambisonics*. Berlin, Germany: Universitätsverlag der TU Berlin, pp. 100–106.

Noisternig, Markus et al. (July 2003). “3D binaural sound reproduction using a virtual Ambisonic approach”. In: *IEEE International Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems*. Lugano, Switzerland, pp. 174–178.

Ocklenburg, Sebastian et al. (2010). “Auditory space perception in left- and right-handers”. In: *Brain and Cognition* 72.2, pp. 210–217.

Pampalk, Elias, Simon Dixon and Gerhard Widmer (2004). “Exploring music collections by browsing different views”. In: *Computer Music Journal* 28.2, pp. 49–62.

Pelzer, Robert et al. (2020). “Head-related transfer function recommendation based on perceptual similarities and anthropometric features”. In: *The Journal of the Acoustical Society of America* 148.6, pp. 3809–3817.

Picinali, Lorenzo et al. (May 2017). “Comparative perceptual evaluation between different methods for implementing reverberation in a binaural context”. In: *Audio Engineering Society Convention 142*. Berlin, Germany, pp. 1–7.

Politis, Archontis (2015). *Vector base amplitude panning library*. URL: https://github.com/polarch/Vector-Base-Amplitude-Panning.

Pulkki, V. (2001a). “Localization of amplitude-panned virtual sources. II: Two- and three-dimensional panning”. In: *Journal of the Audio Engineering Society* 49.9, pp. 753–767.

Pulkki, Ville (1997). “Virtual sound source positioning using vector base amplitude panning”. In: *Journal of the Audio Engineering Society* 45.6, pp. 456–466.

— (Sept. 2001b). “Evaluating spatial sound with binaural auditory model”. In: *Proceedings of the 2001 International Computer Music Conference*. Havana, Cuba: Michigan Publishing, 73–76.

— (2007). “Spatial sound reproduction with directional audio coding”. In: *Journal of the Audio Engineering Society* 55.6, pp. 503–516.

Rayleigh, Lord (1907). “XII. On our perception of sound direction”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 13.74, pp. 214–232.

Reardon, Gregory et al. (Oct. 2017). “Evaluation of binaural renderers: A methodology”. In: *Audio Engineering Society Convention 143*. New York, NY, USA, pp. 1–6.

Reardon, Gregory et al. (Aug. 2018). “Evaluation of binaural renderers: Multidimensional sound quality assessment”. In: *AES International Conference on Audio for Virtual and Augmented Reality*. Redmond, WA, USA, pp. 1–14.

Rentfrow, Peter J. and Samuel D. Gosling (2003). “The do re mi’s of everyday life: The structure and personality correlates of music preferences”. In: *Journal of Personality and Social Psychology* 84.6, pp. 1236–1256.

Roginska, Agnieszka (2018). “Binaural audio through headphones”. In: *Immersive sound: The art and science of binaural and multi-channel audio*. Ed. by Agnieszka Roginska and Paul Geluso. 1st ed. New York, NY, USA: Routledge. Chap. 4, pp. 88–123.

Roginska, Agnieszka, Thomas S. Santoro and Gregory H. Wakefield (Nov. 2010). “Stimulus-dependent HRTF preference”. In: *Audio Engineering Society Convention 129*. San Francisco, CA, USA, pp. 1–11.

Romanov, Michael et al. (May 2017). “Implementation and evaluation of a low-cost head-tracker for binaural synthesis”. In: *Audio Engineering Society Convention 142*. Berlin, Germany.

Sabine, Wallace Clement (1922). *Collected papers on acoustics*. Cambridge, Massachusetts: Harvard University Press.

Satongar, Darius et al. (2013). *Localisation performance of higher-order Ambisonics for off-centre listening*. Tech. rep. October. London, UK: BBC Research and Development White Paper.

Sawhney, Nitin and Chris Schmandt (2000). “Nomadic Radio: speech and audio interaction for contextual messaging in nomadic environments”. In: *ACM Transactions on Computer-Human Interaction* 7.3, pp. 353–383.

Schönstein, David and Brian F. G. Katz (2012). “Variability in perceptual evaluation of HRTFs”. In: *Journal of the Audio Engineering Society* 60.10, pp. 783–793.

Schörkhuber, Christian, Markus Zaunschirm and Robert Höldrich (Mar. 2018). “Binaural rendering of Ambisonic signals via magnitude least squares”. In: *Proceedings of the DAGA*. 44. Munich, Germany, pp. 339–342.

Seeber, Bernhard U. and Hugo Fastl (July 2003). “Subjective selection of non-individual head-related transfer functions”. In: *Proceedings of the 2003 International Conference on Auditory Display*. Boston, MA, USA, pp. 259–262.

Simon, Laurent S. R., Nick Zacharov and Brian F. G. Katz (2016). “Perceptual attributes for the comparison of head-related transfer functions”. In: *The Journal of the Acoustical Society of America* 140.5, pp. 3623–3632.

Spagnol, Simone (2020). “HRTF selection by anthropometric regression for improving horizontal localization accuracy”. In: *IEEE Signal Processing Letters* 27, pp. 590–594.

Stewart, Rebecca (2018). *Bela On Ur Head - head-tracking for binaural audio with Bela*. URL: `https://github.com/theleadingzero/belaonurhead`.

Stewart, Rebecca and Mark Sandler (July 2011a). “An auditory display in playlist generation”. In: *IEEE Signal Processing Magazine* 28.4, pp. 14–23.

Stewart, Rebecca and Mark Sandler (July 2011b). “The amblr: A mobile spatial audio music browser”. In: *Proceedings - IEEE International Conference on Multimedia and Expo*, pp. 1–6.

— (2012). “Spatial auditory display in music search and browsing applications”. In: *Journal of the Audio Engineering Society* 60.11, 936–946.

Suzuki, Yôiti, Satoshi Yairi and Yukio Iwaya (2007). “Effect of large system latency of virtual auditory display on listener’s head movement in sound localization task”. In: *Proceeding of the 13th International Conference on Auditory Display*. Montréal, Canada, pp. 24–31.

Taylor, Eric (1989). *The AB guide to music theory: Part I*. 1st. London, UK: The Associated Board of the Royal Schools of Music.

The Institute of Sound Recording Surrey (2016). *IoSR Matlab toolbox*. URL: https://github.com/IoSR-Surrey/MatlabToolbox.

Thresh, Lewis, Calum Armstrong and Gavin Kearney (Oct. 2017). “A direct comparison of localisation performance when using first, third and fifth order Ambisonics for real loudspeaker and virtual loudspeaker rendering”. In: *Audio Engineering Society Convention 143*. New York, NY, USA, pp. 1–9.

Thurstone, L. L. (1927). “A law of comparative judgment”. In: *Psychological Review* 34.4, pp. 273–286.

Unity (2019). *Ambisonic audio*. URL: https://docs.unity3d.com.

Vazquez Alvarez, Yolanda and Stephen A. Brewster (2010). “Designing spatial audio interfaces to support multiple audio streams”. In: *ACM International Conference Proceeding Series*, pp. 253–256.

— (2011). “Eyes-free multitasking: The effect of cognitive load on mobile spatial audio interfaces”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Vancouver, BC, Canada, pp. 2173–2176.

Vazquez-Alvarez, Yolanda, Ian Oakley and Stephen A Brewster (2012). “Auditory display design for exploration in mobile audio-augmented reality”. In: *Personal and Ubiquitous Computing* 16, pp. 987–999.

Vazquez-Alvarez, Yolanda et al. (2016). “Designing interactions with multilevel auditory displays in mobile audio-augmented reality”. In: *ACM Transactions on Computer-Human Interaction* 23.1, pp. 1–30.

Wan, Yunhao, Alireza Zare and Kyla Mcmullen (Aug. 2014). “Evaluating the consistency of subjectively selected head-related transfer functions (HRTFs) over time”. In: *AES 55th International Conference: Spatial Audio*. Helsinki, Finland, pp. 1–8.

Warusfel, Olivier (2003). *Listen HRTF database*. URL: http://recherche.ircam.fr/equipes/salles/listen/index.html.

Wenzel, Elizabeth M., Durand R. Begault and Martine Godfroy-Cooper (2018). "Perception of spatial sound". In: *Immersive sound: The art and science of binaural and multichannel audio*. Ed. by Agnieszka Roginska and Paul Geluso. 1st ed. New York, NY, USA: Routledge. Chap. 1, pp. 5–39.

Wenzel, Elizabeth M. et al. (1993). "Localization using nonindividualized head-related transfer functions". In: *The Journal of the Acoustical Society of America* 94.1, pp. 111–123.

Wiggins, Bruce (Sept. 2017). "Analysis of binaural cue matching using Ambisonics to binaural decoding techniques". In: *4th International Conference on Spatial Audio*. Graz, Austria.

Xambó, Anna et al. (Sept. 2018). "Jam with Jamendo: querying a large music collection by chords from a learner's perspective". In: *Proceedings of the Audio Mostly Conference on Sound in Immersion and Emotion*. Wrexham, UK: ACM Press.

Yeoward, Christopher et al. (2021). "Real-time binaural room modelling for augmented reality applications". In: *Journal of the Audio Engineering Society* 69.11, pp. 818–833.

Zotter, Franz and Matthias Frank (2012). "All-round Ambisonic panning and decoding". In: *Journal of the Audio Engineering Society* 60.10, pp. 807–820.