An Investigation into the Application of Machine Learning to Spatial Audio for Immersive Media

Daniel J. Turner

PhD

University of York

School of Physics, Engineering, and Technology

May 2023

Abstract

This thesis investigates the challenges of producing spatial audio for immersive media and the utilisation of machine learning to develop novel methods of spatial audio production. Despite growing interest in immersive technology, there is currently little academic literature that captures the perspectives of practitioners. One of the aims of this thesis is to explore and identify the practices and challenges associated with spatial audio production from the perspective of practitioners. A qualitative study is first presented that identifies key features and challenges associated with spatial audio production for immersive media including the time-consuming nature of sound spatialisation, the lack of available spatial sound effects libraries, and the integration of legacy stereo content into spatial productions. These findings were then used to guide the subsequent research in this thesis.

A proof-of-concept system is presented that utilises visual object detection to locate and classify objects within a simple 2D video. The system suggests candidate sound effects files from the chosen repository and generates stereo panning data for the detected objects. The results demonstrate that whilst the use of computer vision algorithms to search sound effects repositories is possible, more robust search methods are required. Furthermore, the results show that whilst panning information can be accurately derived for individual frames containing multiple objects, a more robust method for tracking objects across frames is required. For scenes containing a single object panning data can be accurately derived across multiple frames.

A novel method of upmixing from stereo to b-format is proposed, which uses deep learning to predict time-frequency directional data which is subsequently used to extract and remap time-frequency components into the target spherical harmonics. Results show that whilst the system can learn a generalised mapping for the time-frequency tiles related to the spectrum of the sound sources, it has difficulty generalising to the ambient noise present within the scenes.

Contents

Declaration of Authorship

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at the University of York or any other University. The bibliography contains proper acknowledgement of all sources.

In addition, I declare that parts of this research have been presented as conference and journal publications during the course of the research degree. The related publications are as follows:

- **Chapter 4**: Daniel Turner, Chris Pike, Chris Baume and Damian Murphy (2022). "Spatial audio production for immersive media experiences: Perspectives on practice-led approaches to designing immersive audio content". In *The Soundtrack* 13:1, pp.73-94
- **Chapter 5**: Dan Turner, Chris Pike, and Damian Murphy (2020). "Content Matching for Sound Generating Objects within a Visual Scene Using a Computer Vision Approach". *In Proceedings of the 148th Audio Engineering Society Convention.*

Acknowledgements

Whilst this thesis represents the culmination of 4.5 years of doctoral study, it also represents the end of a 14 year journey since I first arrived at University, guitar in hand, wondering what kind of career in music/music technology awaited me. Well 14 years, two degrees, a PGCE, a stint as an FE lecturer, and a global pandemic later, and I am now about to embark, much older, but still just as excited, on the next stage of the adventure.

Although I may have typed all the words in this thesis, it would certainly not have taken the form it has without several key individuals.

Thanks first and foremost must go to my supervisor, Prof. Damian Murphy, whose guidance, endless enthusiasm, and ability to always see the positive has been a never ending source of inspiration, especially at those times when my own levels of enthusiasm were somewhat lacking.

Thanks must also go to both my industry supervisors at the BBC, Dr Chris Pike, and Dr Chris Baume, for their contributions in helping shape this thesis and always offering a fresh perspective. I am also grateful to everyone I met at BBC R&D for their friendship, advice, and generally making me feel welcome. I was privileged to have the work in this thesis supported by an EPSRC iCASE doctoral studentship in parternship with BBC R&D. Without this support I would not have been able to take this amazing opportunity to work towards a PhD and for that I will forever be grateful.

Thanks must also go to Dr Bruce Wiggins and Dr Adam Hill of the University of Derby, for reopening the door to the realms of science and engineering. I do not think it an exaggeration to say that returning to Derby to undertake my

MSc changed the course of my life.

I would like to thank everyone at the AudioLab for their friendship, particularly Dr Kat Young who was my only office mate before the world decided we should all work from home. Thanks goes to Andrew Chadwick for this assistance in the collection of the IR dataset that was crucial to the latter parts of my work. To Dr Tom McKenzie, and again Dr Kat Young, for allowing me to clog up our group chat these last few weeks with constant thesis writing related questions. It is very much appreciated. Particular thanks to Simon Durbridge, for going through the mill with me and always being on hand for a friendly, insightful, and reassuring chat. We did it!

Finally, Sophie, for feeding me, keeping me hydrated, and making sure I left the house every so often during what felt like a never ending write up period. This is just as much your achievement as it is mine. I could not have done it without you. Lets go climb some mountains.

A special mention must also go to Prof. Chonkous, for always listening to me complain and never asking too many questions.

This thesis is dedicated to John Burden, my Granddad. I'm sorry you never got to see me finish it.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The last decade has seen a significant increase in the production and availability of both industrial and consumer grade Extended Reality (XR) technologies to facilitate *immersive* experiences such as Virtual Reality (VR), Augmented Reality (AR), Mixed Reality (MR), and 360° videos. These technologies facilitate a wide variety of experiences including, training simulators [1], multi-channel music mixes/soundscape recordings [2], 360° video, and videogame-like first person experiences [3]. This has been accompanied by an increase in the production of 3D spatial audio that, alongside the visuals, aims to deliver to the user a sense of being present within a virtual environment [4]. As such, both the subjective quality and objective accuracy with which IMEs are able to generate the target environments impacts their ability to deliver experiences that are perceived to be realistic or, at the very least, plausible. Spatial, tactile, and auditory accuracy can all be considered as more important for some experiences than other. Within an entertainment context, an IME utilising 3D spatial audio with a high subjective quality and high spatial accuracy may result in increased user uptake and a greater amount of commercial success compared to that with a lower perceptual quality and spatial accuracy. Within a training context, the quality of IME production can, arguably, have a farther-reaching impact. XR technologies

can be used to simulate unusual or potentially dangerous situations whilst also supplementing and improving current real-world training regimes. Examples of this may include surgical training [5], safety critical manufacturing [6], and military training [7]. Therefore, the congruencey between the audiovisual stimuli provided by the XR experience and that delivered by the real-world experience may impact the quality of training outcomes and therefore the length of training that is required.

Although sound design for traditional linear media has well documented practices, workflows, and tools [8, 9], spatial audio production, especially within the context of immersive media, can still be considered a relatively new area of practice with less established methods. There is also a current lack of literature addressing the challenges associated with audio production for immersive experiences from the perspective of those working in the industry. By gaining an understanding of the workflows being developed and the challenges faced by practitioners it would allow for research interventions to be targeted and potential impact maximised.

By their very nature, experiences utilising 360° environments are often more complex and may require many more sound cues or sound generating objects to provide not only a requisite level of plausibility but also to fill the additional visual and auditory space. Due to the environment the user is placed within extending to 360°, spatial audio and sound spatialisation plays an important role in creating an engaging and immersive environment by facilitating the positioning of audio sources around the space. The associated sound design often takes advantage of listener expectation to reinforce a sense of immersion [10], such as sounds appearing from specific directions e.g. footsteps from below, and birds from above. Given the extra complexity involved, the design of soundscapes utilising spatial audio can be seen as a more labour intensive task increasing the workload of the sound design teams.

Additionally, there has been recent increased interest, and progress, in the application of machine learning to audio signal processing tasks such as sound scene classification [11–13], sound event detection and localisation [14–16], audio

effects/production [17–19], and audio synthesis [20–22]. Whilst there has been progress in the development of production pipelines to facilitate the capture of 6-degrees-of-freedom (6DOF) audiovisual content [23], and audio-visual reproduction of real environments [24], there is little in the literature related to applying these techniques within a sound design context for experiences which may utilise virtual representations of existing, augmented, or completely synthesised environments.

As such, the aim of the work presented in this thesis is to identify the challenges in producing spatial audio for immersive media experiences (IMEs) from the perspective of practitioners working in the field, and to investigate where machine learning might be used to develop novel methods of spatial audio production.

1.2 Statement of Hypothesis and Novel Contributions

The hypothesis that forms the motivation for the work presented in this thesis is as follows:

> *Machine Learning approaches can be used to assist in addressing challenges associated with the sound spatialisation pipeline for IMEs.*

Details of the work that has been undertaken to investigate this hypothesis are covered in the rest of the thesis as outlined in this introductory chapter. The research conducted as part of this thesis has produced the following novel contributions:

- A qualitative study exploring the defining features of immersive media experiences as a new experience format and which identifies the challenges associated with its production from the perspective of practitioners.
- An investigation into the use of computer vision to derive audio metadata for source position and to search large scale sound effects repositories.

- An IR dataset for locations on a 50-point Lebedev quadrature captured for 9 stereo configurations, up to fourth-order spherical harmonics, and 32 captures from a rigid spherical microphone.

- The estimation of 360° spatial parameters from a stereo signal using a deep learning approach.

- Two approaches to stereo upmix pipelines that utilise predicted spatial parameters to enable the remapping of components in a 360° space.

1.3 Thesis Outline

The thesis is structured as follows. Chapter 2 introduces the fundamentals of acoustics, digital signal processing and sound field capture which will enable the reader to engage with the material in subsequent chapters. The chapter begins with a description of the physical properties of sound waves and how they propagate through space. This is followed by an overview of the human auditory system and specifically how humans decode localisation cues from the pressure signals received at each ear. The chapter goes on to cover the fundamentals of digital audio signal processing with respect to how sound can be represented and manipulated in the digital domain. This includes an introduction to time, frequency, and time-frequency processing methods, such as convolution, the fast Fourier transform, and the short time Fourier transform that be used for both the analysis and processing of digital audio signals. The chapter concludes by introducing a simple definition and model of a sound field and details a summary of sound field recording and encoding techniques. Particular attention is paid to the importance of spatial sampling resolution and the explores the advantages of encoding to the spherical harmonic domain.

Chapter 3 details the relevant background relating to sound design as applied to immersive media experiences. First, the term immersion is defined within the context of this thesis and is followed an explanation of what therefore constitutes an immersive experience and by extension an immersive media experience.

Different types of immersive media experiences are then defined and discussed taking note of the differences between common categories such as AR, VR, MR, and 360° media as well as detailing the roles sound can play within an immersive media experience, particularly how sound can be used to not only immerse the user, but also to help guide them through the experience. The chapter concludes by exploring spatial audio within the context of immersive media experiences and how the use of spatial audio differs between immersive and traditional media.

Chapter 4 presents a qualitative investigation into the challenges associated with spatial audio production for immersive media experiences (IMEs), from the perspective of those practitioners creating this content. The motivation for the work presented in this chapter is to ascertain how practitioners working within immersive/spatial audio approach immersive media sound design and what challenges are faced that differ from those encountered when designing sound for traditional media. The data collection method and participant selection criteria are detailed as well as an explanation on the use of thematic analysis to interrogate the collected data. The generated themes form the basis of a discussion that draws together common topics that emerge across the themes, and consider both the defining features of IMEs, along with what are perceived to be the main challenges by the participants. From the analysis and discussion of the interview data, several areas of potential research are highlighted and discussed.

Chapter 5 builds on some of the conclusions from Cchapter 4 and details the investigation, development and evaluation of an early stage methodology for deriving audio metadata from objects within a 2D visual scene and using this to facilitate automatic stereo panning and candidate sound effects suggestion. The chapter begins by providing some background on using computer vision to affect audio outcomes. This is followed by a detailed description of the system architecture including the computer vision backend, inter-frame continuity check, object trajectory and panning data derivation, and finally candidate sound effects suggestion using the BBC Sounds Effects archive [25] as the target repository. The performance of the system is assessed and the results discussed, including

limitations and recommendations for further optimisations.

Chapter 6 continues the investigation into machine learning approaches to sound spatialisation, whilst seeking to address another of the challenges highlighted in the results of Chapter 4, specifically, the challenges surrounding the perceived lack of spatial audio sound effects libraries and the integration of legacy stereo content into projects requiring spatial audio. The chapter presents the development of a novel deep learning approach for the prediction of time-frequency spatial parameters from stereo signals, which can then be integrated into a number of different stereo upmix pipelines to facilitate the remapping of frequency components to a 360° space. A novel dataset of IRs is presented which was used to synthesise stereo and First-order Ambisonic sound scenes with which to train the network. The optimisation and evaluation pipeline are described along with details of the baseline, and a description of the proposed architecture is presented. The performance of the model is evaluated and discussed. Finally, two example upmix pipelines are described, within which the proposed model can be integrated, and the potential improvements over current approaches are presented.

This thesis concludes with Chapter 7 providing a summary of the key findings of the work presented and their contribution to the field. The hypothesis is restated along with whether the objectives of the thesis have been met. Areas of future work that have been highlighted throughout the thesis are brought together and considered in more detail and finally, this thesis and its findings are considered within the wider research context and its implications discussed.

1.4 Associated publications

Parts of the work detailed in this thesis have been presented in the following publications:

- **Chapter 4**: Daniel Turner, Chris Pike, Chris Baume and Damian Murphy (2022). "Spatial audio production for immersive media experiences: Perspectives on practice-led approaches to designing immersive audio content".

In *The Soundtrack* 13:1, pp. 73-94

- **Chapter 5**: Dan Turner, Chris Pike, and Damian Murphy (2020). "Content Matching for Sound Generating Objects within a Visual Scene Using a Computer Vision Approach". *In Proceedings of the 148th Audio Engineering Society Convention.*

1.5 Associated Datasets

Datasets that have been collected by the author for use within this thesis.

- Chapter 6 Dan Turner and Damian Murphy (2023) "Dataset of stereo and multi-channel IRs for a 50-point Lebedev quadrature". Available at: 10.5281/zenodo.7990195

Chapter 2

Fundamentals of Sound and Audio Signals

2.1 Introduction

As a significant portion of this thesis is concerned with the capture, manipulation, and reproduction of sound fields and audio signals, it is first necessary to provide details on some fundamental concepts pertaining to these areas. This chapter will provide a foundation on the acoustic theory and signal processing that underpins parts of this thesis and also provide the requisite knowledge base from which to explore how machine learning can be applied to audio domain problems.

2.2 Basic Properties of Sound

2.2.1 Sound Waves

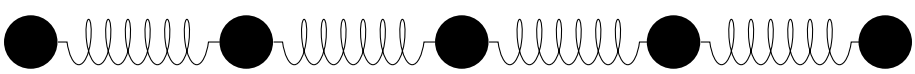


Figure 2.1: The mass-spring model of sound propagation, adapted from [26].

In its simplest form, a sound wave is the displacement of particles, in some medium, from their mean position [27]. The displacement of particles from their equilibrium causes local pressure fluctuations that travel outwards from

the point-of-origin, resulting in areas of high (compression) and low (rarefaction) pressure as molecules in one local area displace molecules in an adjacent area [26, 27]. Sound waves are longitudinal in nature with particle displacement associated with the wave being parallel to the direction of wave propagation. Figure 2.1 illustrates air as a transmission medium using the ball-and-spring model detailed in [26], where the balls represent the molecules of the medium, and the springs represent the inter-molecule forces.

The speed at which the sound wave travels through a medium is dependent on the density and stiffness of the medium and can be represented using the following equation [26]:

$$c = \sqrt{\frac{E}{p}} \tag{2.1}$$

where c is the speed in meters per second ($\mathrm{m \cdot s^{-1}}$), p is the density of the medium ($\mathrm{kg \cdot m^{-3}}$), and E is the Young's modulus (stiffness) of the medium ($\mathrm{N \cdot m^{-2}}$).

For gas, the equivalent to Young's modulus (as gas does not have a Young's modulus) and medium density are derived using the following:

$$E_{gas} = \gamma P \tag{2.2}$$

$$p_{gas} = \frac{PM}{RT} \tag{2.3}$$

Where γ is the adiabatic gas coefficient (1.4 for air), R is the gas constant ($8.31\mathrm{J \cdot mol^{-1} \cdot K^{-1}}$), T is the absolute temperature (in K), and M is the molecular mass of the gas (in $\mathrm{kg \cdot mol^{-1}}$). The speed of sound in a gas is then given by [26]:

$$c_{gas} = \sqrt{\frac{\gamma RT}{M}} \tag{2.4}$$

Equation 2.4 shows that, apart from R and M, which are values specific to the medium, the only factor to affect the speed of sound in gas is the temperature.

Estimating M for air at 2.89 x 10^{-2} $\mathrm{kg \cdot mol^{-1}}$ [28], the speed of sound in air at 20°C can be calculated as follows:

$$c = \sqrt{\frac{\gamma RT}{M}} = \sqrt{\frac{1.4 \times 8.31 \times (20 + 273)}{2.89 \times 10^{-2}}} =\sim 343.4\ \mathrm{m \cdot s^{-1}} \tag{2.5}$$

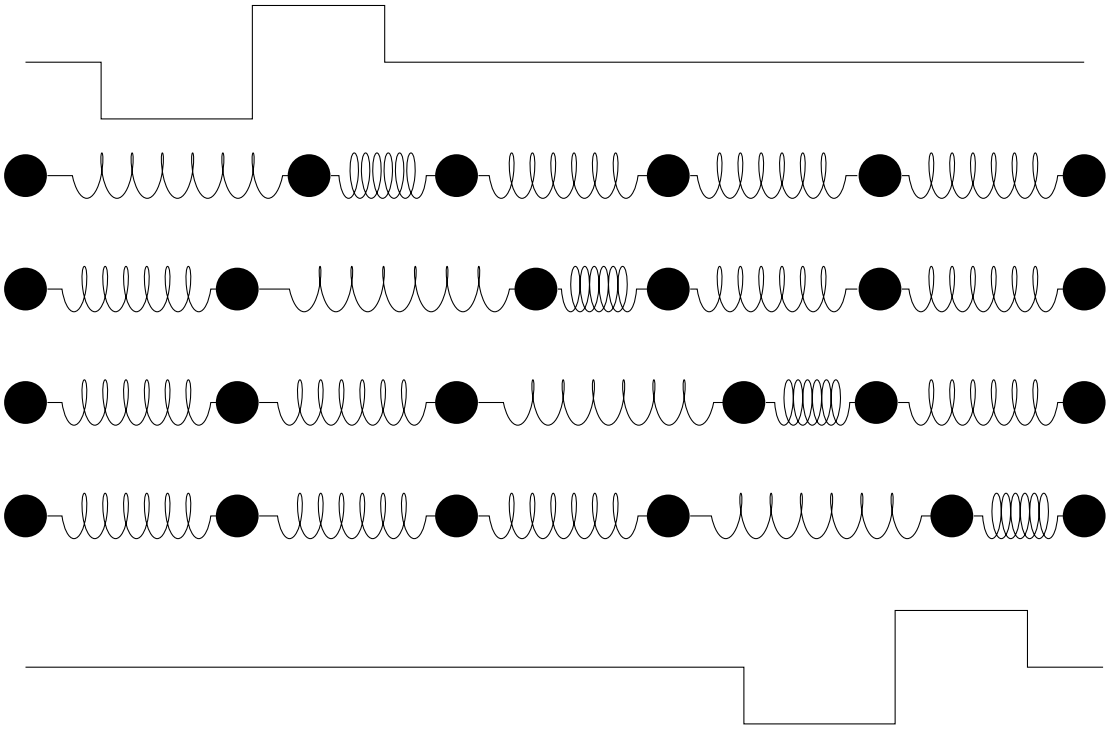


Figure 2.2: The mass-spring model showing propagation of a single sound pulse through a medium, adapted from [26]

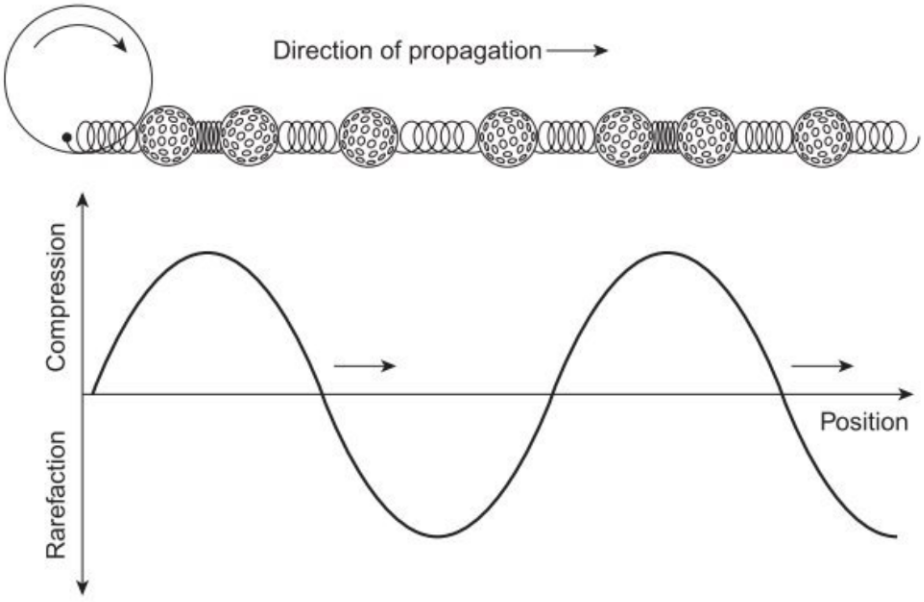


Figure 2.3: The mass-spring model showing propagation of sinusoid through a medium together with its transverse visualisation [from [26]].

2.2.2 Properties of Sinusoidal Signals

The mass-spring model illustrated in Figure 2.2 considers the propagation of a single pulse, however, many of the sounds we hear are periodic in nature. The simplest example of a periodic signal is a sine wave, as it is a vibration at a single frequency. Figure 2.3 shows an example of a sinusoidal signal as an extension of the ball-and-spring model. A sine wave possesses four main properties, some of which are interrelated:

- Wavelength λ (measured in m) is the distance taken to complete one full oscillation, or the distance between points of compression and rarefaction [29].
- Frequency f (measured in Hz) is defined as the number of oscillations per second. Given a constant velocity (e.g. the speed of sound in air) there is an inversely proportional relationship between signal wavelength and frequency [26].
- Amplitude A, often short hand for peak amplitude, represents the maximum change in pressure between points of compression and rarefaction to the medium's equilibrium. This should not be confused with instantaneous amplitude, which can be defined as the value of $x(t)$ at any time (t) [30].
- Phase Φ, which in this context represents the initial phase or phase offset [30].

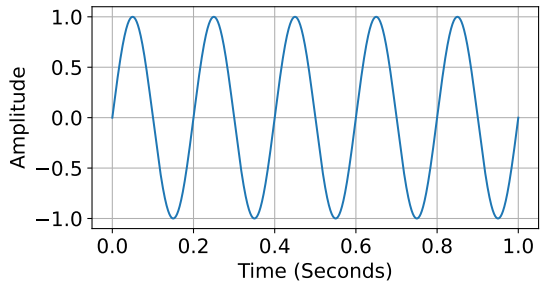
Given the properties outlined above, sine waves can therefore be defined by the equation:

$$y(t) = A\sin(2\pi ft + \Phi) \tag{2.6}$$

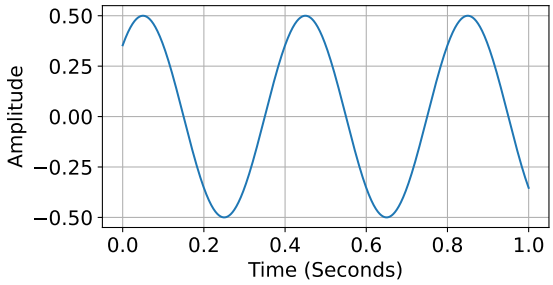
It should also be noted that in the literature, $2\pi f$ will sometimes be substituted for ω, representing frequency in radians as $\omega = 2\pi f$. Figure 2.4 shows two sine waves $A\sin(2\pi ft + \Phi)$, with different values of A, f, Φ, and both with $t \in [0, 1]$.

The relationship between the speed of sound, frequency, and wavelength is given by the following equation:

$$c = f\lambda \tag{2.7}$$



(a) $A = 1$, $f = 5$ Hz, and $\Phi = 0$



(b) $A = 0.5$, $f = 2.5$, and $\Phi = \frac{\pi}{4} rads$

Figure 2.4: Two sinusoids with different values of A, f, Φ and with $t \in [0, 1]$

2.2.3 Sound Propagation

Alongside the properties of a sound wave it is also important to consider how a sound wave propagates through a medium. In Section 2.2.1, it was established that sound waves can be considered as a series of compressions and rarefactions travelling through a given medium. This is possible due to both molecular elasticity and the transfer of momentum from one local group of particles to another [29]. The force required to cause the displacement of particles is the pressure component of a wave and can be defined as the difference between the instantaneous pressure and the static pressure at a given location (x, y, z) [27] given by:

$$p(x, y, z) = \hat{p}(x, y, z) - p_{rest}(x, y, z) \tag{2.8}$$

where p is the resulting sound pressure, $\hat{p}$ is the instantaneous sound pressure, and p_{rest} is the static pressure.

Given the sensitivity of the human ear to sound pressure [26, 31], the overall amplitude or loudness of a sound wave at a given point is measured as the ratio

Sound sources	dB SPL	Pa
Human hearing threshold	0	$2\times$ Pa
Background in TV Studio	20	2×10^{-4} Pa
Quiet library	40	2×10^{-3} Pa
Conversational speech at 1m	60	2×10^{-2} Pa
Busy road at 5m	80	0.2 Pa
Nightclub at 1m from loudspeaker	100	2 Pa
Threshold of discomfort	120	20 Pa
Jet aircraft at 50m	140	200 Pa

Table 2.1: Approximate sound pressure levels and pressure levels of common sounds at specified distances, taken from [32].

of the actual sound pressure p and the threshold of human hearing p_o (a pressure value of 2×10^{-5} Pa [29]). This is quantified as the Sound Pressure Level (SPL), in decibels (dB), on a logarithmic scale [26] and is described by the equation:

$$SPL = 20 \log_{10} \left(\frac{p}{p_o} \right) \tag{2.9}$$

Table 2.1 presents some approximate SPLs of common sounds at specified distances along with their sound pressure values. The relationship between distance and a resulting change in SPL can be expressed by:

$$SPL_2 = SPL_1 + 20 \log_{10} \left(\frac{xyz_1}{xyz_2} \right) \tag{2.10}$$

Where SPL_1 is the SPL measurement at position xyz_1 and SPL_2 is the SPL measurement at position xyz_2. This equation shows that each doubling in distance results in a drop of 6 dB while each halving of the distance results in an increase of 6 dB.

The change in SPL relative to distance is mainly due to fact that sound, in free-field conditions, propagates spherically in three dimensions and as it does its power W, spreads out to cover an ever increasing area as illustrated in Figure 2.5. A measurement that better demonstrates this effect is sound intensity

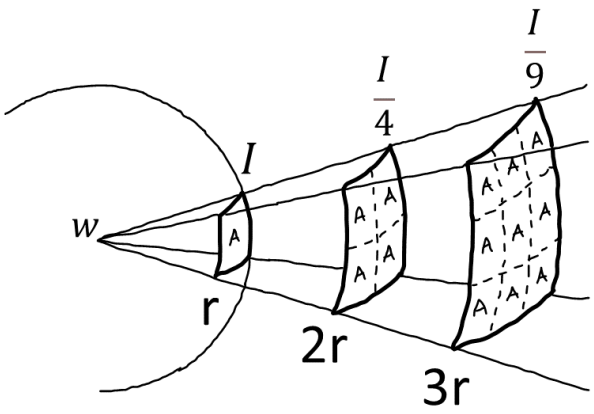


Figure 2.5: Illustration of the inverse square law and how sound intensity I, is inversely proportional to source distance

.

I, measured in watts per square meter ($\mathrm{W{\cdot}m^{-2}}$), as it represents the energy transmitted through a unit area per second and is itself a function of distance, calculated from [26]:

$$I = \frac{W}{4\pi r^2} \tag{2.11}$$

Where I is intensity in $\mathrm{W{\cdot}m^{-2}}$, W is the power of the source in watts, and r is the distance from source (radius).

This equation demonstrates the inverse square relationship between source distance and source intensity measured. Sound intensity, like sound pressure, can also be expressed as a relative measure against the threshold for human hearing, which for intensity is $I_o = 10^{-12}$ $\mathrm{W{\cdot}m^{-2}}$ and can expressed as the Sound Intensity Level (SIL) on a logarithmic scale as:

$$SIL = 10\log_{10}\left(\frac{I}{I_o}\right) \tag{2.12}$$

It should be noted however that this assumes an infinitely small point source exhibiting perfect omnidirectional radiation in a free-field. In reality, sources, no matter how small, have some defined finite area, are rarely perfectly omnidirectional, and real environments are never truly free-field i.e. even sources away

from any rigid boundary (such as the ground) are subject to excess attenuation through atmospheric absorption, which itself varies as a function of frequency and humidity [33]. Given that true free-field environments are rare, sound waves will therefore interact with environment in the form of reflections and diffractions.

2.3 Coordinate Systems

When discussing and working with three-dimensional sound it is important to have clearly defined coordinate systems that can be used to quantify the position of objects within a three-dimensional space relative to a point of origin, usually the listener. There are two commonly used coordinate systems within the context of spatial audio, each suited to a particular context and both transformable to and from each other.

The Cartesian coordinate system (see Figure 2.6) represents three-dimensional space as an ordered set of 3 orthogonal axes that intersect at the point of origin. Positions within Cartesian space are defined according to their location on the x, y, and z axes. This system is often used when viewing the space from an external perspective to that of the listener or subject and is analogous to viewing the scene in third person. The listener is often positioned at the origin facing along the positive side of the *x-axis*, which represents front-back, with y representing the inter-aural axis with positive coordinates to the left and z representing up-down with positive coordinates up.

Unlike the Cartesian coordinate system, which can be used to represent both two-dimensional and three-dimensional spaces, the spherical coordinate system (also see Figure 2.6) specifically represents three-dimensional space where points are positioned according to their azimuth angle θ, elevation angle ϕ, and distance r of the object relative to the listener. Both azimuth and elevation are measured with reference to a fixed point, usually the direction in front of the listener. In this position, both azimuth and elevation will be 0°. The elevation angle has a value range of -90° $\leq \ \phi \ \leq$ 90° with positive above the horizontal plane and negative below it. Azimuth angle value ranges have two conventions: angle values

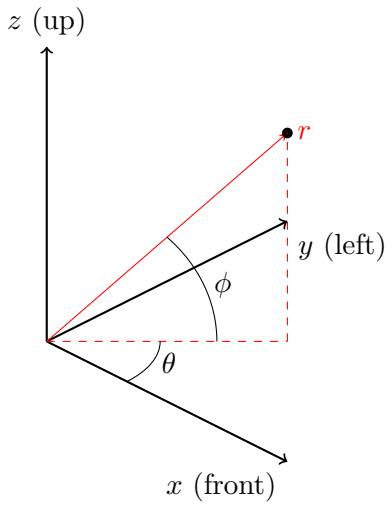


Figure 2.6: Cartesian and Spherical Coordinate Systems

increase with counterclockwise rotation about the $z - axis$ from 0° to 359°, or a range of -180° $\leq$ θ $\leq$ 180°. The spherical coordinate system can be viewed as a more listener-centric coordinate system, analagous to a first person view of the scene, and is more intuitive for listeners orientating themselves within an acoustic environment when compared to the Cartesian coordinate system [34]. Spherical coordinate systems are often useful for systems utilising fixed loudspeaker arrays as an easier way to describe the position of the loudspeakers relative to the listener, where Cartesian coordinates are effective for providing an initial reference frame for objects placed within an environment not necessarily with respect to the user.

Cartesian coordinates can be derived from spherical coordinates using the equations:

$$x = r\cos(\theta)\cos(\phi) \tag{2.13}$$

$$y = r\sin(\theta)\cos(\phi) \tag{2.14}$$

$$z = r\sin(\phi) \tag{2.15}$$

With spherical coordinates being derived from Cartesian coordinates as follows:

$$\theta = \arctan \frac{y}{z} \tag{2.16}$$

$$\phi = \arctan \frac{z}{\sqrt{x^2 + y^2}} \tag{2.17}$$

$$r = \sqrt{x^2 + y^2 + z^2} = \|\mathbf{r}\| \tag{2.18}$$

However, given that IMEs are often viewed from a user-centric perspective and can also include multiple users simultaneously, they also require a coordinate system that describes the relative position and movement of the users within an environment. Additionally, IMEs need to also ensure that the location from which the audiovisual content is viewed is congruent to the expectations of the user. For example, unless an intentional part of the experience, it would be strange for the audiovisual perspective provided to the user was one that had them at ground level, as opposed to head height. For this purpose, it is common for audio objects, and the position of the user, to be represented as allocentric Cartesian coordinates [35], which can then be transformed into user-centric head-related coordinates, shown in Figure 2.7. The visual content seen by the user and the orientation of the spatial sound scene relative to the user can then be manipulated utilising rotational movements around the head-related Cartesian axes by the user looking up/down (pitch), left/right (yaw), or tilting their head side-to-side (roll). The movement of the user within the environment is described by translation movements such as moving backwards/forwards (surging), left/right (strafing), or up/down (elevating). IMEs, therefore, often require the use of both allocentric-centric and user-centric coordinate systems to describe the position of the objects within an environment and the relative position and orientation of the user within that environment.

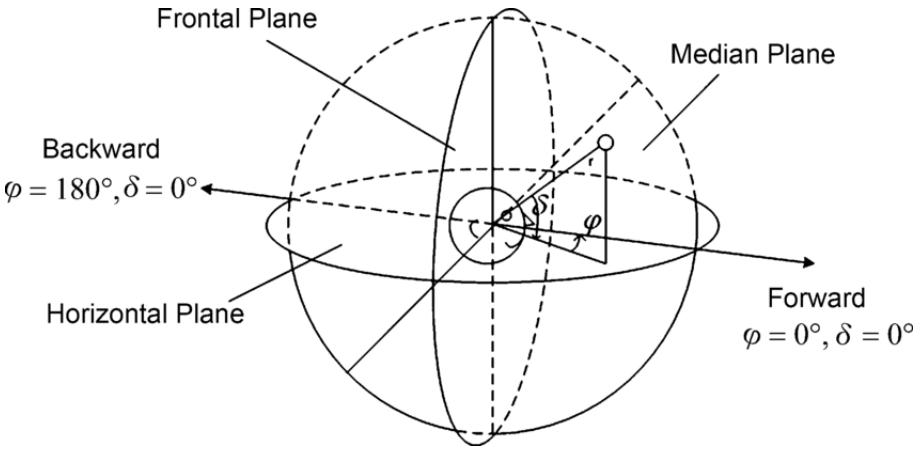


Figure 2.7: Head-related coordinate system, taken from [36]

2.4 Spatial Hearing and Auditory Perception

The human auditory system is remarkably adept at extracting and interpreting complex information about the acoustic environment encoded within sound waves, the physical mechanisms of which were discussed earlier in this chapter. The auditory system processes and decodes information encoded within the changes in sound pressure and allows us to perceive information about the environment, detect objects and activities around us, orient ourselves within the environment, and acoustically communicate with each other through speech or other methods such as music.

This section provides a basic introduction to human auditory perception including a description of the human hearing system, a review of spatial hearing mechanisms, and a brief summary of auditory perception relating to frequency. This section does not aim to be an exhaustive review of the research to date on the human auditory system, but, rather provides the necessary background information relevant to this thesis. For a more detailed and comprehensive overview, see [34, 37–39] for spatial hearing and perception and [40, 41] for a detailed review of auditory distance perception.

2.4.1 Basic Concepts

While this is not a thesis on the philosophy of perception, it is useful for a moment to consider the thought experiment of a tree falling in a forest, and whether, without any hearing-enabled organisms present, it makes a sound. Blauert [37] differentiates between mechanical vibrations that result in the pressure waves discussed previously and what we perceive as a result of those pressure waves interacting with our auditory system. The term *sound event* is used to describe the former, while *auditory event* is used to describe what is perceived auditorily (heard). As such, the hypothetical tree could be said to result in a *sound event*, but not an *auditory event*. This is an important distinction as the information associated with an auditory event is not always congruent to the sound event that it results from. Localisation, by extension, is the process by which an *auditory event* with a location in the auditory space is associated with a sound event in the acoustic environment [34]. The localisation of a sound includes perception of its direction, distance, and extent.

A reason why human (and many non-human animal) auditory systems are so effective is because most possess two ears, one either side of the head, each acting as a data collection point. Binaural hearing is defined as the process whereby the differences in the signals arriving at the two ears are used to resolve the position of a sound relative to the listener. Monaural hearing therefore refers to situations where interaural differences are either not present, or ignored. Although the literature shows localisation cues derived from interaural differences improve our ability to localise sound (see [37, 38] for a detailed review), certain monaural cues are also effective for resolving the positions of sounds. When listening to sounds on headphones, they will often be perceived as originating from inside the head, unless specific binaural processing or recording techniques have been used. The term *lateralisation* is used to describe the localisation of auditory events inside the head as the source has a perceived lateral position, but no associated distance [42].

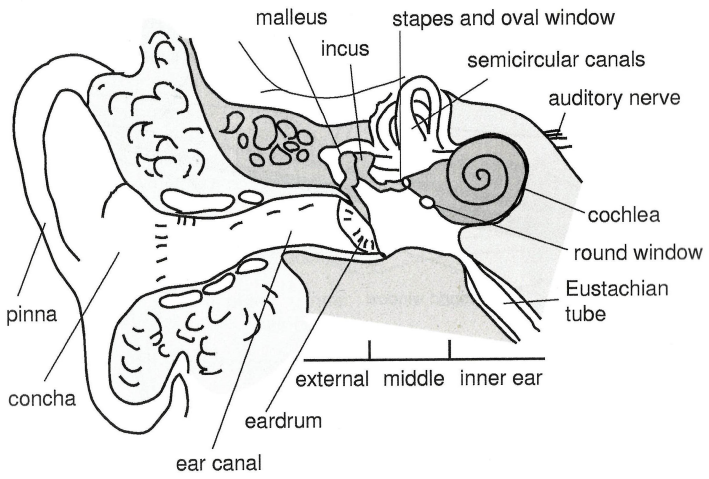


Figure 2.8: The anatomy of the auditory system. Reproduced from [43].

2.4.2 The Auditory System

This subsection briefly discusses the basic elements of the human auditory system; for a more complete review, see [34]. The ear consists of three main sections; the outer ear, middle ear, and inner ear (see Figure 2.8.) The outer ear consists of the pinna, the concha, the ear canal, with the eardrum as a separation point between the outer and middle ear. The outer ear can be described as passive in the sense that it does not itself react to sound or generate sound energy, but instead carries sound waves to the eardrum and the middle ear. The pinna does however affect the incoming sound waves at higher frequencies in a way that assists us in the localisation of sound sources. This and other aspects of spatial hearing are discussed in more detail in Section 2.4.3.

The ear canal (external auditory meatus), due to having one hard boundary (the eardrum) and one bound-unbound boundary (the entrance to the ear canal), acts as a quarter wavelength resonator, which results in a resonance at around 3-4 kHz [26].

The eardrum, or *tympanic membrane*, can be viewed as a signal converter between the outer and middle ear that converts acoustic pressure into mechanical vibrations, which are then passed to the middle ear.

The middle ear is a small air-filled cavity which transmits mechanical vibra-

tions from the eardrum through the ossicles to the oval window. The ossicles are three small bones: the malleus (hammer), incus (anvil), and stapes (stirrup) and the oval window forms the boundary and transmission point between the middle and inner ears. The function of the middle ear is to both transmit the vibrations from the eardrum to the fluid which fills the cochlea and to protect the physical hearing system from the effects of harmful levels of sound pressure. To transmit the vibrations from the middle to inner ear, the ossicles act as a mechanical impedance transformer which transforms a small pressure with larger velocity acting on the eardrum to a high pressure with smaller velocity acting on the oval window. Without this transform, only a minimal amount of the sound energy would proceed from the middle to inner ear due to the much higher impedance of the cochlea fluid relative to air. This process improves pressure transfer by a factor of 30dB [26].

The inner ear consists of the cochlea and semicircular canals. The latter assists in balance and plays no part in the auditory system. The cochlea acts as another signal converter, taking the mechanical vibrations passed from the middle ear and converting them to nerve impulses that are passed to, and processed, by the brain. The cochlea is a coiled structure within which is the basilar membrane. An early study by von Bekesy [44] showed that each point along the basilar membrane resonates at a different frequency, with the area approaching the apex responding to lower frequencies and the area approaching the base responding to higher frequencies. As the cilia are stimulated, they trigger the vestibulocochlear nerve which transmits frequency and temporal information to the brain. This allows humans to perceive differences in timbre and pitch. The range of human hearing is often quoted as being from 20 Hz to 20 kHz, however, due to the structures of the outer and middle ear the sensitivity of the hearing system varies with frequency. The 3-4 kHz resonance of the ear canal results in a greater sensitivity to frequencies within this range. This also happens to cover frequencies within the human vocal range that are important for speech intelligibility. The frequency-dependent sensitivity of the auditory system has been studied extensively [45–47], resulting in the widely recognised

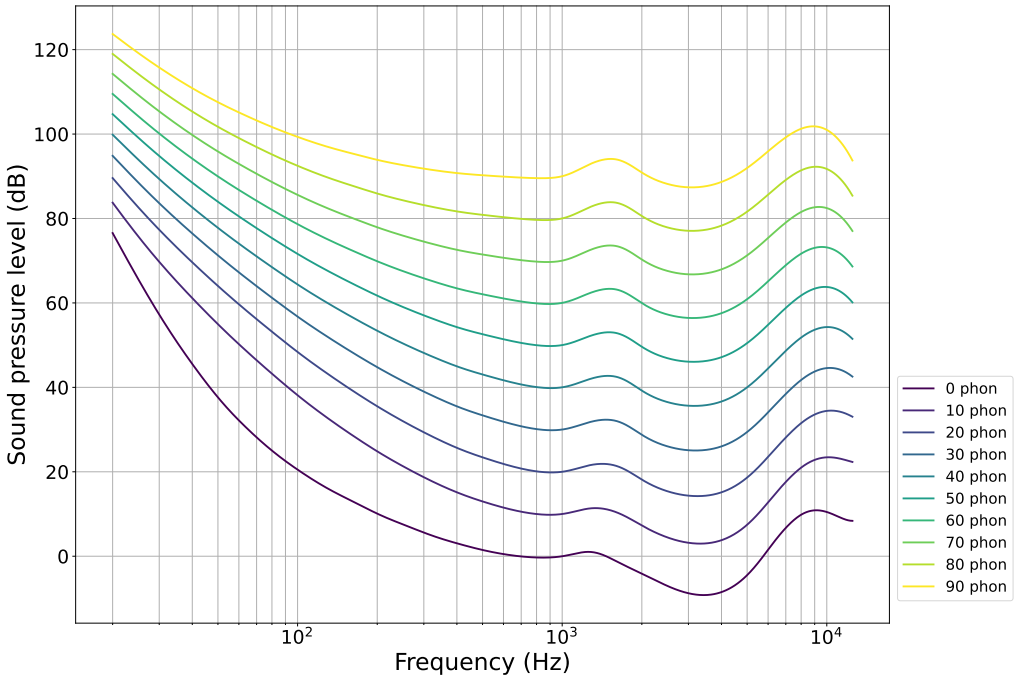


Figure 2.9: ISO 266 equal loudness curves illustrating how the sensitivity of the human hearing system varies as a function of frequency

equal-loudness-curves as shown in Figure 2.9. These curves show the relative SPLs required for tones at different frequencies to be perceived equally as loud as a 1 kHz tone at a particular reference level. It is noted that there is a trough around 3-4 kHz, confirming a heightened sensitivity to frequencies in that range.

Another effect of the physical structure of the human hearing system is a limited ability to resolve spectral components which are close together in frequency. As previously discussed, each point on the basilar membrane resonates according to a specific frequency; however, when a signal causes a particular point to resonate, it also causes an area either side of it to resonate. As such, the basilar membrane will fail to resolve each individual frequency component encoded within the signals presented to the ears for frequencies particularly close together. This may result in some sounds being made inaudible by any other sounds present in the signal; this phenomenon is known as *spectral masking*. The frequency region where the cilia respond strongly to frequencies near their own resonant frequency is known as the *critical band* [48]. The *critical bandwidth*,

is then determined by the minimum difference in frequency required for two sinusoids to be perceived as two separate and smooth tones. Fletcher proposed that the frequency resolution of the ear could be modelled as a bank of bandpass filters, referred to as *critical-band filters*. This led to the development of numerous critical band scales used in the design of auditory filters. The Bark Scale, proposed by Zwicker [49], is derived from listening tests in which a narrow-band noise with a fixed centre frequency is referenced against band-limited noise whose SPL and centre frequency are equal to the reference signal. The bandwidth is then increased until its perceived loudness is greater than that of the reference signal. The relationship with Barks (z) and frequency can be approximated from [50]:

$$z = \left(\frac{26.8}{1 + \frac{1.96}{f}} \right) - 0.53 \tag{2.19}$$

The Equivalent Rectangular Bandwidth scale (ERB) [51] is another critical band scale. The ERB critical bands are estimated by measuring the detection threshold of a sinusoid masked by notched noise. The ERB scale relationship to frequency is given by:

$$R_{ERB} = 21.3 \log_{10} \left(1 + \frac{f}{228.7 \text{ Hz}} \right) \tag{2.20}$$

whilst width of the critical band can be estimated as:

$$\Delta f_{ERB} = 24.7 + 0.108 f_c \tag{2.21}$$

While both the Bark and ERB scales are based on loudness measurements, other scales are measured using alternative metrics. Stevens [52] proposed the *Mel Scale* ('Mel' being short for melody) based on the perception of pitch. The scale was derived by listeners adjusting tones to a specified fraction of a reference tone. For example, if 1 kHz was the reference tone, listeners may be asked to increase a second tone until they perceive it to be half the pitch of the reference. This relationship between frequency and mels is expressed by:

$$mel = 2595 \log_{10} \left(1 + \frac{f}{700} \right) \tag{2.22}$$

Figure 2.10 shows a comparison of the discussed scales using normalised frequency. As discussed in later sections, perceptually motivated frequency scales are widely used in audio signal processing and machine learning.

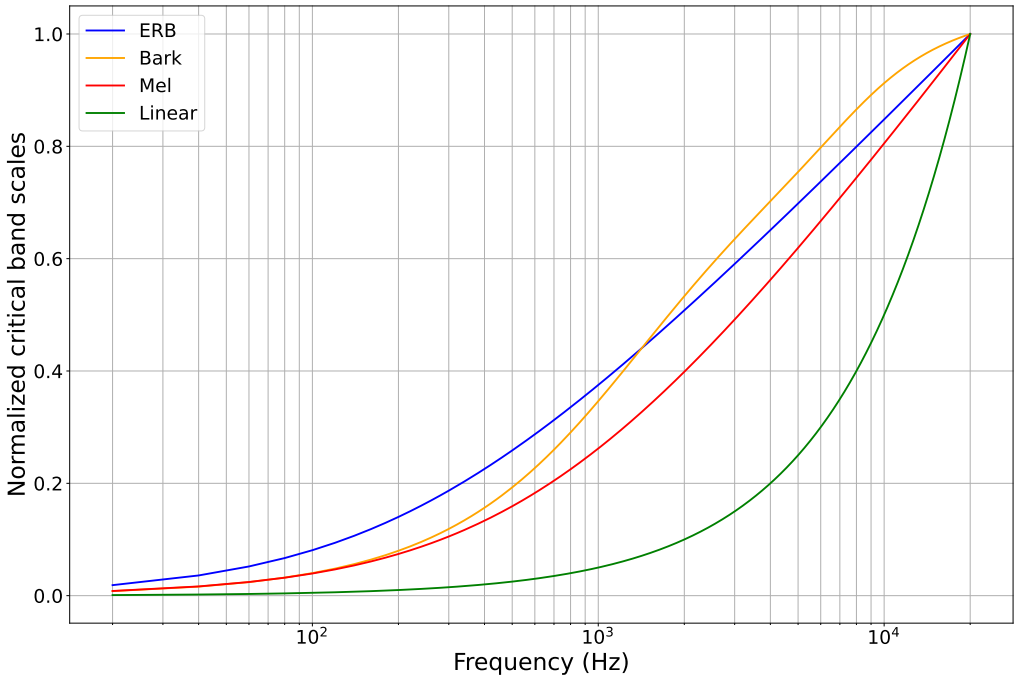


Figure 2.10: Comparison between Bark, ERB, Mel, and linear frequency scales. Units have been normalised.

2.4.3 Directional Localisation Cues

The localisation of an event is guided by a number of different localisation cues, both binaural and monaural [37]. The effectiveness of different cues vary as a function of frequency and direction. Localisation on the horizontal plane is largely informed by the binaural cues that result from time and level differences between the signals arriving at each ear [53]. These cues are referred to as the *interaural level difference* (ILD), the *interaural time difference* (ITD), and the *interaural phase difference* (IPD). It should be noted that whilst ITDs and IPDs can be considered as two different cues, they are not independent of each other as the frequency-dependent IPDs are a result of the frequency-independent ITDs.

To aid the illustration of ILDs and ITDs, consider Figure 2.11, which shows a sound source located at an azimuth of $\theta = 45°$. First, consider that the path between the source and the ipsilateral ear is shorter than that of the source to the contralateral ear. The difference in path length between the source to each of the ears introduces a time difference of arrival between the two ears which results

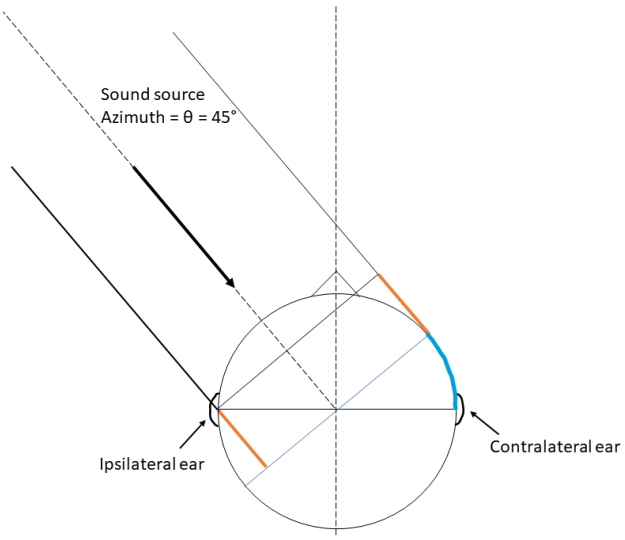


Figure 2.11: Illustration of a simple head model with a source location at 45°. Model highlights additional path length and shadowing introduced by the head for off axis sounds at the contralateral ear. Adapted from [55].

in the ITD. ITDs have been shown to contribute to the localisation of frequencies up to around 1.5 kHz [37, 42, 54]; this is due to frequencies with a wavelength greater than the diameter of the head diffracting around the head. Assuming a perfectly spherical head and $c = 343ms^{-1}$, the wavelength corresponding to the head is as follows:

$$f = \frac{c}{0.175} = 1.96\text{kHz} \tag{2.23}$$

with 0.175 m being the average diameter of the human head according to Kuhn [54].

ITDs can range from 0 μs to approximately 600 - 700 μs [34] with humans being capable of differentiating ITDs as small 10 μs [56, 57]. The maximum ITD is dictated by the maximum path difference between the ipsilateral and contralateral ears, which occurs at 90°, and can be calculated as follows [26]:

$$ITD = \frac{r(\theta + \sin(\theta))}{c} \tag{2.24}$$

$$ITD_{max} = \frac{0.0875 \times (\frac{\pi}{2} + \sin(\frac{\pi}{2}))}{343 \text{ m} \cdot \text{s}^{-1}} \tag{2.25}$$

$$ITD_{max} = 6.56 \times 10^{-4} s = 656 \ \mu s \tag{2.26}$$

While ITDs are predominately a function of angle of incidence (given the assumption of a stable speed of sound and a static head size), IPD, though related, also varies as a function of frequency. Alongside the ITD, the auditory system also uses the IPD caused by the ITDs up to frequencies of around 1.6 kHz [37]. However, given that the phase change due to path length increases as a function of frequency, it becomes less reliable once the phase difference between the two ears is greater than 180°. The relationship between phase shift, frequency, and angle of incidence is given by:

$$IPD = 2\pi fr(\theta + \sin(\theta)) \tag{2.27}$$

IPDs greater than 180° become ambiguous as the ear-brain system struggles to resolve which signal is leading and which is lagging [38]. The predominant reason for this is that the human auditory system is not capable of detecting absolute phase shift and instead compares the relative IPDs. Figure 2.12 illustrates this using a 500 Hz sound wave as an example. Given a 500 Hz sine wave has a period equal to 2 ms, a delay of 0.5 ms would result in a 90° phase shift. If the signal was delayed by 2.5 ms that would equate to a 450° phase shift, however the auditory system would still interpret this as a IPD of 90°. To illustrate that IPD also varies as a function of frequency, the same time delay of 0.5 ms applied to a 1 kHz sine wave would result in a phase shift of 180° (shown in Figure 2.13). ITD cues are also able to be extracted from the delays between temporal envelopes of signals and in some cases have been shown to be perceivable up to 3 kHz [58].

Considering again Figure 2.11, the path to the contralateral ear also has the head as obstacle. At frequencies higher than approximately 800 Hz, the head becomes an appreciable barrier as wavelengths at these higher frequencies become smaller in relation to the head. Rather than diffracting around the head, as is the case with lower frequencies, they are scattered and reflected by the head which results in an acoustic shadowing effect. This causes the level at the ipsilateral ear to be greater than that at the contralateral ear, resulting in ILDs. Therefore, the ILD can be expressed as the difference in SPL between the two ears [59]:

$$ILD = 20\log\left|\frac{Pl(f)}{Pr(f)}\right| \tag{2.28}$$

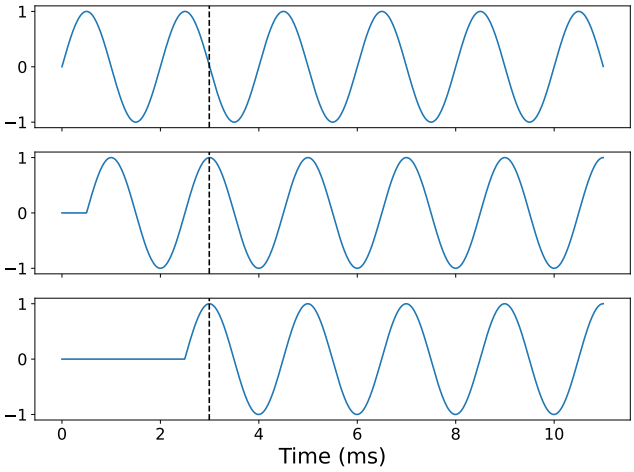


Figure 2.12: Relative phase shift for a 500 Hz sinusoid delayed by 0.5 ms and 2.5 ms. Dotted line indicates that although each delay time represents a phase shift of 90° and 450° respectively, the auditory system would interpret both as 90° as humans are unable to detect absolute phase shift.

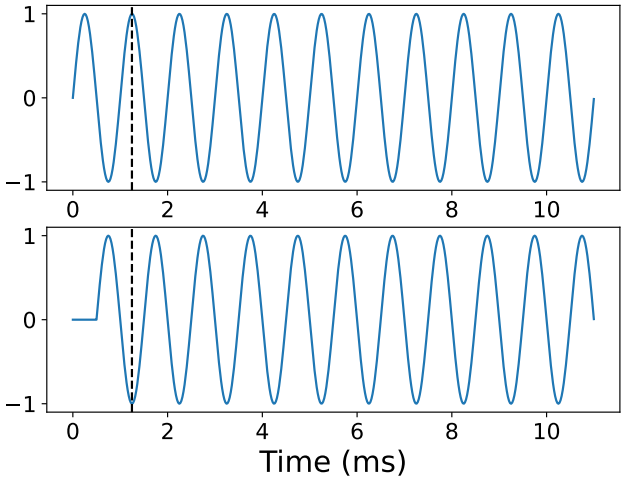


Figure 2.13: Relative phase shift for a 1 kHz sinusoid delayed by 0.5 ms. Dashed line indicates a 180° phase shift, compare this to the 90° phase shift for 500 Hz given the same time delay illustrates that phase shift varies with frequency.

Where $Pl(f)$ and $Pr(f)$ are the Fourier domain representation of the measured sound pressure at the left and right ears respectively.

Since ILDs are present predominately for frequencies where the head is an appreciable barrier to the sound wave, they are only reliable down to the point where head size is equal to roughly $\frac{2}{3}$ of frequency wavelength. Once the head is equal to approximately to $\frac{1}{3}$ of frequency wavelength, this acts much less usefully as a method for resolving direction. For a head with a diameter of 17.5 cm and a source at azimuth $\theta = 90° = \frac{\pi}{2}$, this equates to a minimum frequency of:

$$f_{min(\theta=\frac{\pi}{2})} = \frac{1}{3}\left(\frac{c}{d}\right) = \frac{1}{3} \times \left(\frac{343 \text{ ms}^{-1}}{0.175}\right) = 653 \text{ Hz} \tag{2.29}$$

Therefore, ITDs are predominately utilised for localisation at lower frequencies, while ILDs are utilised at higher frequencies. Within the crossover region of these two cues, our ability to resolve horizontal direction is compromised due to the frequencies being too high for reliable ITD cues and too low for reliable ILD cues [26]. Although used as a directional localisation cue at higher frequencies, Weiping et al. [60] measured *Just Noticeable Differences* (JND) of the ILD as less than 3 dB with sinusoids at frequencies below 2 kHz for base ILD values of close to 0. At higher frequencies and higher base ILDs, the JND is higher, with values ranging from 3 dB to 7 dB. This suggests that at low frequencies ILD cues are primarily used as an auditory distance cue given that the ILD is often greater for nearby sources, especially those within 1 m of the listener.

Both time and level cues present a robust representation of the lateral position of a sound source and, according to the Duplex Theory [53], are all that are required for localisation. However, for each set of interaural difference cues there exists a *cone of confusion* [61], illustrated in Figure 2.14, where for all points on the surface of the cone the interaural cues are theoretically identical, though this assumes a spherical head model and perfectly symmetrical ear positions. Resolution of source direction in these regions is challenging as interaural cues can be considered ambiguous. Common localisation errors resulting from the cone of confusion are front-back and up-down errors where the sound is localised on or near the surface of the cone, but at the wrong location [62, 63].

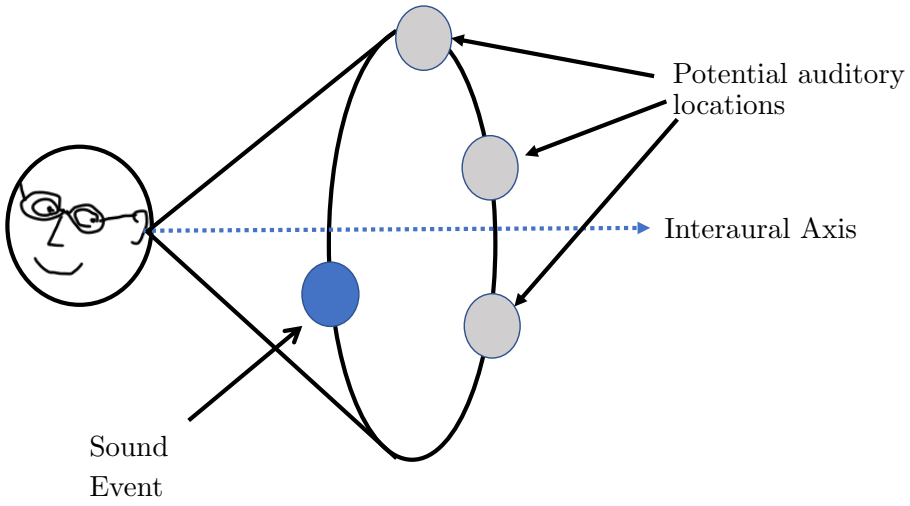


Figure 2.14: Illustration of the cone of confusion where sounds on the surface of the cone have identical interaural differences and may result in localisation errors without additional cues. Adapted from [37].

To resolve the position of a sound source in these regions, spectral and dynamic cues are utilised which are derived from angular dependent filtering and head movement respectively. Spectral cues are a result of reflections from the pinnae introducing delays that range from 100-300 μs [64], which act as an angular dependent filter. Additionally, dynamic changes in head orientation (head movement) alters both the interaural differences and spectral cues resulting in a dynamic change to the interaural transfer function. A number of studies have shown that utilisation of head movement results in a higher localisation accuracy [62, 65–67] with similar results being observed for studies investigating headphone-rendered spatial audio using head tracking [68–70]. Elevation localisation also utilises the spectral cues introduced by pinnae, but also additionally uses the effects of the shoulders and torso, which provide reflections from sources above the horizontal plane and acoustic shadowing for sources below the horizontal

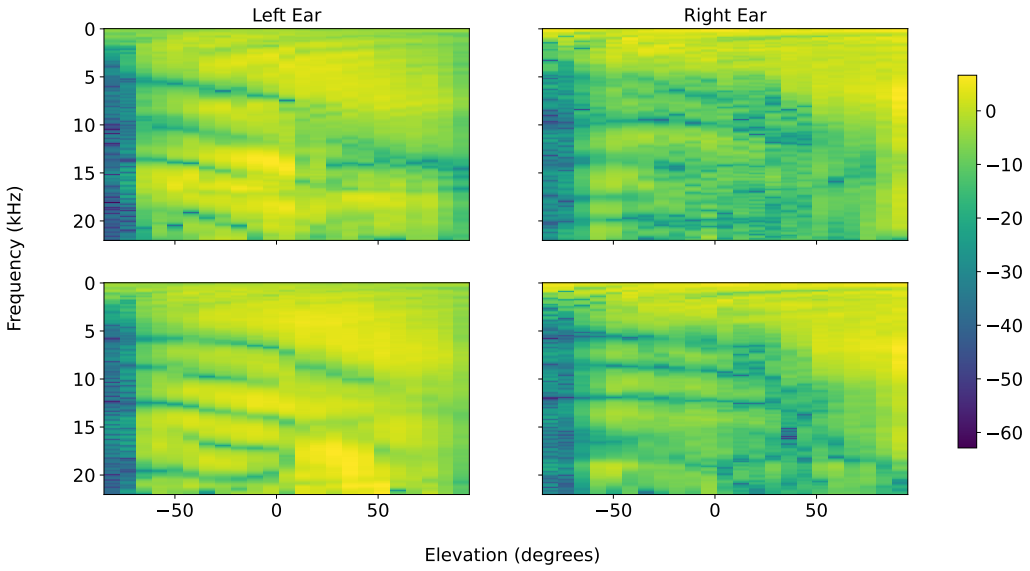


Figure 2.15: Spectral variations vary with elevation angle for 2 different subjects extracted from the SADIE II Database [72] with source azimuth angle of left 45 °and changing elevation. Top: Subject H3. Bottom: Subject H4.

plane. Furthermore, it has been found that cues in specific frequency bands are related to specific positions, with front-back cues located in the region of 8-16 kHz and up-down cues in the region of 5.7-11.3 kHz [39]. For broadband sounds, cues are required in the 4-16 kHz range [71].

As spectral cues are largely determined by the shape of pinnae, which vary between individuals, this results in variations in the spectral filtering unique to each individual. Figure 2.15 shows the spectral variations for a source at azimuth $\theta = 45°$ at varying elevations for two subjects from the SADIE II Database [72]. Particular attention is drawn to the spectral differences between sounds occurring below the listener, compared to those originating from above the listener.

2.4.4 Head Related Transfer Function

The cumulative effects of the head, ears, and torso on a sound wave which results in the acoustic cues used to resolve source position can collectively be represented as time-domain head-related-impulse-responses (HRIRs) or the frequency domain head-related-transfer-functions (HRTFs). For given source and listener positions,

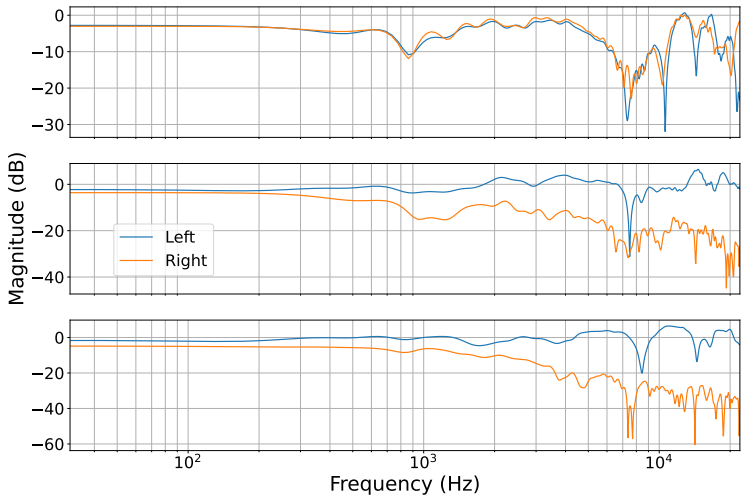


Figure 2.16: HRTF magnitude responses for the left and right ears captured for sources at directions 0° (top), 45° (middle), and 90° (bottom). Derived from data captured from a subject measurement from the SADIE II Database [72].

the HRTF represents the spectral and temporal features of the signals at each ear, as well as the frequency-dependent ILDs and IPDs. As interaural and spectral cues vary as a function of source angle, HRTFs also vary as a function of source position relative to the listener. Figure 2.16 shows a set of HRTFs measured from both ears for three directions. As shown, the HRTFs for both ears are broadly similar to one another for a source at 0°. However, as the source moves across the horizontal plane spectral and level changes occur in both HRTFs. Most notably a drop in level of higher frequencies at the contralateral ear is observed, demonstrating the ILD cue described in section 2.4.3.

Two common methods of HRTF measurement are to either place specialist microphones at the position of the ear drum [73], or to block the ear canal and place a microphone at its entrance [72, 74]. A known signal can then be captured, such as the impulse (or dirac delta) signal δ presented in Section 2.5. Both methods are similar to those employed to capture the impulse response of an environment, which is explored in greater detail in Section 2.6.6. The blocking of the ear canal serves to reduce ear canal resonances and has also been shown to reduce the magnitude variations in measurements between individuals. It

should be noted, however, that the HRTF usually refers to the transfer function of the physical auditory system in isolation without capturing the response of the environment. For this reason HRTFs are normally measured in anechoic conditions. When captured in non-anechoic conditions, they are usually referred to as binaural room impulse responses (BRIRs) as they are a transfer function that collectively encodes both the HRIRs and the room impulse response at a given position [75].

2.4.5 Distance Perception

To completely localise a sound within a space, we not only need to resolve its angular position but also the distance at which it is placed. In general, it has been found that perceived distance tends to be overestimated for sources in peripersonal space (within approximately 1m from the listener) and underestimated for sources in extrapersonal space (farther than 1m from the listener) [41, 76, 77]. Auditory distance judgements tend to be most accurate for sources approximately 1m from the listener [41] and when close sources are positioned laterally relative to the listener [40]. There are multiple cues available for perceiving the distance between a listener and a source, each with varying reliability based upon the distance and direction of the source, the properties of the environment, and the sonic characteristics of the stimulus itself [40]. Kolarik et al. [40], presents two categories of distance cues; absolute cues and relative cues. Absolute cues provide adequate information for distance to be judged from a single presentation of a sound, whereas relative cues allow sounds at different distances to be discriminated.

Section 2.2.3 described the relationship between SIL in a free-field environment being characterised by the inverse square law, where a doubling of source distance results in a 6 dB level reduction. Given this relationship, where the level of a source at a receiver decreases with increased distance and visa versa (assuming consistent level at the source), the human auditory system uses overall level as a relative distance cue [40, 76]. Absolute distance judgements based solely on level are often unreliable given they may be influenced by variation in the level at the

source [41]. It must also be noted that the rate of change of 6 dB per doubling of distance is lessened in reverberant environments, with Zahorik [76] measuring a rate of 4 dB/doubling in an auditorium environment. The radiation pattern of the source and the relative position of the receiver may also effect the rate of change.

A number of studies have also documented differences in the rate of change of perceptual and physical distance when level is the primary cue. Studies by Kearney et al. [77] and Cocran et al. [78] found that perceived distances to a source increased at a lower rate than physical distance increases, while Simpson and Stanton [79] found that sources closer than 1 m require less of a physical change in distance in order for participants to register that the source was moving towards them. This suggests that when level is the primary cue, inverse square law, or the lesser proportional changes that take place in non-free-field environments, may not exactly correlate with our perception of distance changes. One of the reasons suggested for this is that our ability to discriminate relative changes in distance is based on our ability to discriminate changes in sound pressure. Depending on experimental conditions, the recorded thresholds for changes in sound pressure equate to a relative change in distance of between 5% to 25% of the reference distance. Miller [80] observed that, for broadband noise, the smallest detectable change was approximately 0.4 dB and this was observed to be the case 50% of the time for intensities greater than 30 dB above the threshold for hearing. For sinewaves the threshold has been found to be higher at between 1-2 dB and varies with frequency and level [81, 82]. However, results directly measuring the perceptual the threshold for relative changes in distance have often been found it to be much higher with observations of 13% [79], 20% [83], and 25% [84] of the reference distance, although a study by Ashmead et al. [85] found that changes in distance for a white noise burst to be approximately 6%.

Begault [86], however, argues that our perception of distance is better predicted by perceptual loudness than by objective intensity. Results showed that when presented with four different level increases (3 dB, 6 dB, 9 dB, and 12

dB) and asked to select the preferred level increase that for a given reference corresponds to a halving of the perceived distance, 69% of participants chose either a 12 dB or 9 dB increase when compared to a 6 dB increase. This aligns with the generally accepted phenomena that a level increase of 10 dB equates to a doubling of perceived loudness [39]. This concurs with observations that perceptual distance increases at a lower rate than physical distance when the level reduces according to inverse square law.

Whilst level/intensity can be effective as a relative cue, it is often unreliable when making judgements relating to absolute distance. This is because the level experienced at the ears is dependent not only on distance, but also on the acoustic power and radiation pattern of the source. However, it is not usually the case that a sole increase in level would be mistaken for a change in distance. In many situations, multiple cues are often available from which we can derive a distance estimate. Studies by Zahorik & Wightman [87] and Altmann et al. [88] propose that the acoustic power of a sound source is estimated from the reverberant sound energy and, as this remains approximately constant across distance within indoors environments, loudness judgements for sound sources with a fixed acoustic power remain consistent at variable distances. Altmann et al. [88] showed that loudness consistency was generally observed in a room with strong reverberation (RT60 = 1.03 s) but not in a room with weak reverberation (RT60 = 0.14 s), although distance judgements were found to be similar across both environments.

Alongside facilitating loudness judgements, reverberant energy also aids in distance judgements in the form of the direct-to-reverberant energy ratio (DRR) at the ears [40, 76, 89, 90], which decreases as a function of increasing distance. Given a large enough room, reverberant energy is considered diffuse and as such maintains constant energy irrespective of source location. In a small auditorium, as used by Zahorik [89], it was observed that the level of the reverberant sound reduced by around 1 dB for each doubling of source distance. The magnitude of the reverberant energy is dependent on room dimensions, objects within the room, and the absorption properties of the walls, floor, and ceiling.

In contrast to level cues, DRR has been shown to act as an absolute distance cue [91] utilising a frequency dependent, but direction independent, DRR-to-distance mapping based on the DRR at the ipsilateral ear [92]. Distance judgements have been found to be most accurate in situations where both level and DRR cues are available [93, 94], which should be unsurprising given that most natural environments will provide both of these cues and humans would have evolved to use a combination of the two. Much like level change discrimination, the ability to detect changes in DRR has also been shown to contribute to judgements relating to changes in distance. Sensitivity to changes in DRR depend on the reference DRR value [95] and has been shown to be highest around the critical distance, which is the distance at which the direct and reverberant signals have equal energy and where DRR = 0 dB [96]. Sensitivity to DRR changes was observed to be lower for both high and low DRR values, which equates to scenarios where sound sources are considerably closer to, or farther away, from the listener. An investigation by Zahorik [89], however, found that sensitivity to DRR changes were approximately equal across a range of positive DRR values, with JND's of 5-6 dB for values between 0 and 20 dB for stimuli consisting of speech, noise, and frontally and laterally presented impulses. As Kolarik [40] notes, this discrepancy may be attributed to differences in experimental procedure and stimuli. In reverberant environments, the level and time of arrival of early reflections can also provide information on a sources distance [59]. For example, sources closer to the listener will result in a greater initial time delay between the direct sound and the early reflections.

For sounds within 1m of the listener and for sounds farther than 15m from the listener, spectral content can also be used to inform distance judgements [40]. For sounds within approximately 1 m of the listener this is due to the frequency and distance dependent diffraction of sound around the head. In a study by Brungart [97], participants judged the distance of proximal sounds in anechoic conditions for broadband (0.2 - 15 kHz), high-passed (3 - 15 kHz), or low-passed (0.2 - 3 kHz) noise bursts. Results showed that accurate distance judgements required spectral components below 3 kHz [97]. A similar study by

Kopco and Shinn-Cunningham [92] obtained distance judgements for sounds between 0.15 and 1.7 m using noise bursts that varied in centre frequency between 300 Hz and 5700 Hz and in bandwidth between 200 Hz and 5400 Hz. It was found that as low frequency energy was removed from the stimuli the accuracy of distance judgements decreased for both frontal and laterally presented sounds. The variable bandwidth was shown not to affect the mean distance judgements. The results also support those presented in [97], that judgements were relatively accurate for stimuli containing energy at frequencies around 300 Hz and were accurate for stimuli with energy only at 5700 Hz, supporting the conclusion that it is the low-frequency cues provided by diffraction around the head that aid in distance perception at close distances.

For sound sources farther than approximately 15 m from the listener, the spectral content is predominately altered by air absorption with high-frequencies undergoing greater attenuation than low-frequencies [33]. Sounds with decreased high-frequency content relative to low-frequency content are often perceived as being farther away [98]. A study by Butler [99], utilised recorded broadband, low-pass, and high-pass noise in the ear canals of humans in an anechoic or reverberant room. These were then used a stimuli and played back to participants over headphones. For both anechoic and reverberant conditions the low-pass filtered noises were consistently judged as being at a greater distance from the participants. The broadband noise was perceived as being in the middle of the range of perceived distances. However, a similar study by Little et al. [98] argued that spectral differences in the stimuli used in [99] could not be produced by physical changes in distance to a sound source and therefore lacked ecological validity. Little utilised shaped broadband noises low-pass filtered at 5, 6, and 6.7 kHz, arguing that these spectral differences were more akin to those that could be caused by changes in source distance. The results, however, did concur with previous findings that decreases in high-frequency energy were associated with greater perceived distances, but only after several trials, which suggests that spectral content is a relative distance cue.

Related to, but often treated as distinct from, distance perception is the notion

of externalisation. The term externalisation is commonly used with reference headphone reproduction. When listening to sources in a real environment there is almost always an inherent distance between the listener and the source, which results in the source having a perceived distance and being perceived as external to the listener [100]. However, listening via headphones confounds this since the signals are being delivered directly to both ears at a very close proximity and thus bypasses the filtering properties of the head and ears [100]. For traditional amplitude panned stereo, this results in signals that contain the lateral cues outlined in Section 2.4.3, but lack the spectral cues required to facilitate a perception of distance, which in turn results in an *in head* listening experience [101].

Externalisation, therefore, is often used to describe the ability of an audio reproduction system (usually headphones) to deliver the cues necessary to deliver the perception of distance and thus cause the sound to appear as if it is external to the listener[101]. In this sense, externalisation can be considered as an extension of distance perception, but within the context of audio reproduction systems where distance cues are modelled through signal processing. Lastly, whereas directional localisation cues can be viewed as a result of the effect our anatomy has on incident sound waves, auditory distances cues, predominately, result from the effect the environment has on the temporal and spectral characteristics of the signals reaching the ears.

2.5 Audio Digital Signal Processing

2.5.1 Audio Sampling

Consider again the sinusoids represented by Equation 2.6 and illustrated in Figure 2.4. Both signals are said to be measured in *continuous-time* as the signal is observable for any $t \in \mathbb{R}$ or the specific case of Figure 2.4, $t \in [0, 1]$. However, to represent this signal digitally, the amplitude of the signal must be sampled at regular intervals in time. This is due to the finite precision inherent in digital systems. Once the signal has been sampled at a discrete set of N time points it

is now represented by a discrete set of values and is referred to as a *discrete-time signal*. Rewriting Equation 2.6 for a *discrete-time* signal gives us:

$$y[n] = A\sin(2\pi fn + \phi) \tag{2.30}$$

Instead of t for a continuous time value, we have n for the nth sample.

The conversion from a *continuous-time* signal to a *discrete-time* signal is carried out by an analogue-to-digital converter (ADC). Conversely, discrete digital signals can also be converted into continuous analogue signals by digital-to-analogue converters (DAC). The analogue-to-digital conversion is usually done through methods such as pulse-code modulation [102]. The number of samples taken per second is determined by the sampling frequency (also referred to as the sampling/sample rate), f_s with the *Nyquist theorem* stating that to accurately sample a frequency f, f_s must be at least $2f$. Therefore, $f_s \geq 2f_{max}$ where f_{max} is the highest frequency to be sampled. If a 1 Hz sinusoid is sampled at $f_s = 30$ Hz the digital approximation would be that depicted in Figure 2.17. As seen, rather than having continuous values in times the signal now consists of discrete points of data at regular intervals in time. Each sample is converted into a binary number which represents its amplitude and results in the sample being *quantised* to the closest available value. The number of bits assigned to each sample, known as bit depth, determines the precision at which the ADC can map the continuous analogue level to discrete digital values with each bit $nbits$ resulting in 2^{nbits} discrete amplitude values. Within the context of ADCs, the sampling frequency determines precision in time, and bit depth determines precision with respect to amplitude. For example, a system with 8 bits corresponds to 256 discrete values, whereas 16 bits (the common bit-depth for music on Compact Disc) provides 65,536 discrete values.

2.5.2 Impulse Response

The unit impulse, or the Dirac delta function $\delta(t)$ (continuous time) or $\delta[n]$ (discrete time) is a signal that theoretically contains energy at all frequencies

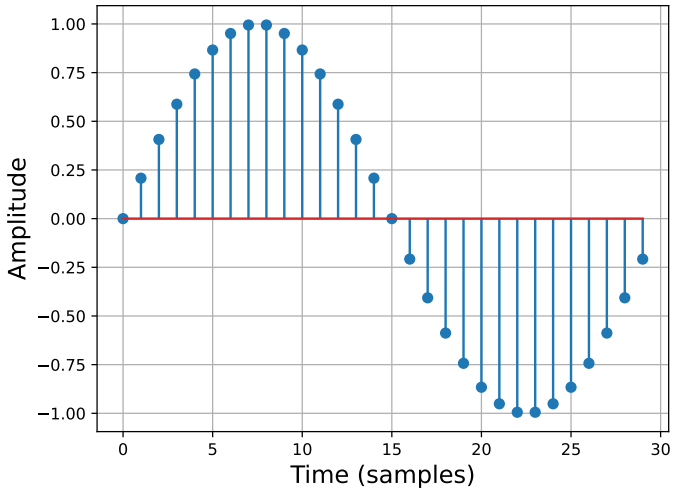
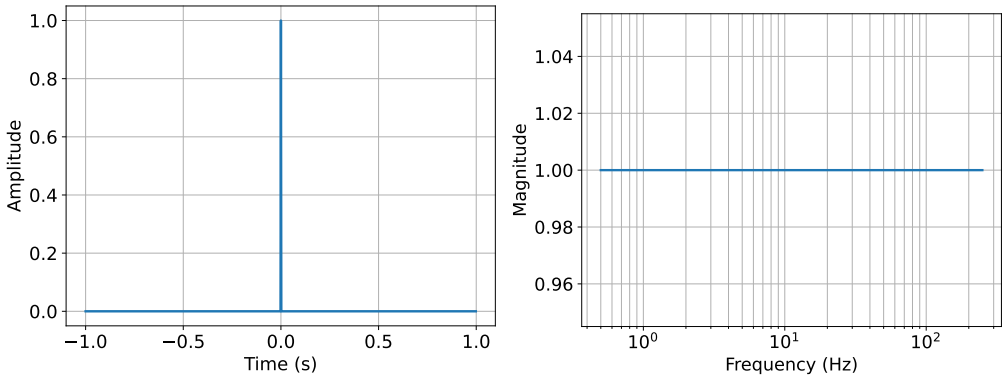


Figure 2.17: Illustration of a 1Hz sinusoid digitally sampled at 30 Hz

when t or $n = 0$ and no energy at all other times. It can be represented as:

$$\delta(n) \triangleq \begin{cases} 1 & n = 0 \\ 0 & n \neq 0 \end{cases} \tag{2.31}$$

Figure 2.18a shows the discrete-time unit impulse and Figure 2.18b shows the associated frequency-domain representation obtained through the FFT of $\delta(n)$.



(a) Time-Domain

(b) frequency-domain

Figure 2.18: Time and frequency-domain representations of the Unit Impulse

A linear time invariant system can be represented by its response to $\delta[n]$. The time-domain representation is the *impulse response* $h(t)$ or $h[n]$ and the frequency-domain representation is the *transfer function* $H(\omega)$ (continuous-frequency)

or $H(k)$ (discrete frequency). Recall previously that the HRTF refers to the frequency-domain representation and the HRIR the time-domain representation. For a static source and head position, the physical hearing system can be considered a linear time invariant system.

The impulse response $h(n)$ can be represented as [30]:

$$h(n) \triangleq \mathcal{L}_n\{\delta(\cdot)\} \tag{2.32}$$

Where $\mathcal{L}$ is the linear-time invariant system at time n and $\delta(\cdot)$ is unit impulse occurring at time 0.

The generation of a unit impulse for the measurement of digital systems, such as digital filters, is a trivial task. However, the generation of a unit impulse for the measurement of audio systems and acoustic spaces, such as an indoor or outdoor environments or audio reproduction systems, is not so straight forward. In practice, it is not possible to reproduce a perfect unit impulse using conventional methods. This is due to the restraints of loudspeaker technology in producing instantaneous impulses with high enough power. Impulse-like sounds can be generated using transient producing sources such as a starter pistol [103], balloon pop, or hand claps, however, these are acknowledged to be less than ideal.

Common approaches which have greater precision and reproducibility are methods that utilise a known excitation signal that is finite in length and is known to contain the full spectrum of frequencies of interest. Early methods included the *maximum length sequence* [104] and *inverse repeat sequence* [105], which use pseudo-random noise as the excitation signal followed by a circular cross-correlation to retrieve the impulse response. Both methods suffered from distortion evenly spread throughout the IR resulting from the imperfect loudspeaker reproduction. One solution was to lower the playback level, which in turn increased the signal-to-noise ratio with respect to the excitation signal (signal) and the artefacts caused by the imperfect loudspeaker reproduction (noise). This may, however, cause a decrease in the signal-to-noise ratio with respect to the excitation signal and the background noise present in the environment.

A more recent method proposed by Farina [106] suggests the use of an

exponential sine sweep (ESS) as the excitation signal, which can be defined as:

$$\vartheta(t) = \sin\left[\frac{\omega 1 \times T}{ln\left(\frac{\omega 2}{\omega 1}\right)} \times \left(e^{\frac{t}{T} \times ln\left(\frac{\omega 2}{\omega 1}\right)}\right) - 1\right] \tag{2.33}$$

Where $\omega 1$ and $\omega 2$ are the start and end frequencies respectively in radians and T is the duration of the sweep.

IRs are extracted through the convolution of the measured signal $y(t)$ with inverse filter for the ESS $\vartheta(t)$:

$$h(t) = y(t) * i_{\vartheta}(t) \tag{2.34}$$

Doubling the duration of the sweep will increase the signal-to-noise ratio by approximately 3 dB. Furthermore, once the "deconvolution" process has extracted the IR any harmonic distortion artefacts present are grouped and appear prior to the start of the "main" impulse in the form of smaller impulses. These are then able to be removed by simple truncation of the signal. The ESS has been shown to improve signal-to-noise ratio when compared to previous methods and perform better in quieter environments [106, 107]. However, though longer sweeps result in an improved signal-to-noise ratio, the risk of interference from other sound sources present in the environment is increased.

2.5.3 Convolution

As the IR represents the response of a system to $\delta[n]$ system responses to any input can be defined by the convolution of an input signal $x[n]$ with the IR $h[n]$ given by [34]:

$$y[n] = x[n] * h[n] + \zeta[n] \tag{2.35}$$

Where $y[n]$ is the measured signal, $x[n]$ is the input signal, $\zeta[n]$ is the noise present in the system, and $*$ denoting the convolution operator, which can further be defined for discrete time as the *convolution sum*:

$$y[n] = \sum_{m=0}^{n} h(m)x(n-m) \tag{2.36}$$

Equation 2.36 also shows that the resulting output from a convolution operation will be of length $M + N - 1$, where M is the length of the IR and N the length of the input signal.

Within the context of audio signal processing, convolution is often used to impart the response of a given system (equalisation filter, room response, BRIR, HRIR) onto a given input signal. Figure 2.19 shows an example convolution of two signals. In this thesis it is extensively used in the synthesis of stereo and B-format sound scenes, explained in greater detail in section 2.6. It is also important to note that convolution in the time-domain is equivalent to multiplication in the frequency-domain; as such, Equation 2.36 can be represented as the frequency-domain operation:

$$Y(\omega_k) = H(\omega_k)X(\omega_k) \tag{2.37}$$

where $Y(\omega_k)$, $H(\omega_k)$, and $X(\omega_k)$ are discrete frequency domain representation of $y[n]$, $h[n]$, and $x[n]$ respectively.

2.5.4 Spectral Analysis

Until this point, sound has been discussed predominately as a function of time and/or containing a single sinusoidal frequency component. In reality, most naturally occurring sounds are far more complex than sinusoids. However, Fourier proposed a theorem that any periodic signal $f(t)$, no matter the complexity, can be modelled as a combination of sinusoids of varying frequencies, amplitudes, and phases. This is mathematically represented as [26]:

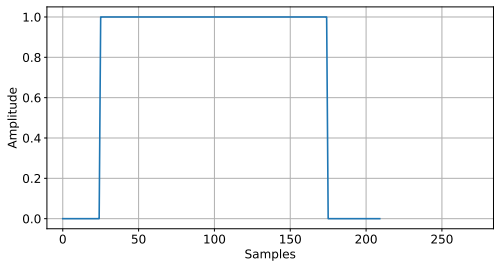
$$f(t) = a_0 + \sum_{n=1}^{\infty} a_n \cos(n\omega_0 t) + b_n \sin(n\omega_0 t) \tag{2.38}$$

Where a_0 is the d.c. offset of the signal, a_n and b_n are the level/amplitude of the nth cosine and sine harmonics respectively, and ω_0 is the angular frequency ($2\pi f_0$). The set of sinusoids that make up a periodic signal are called the Fourier series with all harmonics being integer multiples of the fundamental frequency ω_0.

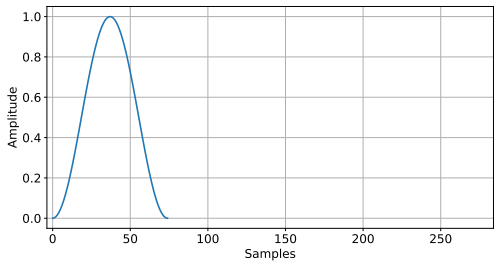
The use of Euler's formula allows both the sine and cosine to be combined in the form of a complex exponential:

$$e^{j\theta} = \cos(\theta) + j\sin(\theta) \tag{2.39}$$

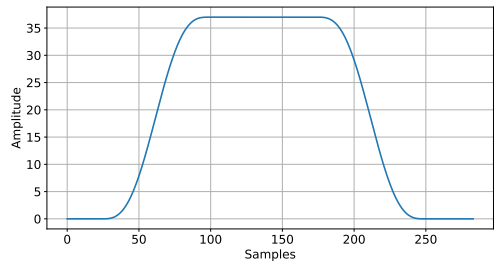
where θ is in radians.



(a) $h[n]$



(b) $g[n]$



(c) $h[n] * g[n]$

Figure 2.19: The Resulting output ($h[n] * g[n]$) of the two convolution of signals $h[n]$ and $g[n]$

This allows the re-expression of the Fourier series as a complex exponential [26]:

$$f(t) = \sum_{n=-\infty}^{\infty} C_n e^{jn\omega_0 t} \tag{2.40}$$

where C_n are the complex coefficients that describe the two previous coefficients a_0 and b_0. The absolute value of C_n, $|C_n|$ represents the magnitude of the nth frequency component while the angle, $\angle C_n$ represents the phase of the nth frequency component.

The summation of harmonic sinusoids to construct a signal is termed Fourier synthesis. Providing the values for a_n and b_n are known this can be used to represent any periodic signal. However, to represent a signal completely may require an infinite number of harmonics as suggested by Equation 2.38. Furthermore, different signals will require different values of a_n and b_n; for instance the Fourier series coefficients for a square wave are defined as:

$$a_0 = 0 \tag{2.41}$$

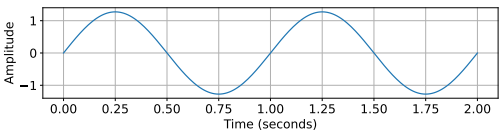
$$a_n = 0 \tag{2.42}$$

$$b_n = \begin{cases} 0, & \text{if } n \text{ is even} \\ \frac{4}{n\pi}, & \text{if } n \text{ is odd} \end{cases} \tag{2.43}$$

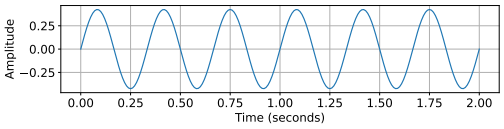
Figure 2.20 shows the fundamental frequency f_0 and the first three harmonics of a square wave along with the result of the summation of these first four components. This is compared with a square wave containing 50 harmonics alongside an idealised version wave that would result from the summation of an infinite number of additional harmonics. This demonstrates the effects of not possessing enough coefficients to accurately model a given signal.

Alongside Fourier synthesis, which allows the creation of a waveform from known Fourier coefficients, Fourier analysis allows for the measurement and extraction of the individual frequency components, $F(\omega)$ of a given periodic signal. In practice, not all signals are periodic, however aperiodic signals can be treated as periodic signals with a finite length [108]:

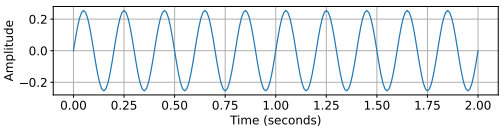
$$F(\omega) = \int_{t=0}^{T-1} f(t) e^{-j\omega t} dt \tag{2.44}$$



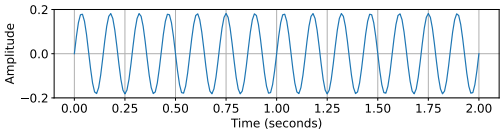
(a) Fundamental frequency



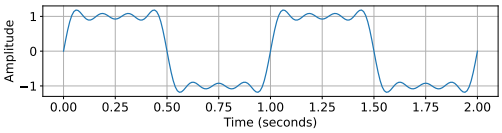
(b) 3rd harmonic



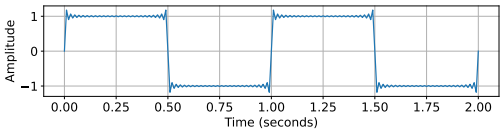
(c) 5th harmonic



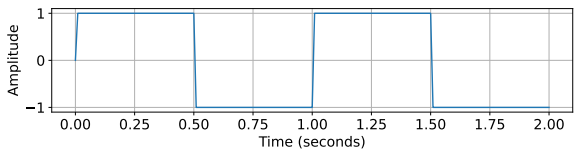
(d) 7th harmonic



(e) Resultant Waveform from the first 4 partials of a square wave



(f) Square wave resulting from 25 partials.



(g) Idealised square wave

Figure 2.20: Fourier synthesis of a square wave, showing the first four partials, the resulting waveform from their summation, the waveform resulting from 25 partials, and a idealised square wave.

With the discrete Fourier analysis being represented as:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{\frac{-j2\pi nk}{N}} \tag{2.45}$$

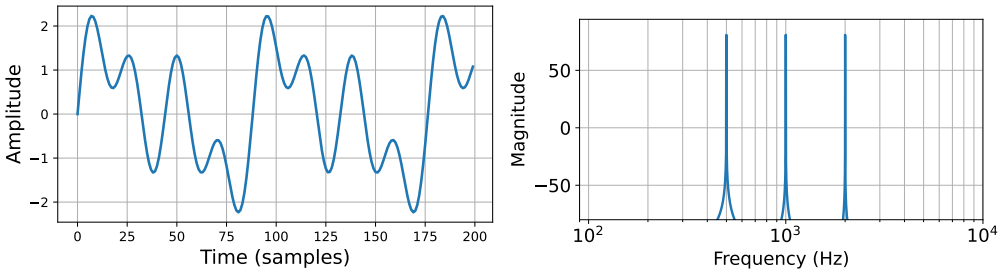
In the literature, it is more common to see Fourier analysis and Fourier synthesis referred to respectively as the *Fourier transform* and *inverse Fourier transform* [108]. The Fourier transform (analysis) takes a time-domain signal as input and converts it to an equivalent spectral representation. The inverse Fourier transform (synthesis) takes a spectrum and converts it into an equivalent time-domain representation. Transforms from either domain and back again should result in a lossless reconstruction of the original signal. The discrete form is therefore known as the discrete Fourier transform (DFT), and is the more commonly used form, as Fourier transforms are often performed on digital signals.

For N discrete time samples the DFT returns N equally spaced frequency bins $X[k]$ with bandwidths and centre frequencies determined by the sampling frequency and length of the signal in samples. While $x[n]$ represents the discrete time-domain representation of a signal, $X[k]$ represents the discrete frequency-domain representation. The DFT is a lossless operation, meaning the original time-domain representation of a signal is recoverable from the frequency-domain representation using the inverse DFT (iDFT)[108]:

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X[k] e^{\frac{jk\omega n}{N}} \tag{2.46}$$

The DFT is, however, a computationally expensive operation requiring N^2 individual computations. A more efficient method was devised to compute the DFT, aptly named the fast Fourier transform (FFT) [109], requiring a total of $N \log_2 N$ individual computations. Due to its increased efficiency, the FFT is now almost always used to compute the DFT for signal processing applications. Figure 2.21a shows a time-domain representation of a signal containing three frequencies components $f_1 = 500$ Hz, $f_2 = 1000$ Hz, and $f_3 = 2000$ Hz with Figure 2.21b showing an approximation of the frequency spectrum obtained from an FFT of the time-domain signal in Figure 2.21a. For the remainder of this thesis the term FFT will be used to refer to the application of the Fourier

transform to a time-domain representation of a signal, and iFFT to refer to the inverse Fourier transform of a frequency-domain representation of a signal.



(a) Time-domain (b) Frequency-domain

Figure 2.21: Sinusoidal signal with three frequency components at 500 Hz, 1000 Hz, and 2000 Hz represented in both a) Time-domain b) Frequency-domain.

2.5.5 Time-frequency processing

Unless the signal under analysis is either very simple, as with the previous sinusoidal signal examples, or very short, it is likely to have some non-stationary characteristics, such as spectral content, that varies over time. To facilitate a more accurate frequency analysis it is often useful to view the signal as a function of time and frequency. This is usually achieved through *frame-based processing* where an input $x(n)$ is divided into a number of much shorter frames with each frame being processed separately. When using time-frequency analysis it is assumed that the spectrum of a signal can be considered stationary if measured over short enough intervals.

The time-frequency representation of a signal can be derived from an extension to the FFT operation described in Section 2.5.4, and is referred to as the *Short-Time Fourier Transform* (STFT) given by [110]:

$$X_m(\omega) = \sum_{n=-\infty}^{\infty} x(n + m\xi)w(n)e^{-j\omega(n+m\xi)} \tag{2.47}$$

where $x(n)$ is the input signal at sample n, $w(n)$ is a window function of length M, $X_m(\omega)$ is the FFT of the windowed data centred about time $m\xi$, and ξ is the hopsize, in samples, between successive FFTs. This form of the STFT utilises

the *Discrete Time Fourier Transform* (DTFT) as its length N is permitted to approach infinity. Given that the length N of an FFT dictates the frequency-domain resolution, this results in the DTFT being a function of continuous frequency as N approaches infinity. In reality, the window $w(n)$ is always of finite length and usually centered about time zero which results in a re-expression of the DTFT as the DFT (in our case utilising the FFT):

$$X_m(\omega_k) = \sum_{n=-N/2}^{N/2} x(n + m\xi)w(n)e^{-j\omega(n+m\xi)} \tag{2.48}$$

where $X_m(\omega_k)$ indicates the now discrete nature of frequency sampling within the number of bins, k.

Assuming a fixed sampling rate, the length M of window $w(n)$ defines the resolution of the STFT in both frequency and time. This presents a compromise between the two dimensions. A longer window length will result in greater frequency resolution but at the expense of lower resolution in time, which may cause transient or shorter term non-stationary events to go undetected. A shorter window length will be more sensitive to temporal changes in the signal but less detailed with respect to how those changes are represented in the spectrum with each frequency bin ω_k containing the approximated energy for a wider band of frequencies. In practice, the window length is often selected through iterative experimentation based on the signal under analysis and the dimension that is most of interest. Typical window lengths in audio signal processing applications vary between 1 ms and 100 ms [34].

A common use of the STFT is the generation of spectrograms, which represent the intensity of the STFT magnitude. As spectrograms usually show the log-magnitude intensity (dB) across time and frequency, and since SPL(dB) can be approximated to perceived loudness as explored in Section 2.2, they can be said to provide an approximate display for how the human auditory system would perceive a given signal. This is assuming the choice of an appropriate window length.

Another application of the STFT is to facilitate the linear and time-varying processing of a given signal. In this scenario, once the frame $x(n + mR)$ has been

transformed into the frequency-domain giving $X_m(\omega_k)$ the desired processing is applied, which is represented as:

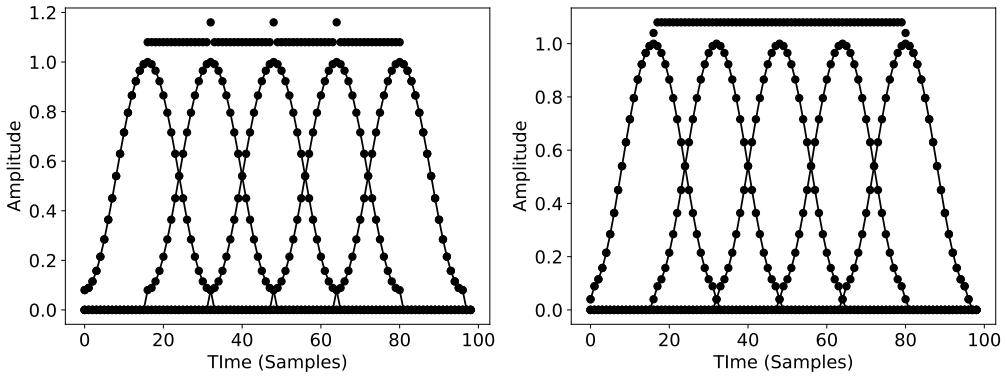
$$Y_m(\omega_k) = H_m(\omega_k)X_m(\omega_k) \tag{2.49}$$

with $H_m(\omega_k)$ as the frequency response of the spectral processing to take place.

It should be noted that when applying spectral processing the FFT size N must be greater than or equal to the $M + L - 1$ where L is the length of the processing filter, to avoid time aliasing [110]. This is achieved through zero-padding the time-domain signal which equates to interpolation in the frequency-domain where each frequency bin is replaced by N/M bins. Additionally, due to the time and frequency modifications introduced by the analysis windows, the frames must overlap to ensure an accurate reconstruction of the desired output signal. A commonly used method is Overlap-Add (OLA) processing. In simple terms it is a sequence of FFTs which may be modified, inverse-transformed, and then summed to create the reconstructed output signal. To ensure successful reconstruction of the input signal or processed version of the input signal, the window $w(n)$ must have *Constant Overlap-Add* (COLA) at hopsize ξ defined as:

$$\sum_{m=-\infty}^{\infty} w(n - m\xi) = 1, \ \forall n \in \mathbb{Z} \tag{2.50}$$

whilst ensuring COLA is vital for processing that requires reconstruction of a frame-processed audio signal, it is less important if using the STFT solely for analysis or visualisation of a signal [110]. Figure 2.22 shows two examples of overlap-add reconstruction using a standard Hamming window and highlights the importance in checking whether a window meets COLA if using a window defined in a library. Those that do not meet COLA will exhibit discontinuities in each frame of the overlap add. For a review of windowing functions, see [110].



(a) Non-COLA (b) COLA

Figure 2.22: Overlap-Add example for a Hamming window with a length of 33 and a hopsize of 16. (a) shows the results of a non-COLA window resulting in discontinuities at the edge of each overlap-add. Whilst (b) shows the COLA solved for a window with odd length M. Generated from code adapted from [110].

2.6 Soundfield Recording, Encoding, & Reproduction for IME Production

IME productions utilise a range of different approaches to spatial sound recording, representation, and reproduction to facilitate the creation of a desired acoustic environment. These can broadly be categorised with respect to how they represent soundfield data. Common categorisations are channel-based audio (CBA), scene-based audio (SBA), object-based audio (OBA), and binaural-based audio (BBA) [111–113].

2.6.1 The Soundfield

The term soundfield often refers to the sound waves present within a given space [114, 115].The simplest of soundfields can be described by a single sinusoidal plane wave, with frequency f, propagating through a free-field. Although sound sources are often described as point sources radiating spherically, it is mathematically simpler to assume waves are planar given sufficient distance from the source, and in practise modelling and synthesis of soundfields is often done based on planar waves. Given a single sinusoidal plane wave, sound pressure p measured

in spherical space at point $\mathbf{r}$ can be defined as [116]:

$$p(\mathbf{k}, \mathbf{r}) = e^{-jkr} \tag{2.51}$$

where $\mathbf{k} = [k \ \theta_k \ \phi_k]$ is the wave vector, describing the wave's direction of propagation and k is the wave number in radians per meter. It should be noted that the negative exponent as the measured direction of arrival of a wave is considered to be opposite to its direction of travel.

The capture or encoding of a soundfield located in a physical space (as opposed to a virtual or synthesised space) is usually undertaken using one or more microphones. Given the many methods used to capture the various properties of soundfields, both perceptual and physical, this thesis defines a soundfield in the broadest sense as any bounded space where at least a single airborne pressure wave is present. The difference between a soundfield consisting of a single pressure wave and a complex real scene that is the superposition of the aforementioned components is a matter of scale. With respect to soundfield capture, encoding, and reproduction of a soundfield, this thesis again uses these terms in the broadest sense. Therefore, the difference between a single pressure measurement (and the reproduction of that signal over a single loudspeaker) and multiple measurements using multiple microphones or multichannel microphones (and the subsequent reproduction over multiple loudspeakers) is treated as a difference in spatial resolution stemming from the *spatial sampling density* [117].

2.6.2 Basics of Soundfield Recording

The easiest method to capture information that can then be used to reproduce, at least in part, a given soundfield, is through the use of one or more microphones. Given the ubiquity of mobile phones it is a reasonable assumption that many people have the means to record at least a low spatial resolution representation of the soundfield they are present within. The minimum spatial resolution results from the capture of a soundfield using a single microphone, as depicted in Figure 2.23. The signal that results from recording of a soundfield is a combination of the sound incident on the microphone and the characteristics of the microphone

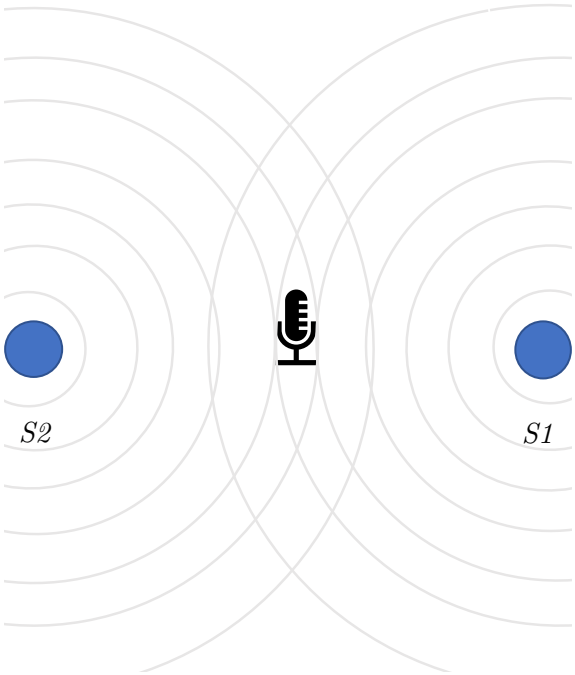


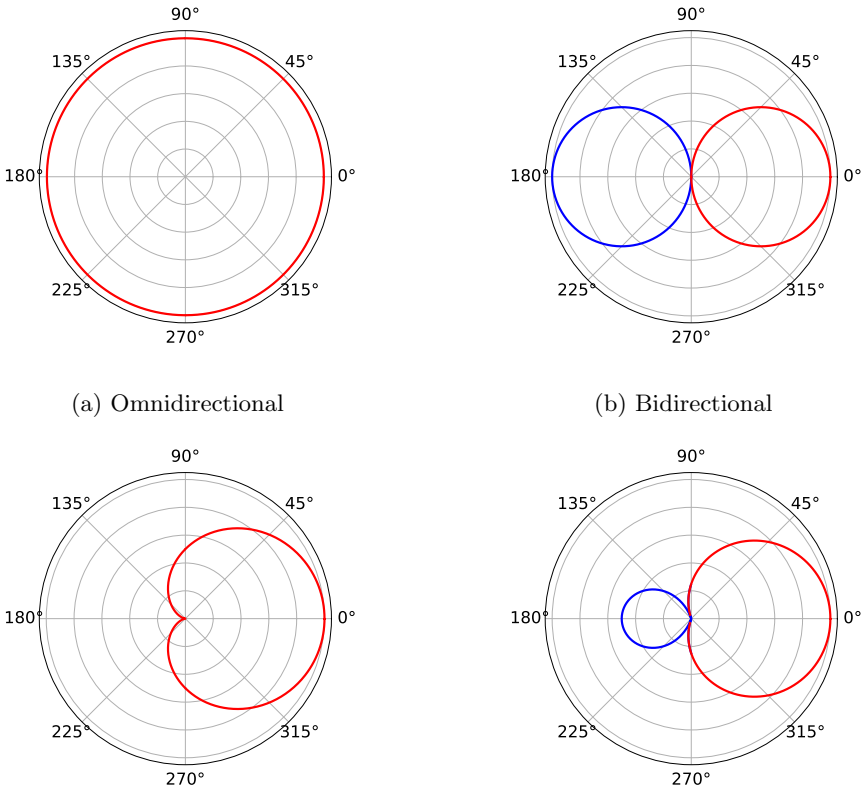
Figure 2.23: Basic model of a soundfield based on two point sources, S_1 and S_2, and one receiver.

such as its frequency response and polar pattern. The polar pattern describes the relative sensitivity of the microphone, and therefore its output level, to sound at various angles of incidence. Figure 2.24 shows 2D representations of some common polar patterns, with red representing reversed polarity. The combination of an omnidirectional (Figure 2.24a) and a bidirectional (Figure 2.24b) polar pattern can also be used to derive an infinite number of directional patterns that scale between the two. The polar response curve can be derived from [118]:

$$r = |\Gamma + \varrho \cos(\theta)| \tag{2.52}$$

where r is the radial distance from the origin of the polar plot and represents relative output; Γ and ϱ are fractional coefficients of the pressure (omnidirectional) and pressure gradient (bidirectional) polar patterns respectively with $\Gamma + \varrho = 1$, and θ is the angle of incident sound relative to the principal axis of the microphone.

Consider again Figure 2.23, if an omnidirectional microphone was used to capture the sound scene, and assuming sources S_1 and S_2 are equidistant from



(a) Omnidirectional

(b) Bidirectional

(c) Cardioid

(d) Hypercardioid

Figure 2.24: Four common microphone directional pickup patterns. Red denotes positive polarity and blue denotes negative polarity.

the microphone, both would equally contribute to the resulting signal. However, if a directional microphone was used pointing towards S_1 then S_1 would appear more prominent in the recorded signal than S_2 and visa versa. In the case of bidirectional (Figure 2.24b) pickup patterns the sound located directly behind the microphone would also contribute equally to the resulting signal, however, the signal captured from the rear of the pickup pattern would be out of phase with the signal captured from the front of the microphone. This suggests that use of a single microphone results in the lowest spatial resolution capture of a soundfield, since the approximate sound pressure at that point in space can be encoded for sound waves from all directions (assuming an omnidirectional polar pattern); however, there is no encoding of spatial information about the scene. It

could be argued that using a directional polar pattern, for instance a cardioid pattern, would encode some very limited spatial information given the directional sensitivity of the microphone. This would not however be detailed enough to build up an accurate representation of how the sound pressure incident on the microphone varies over angle for a given scene. Furthermore, it only provides a single channel for reproduction.

A higher spatial sampling density is required to obtain a soundfield recording with a higher spatial resolution. This can be achieved by increasing the number of discrete sampling points in space i.e. microphones, in appropriate configurations within the space. A higher spatial resolution for reproduction is similarly achieved by increasing the number of independent reproduction channels i.e. loudspeakers. However, when using capture, processing, and rendering methods relating to BBA, it is possible to reproduce spatial sound with a higher spatial resolution without the need for additional reproduction channels.

2.6.3 Channel-based Audio

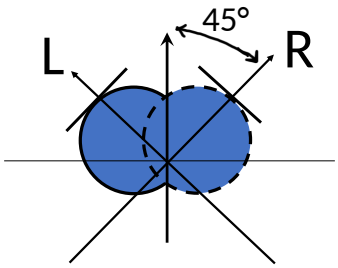
CBA methods can be considered as being loudspeaker-centric, where audio content is either captured or processed such that it is represented by a number of signals, each intended to be delivered to a specific loudspeaker within a pre-defined arrangement, typically without the need for any further modification [113]. This results in CBA being a relatively straight forward reproduction format as the signals are already rendered for each loudspeaker. However, CBA requires that the content be reproduced over the same loudspeaker configuration for which it was created, meaning that a separate version must be generated for each specific loudspeaker configuration. As such, a set of industry recommendations have been established for common configurations such as 2.x, 5.x, and 7.x [119], where x in this instance represents the number of low frequency effects (LFE) channels and the number preceeding the decimal point represents the number of full range loudspeaker channels. In cases where CBA is required to be reproduced over a different configuration there are solutions, such as MPEG-H [120], which facilitate upmix/downmixing between different configurations.

CBA requires a minimum of two channels to encode spatial information. The two-channel case is typically referred to as stereo, two-channel stereo, or two-channel stereophony. Research into stereophonic sound started as early as 1881 [121], however, the development of modern approaches to two-channel stereo are considered to have stemmed from the work of Blumlein in the 1930s [122]. By placing two or more microphones in a soundfield, a spatial representation can be captured which encodes some of the time and level localisation cues discussed in Section 2.4 as *Inter-Channel Time Differences* (ICTD) and *Inter-Channel Level Differences* (ICLD) [123]. As such, CBA can be considered a perceptually motivated approach to spatial sound as it aims to render perceptually relevant cues as opposed to a physically accurate approximation of the soundfield [124].

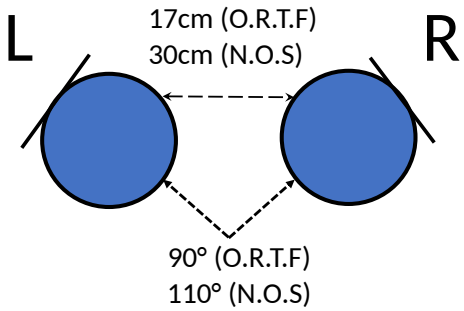
Two-channel stereo is often used to create a frontal sound stage that exists between the bounds of the two loudspeakers. There are a variety of different stereo microphone techniques that utilise different combinations of space, microphone orientation and polar pattern in order to influence the spatial attributes of the recorded sound field. Stereo microphone techniques can broadly be classified into three categories: coincident, near-coindicent, and spaced configurations. Several common stereo microphone configurations are discussed, but for a more detailed review see [123] and [125].

A spaced pair (Figure 2.25a) consists of two identical microphones with matching polar patterns, commonly cardioid or omnidirectional. Although cardioid polar patterns are much less sensitive to sources arriving from behind the array, omnidirectional may be preferred as pressure microphones are not subject to the proximity effect and therefore exhibit a more consistent low frequency response over different distances. Both microphones are orientated towards the direction of the intended sound scene, parallel to each other, and spaced anywhere from 10 cm to up to several meters. Due to the small ICLD between a closely spaced configuration, it can be considered to encode the spatial scene using only ICTD, although once the spacing becomes 1 to 2 meters it will also introduce greater ICLDs alongside enhancing the ICTDs. This results in a stereo representation with enhanced width [123].

(a) Spaced pair



(b) XY pair, adapted from [123]



(c) Configurations for common near coincident techniques

Figure 2.25: Diagrams of common stereo microphone techniques a) Spaced pair b) XY coincident pair c) Near coincident pairs such as O.R.T.F and N.O.S

Spaced pairs utilise both ICTD and ICLD to create a stereo image, both of which are a function of the incident angles of the sound sources. The greater the distance between the two microphones, the greater the possible range of inter-channel difference values. However, in cases where the distance between microphones is several meters, it may be beneficial to add a third microphone in the middle of the configuration in order to ensure consistent coverage of the sound scene.

Coincident configurations, such as the X/Y pair shown in Figure 2.25b, use a pair of identical microphones with their respective capsules placed as close to one another as possible, but without touching. The orientation of each microphone will then be dependent upon the technique being used. The X/Y pair consists

of two cardioid pattern microphones that are usually stacked one on top of the other so they occupy the same region horizontally, with capsules orientated between 90° to 135° away from each other [125]. The Blumlein pair utilises two bidirectional microphones positioned so their pickup patterns are orthogonal to one another on the horizontal plane. Due to the close proximity of the capsules, coincident configurations are considered to encode the spatial scene using only ICLDs. Additionally, the lack of appreciable time and phase differences between the microphones also results in the both signals having good mono compatibility when mixed down.

Near-coincident configurations (Figure 2.25c) combine both ICLD and ICTD resulting in a stereo scene that has both stable localisation of sources and a sense of space and depth. The lower ICTD when compared to spaced configurations also make near-coincident recordings mono compatible. Similar to coincident configurations, the microphones are orientated at an angle facing away from each other, which varies depending on the configuration being used, but additionally they are also spaced apart at distances that produce appreciable ICTDs. The O.R.T.F cofiguration typically uses a distance of 17 cm between capsules at an angle of 110°, while the N.O.S configuration uses a distance of 30 cm and an angle of 90°.

CBA can also be synthesised through the manipulation of identical mono signals that are sent to two or more loudspeakers. This process, known as *panning*, alters the signals that drive the left and right loudspeakers in order to create ICLD and ICTDs. This is commonly achieved through *amplitude panning*, which refers to the manipulation of the relative amplitudes of the mono signals sent to each loudspeaker [126]. By manipulating the signals that drive each loudspeaker, sources can be made to appear as if they are positioned between the loudspeakers, this is referred to as phantom imaging [126]. The *sine law* was an initial model proposed by Bauer [127] which predicts phantom positions θ_s according to the gains of each loudspeaker placed at $\pm\theta_L$:

$$\frac{\sin(\theta_s)}{\sin(\theta_L)} = \frac{G_L - G_R}{G_L + G_R} \tag{2.53}$$

The *tangent law* was also developed around a similar time and is defined by [128]:

$$\frac{\tan(\theta_s)}{\tan(\theta_L)} = \frac{G_L - G_R}{G_L + G_R} \tag{2.54}$$

The key difference between the two is that the tangent law considers the propagation path around the head [129], however the perceptual difference between the models has been found to be negligible.

Given that the standardised positions for two-channel stereo +/- 30°either side of the central listening position [119], Wiggins [42] notes that the application of pair-wise panning to a surround sound configuration would require a minimum of six loudspeakers. In this case each pair of speakers can operate as a two-channel subsystem within the larger array. However, pair-wise panning is still often used for irregularly spaced configuration such as 5.1 and 7.1. In these cases, unstable phantom imaging occurs when sources are panned between speaker pairs where the separation angle is greater than 30° [42]. CBA consisting of a number of discrete channels greater than two is often referred to as multi-channel audio or multi-channel surround sound but still largely follows the same principles.

With respect to the spatial sampling density, stereo microphone techniques result in the sound field pressure being sampled at two positions, which allows the encoding of lateral directional information. The term *lateral* is used as opposed to *location* because whilst it is possible to identify if a source is coming from the left or the right, there is insufficient information to make an adequate judgement on its elevation and whether it is in front or behind the recording array. There are other approaches to channel-based multichannel recording which provide a greater sampling density through the use of additional microphones, such as the Surround Decca Tree [130], Williams Multi-Microphone Array (MMA) [131], and ORTF-3D [132].

2.6.4 Object-Based Audio

OBA provides a greater flexibility than CBA as it allows an object-based represented scene to be used with varying loudspeaker configurations, without the need for the original encoded material to be modified [112, 133]. When authoring OBA

content, the scene is represented by multiple audio objects, with each comprising of one or more audio signals and associated metadata. Examples of associated metadata are the location of the object, the trajectory of an object, and the gain level associated with the object. Some cinema mixing formats include object metadata that allows dynamic reconfiguration of the object renderer to facilitate the object only being rendered by certain loudspeakers or loudspeaker zones i.e *screen zone*, *sides zone* [134].

As detailed in Section 2.3, the rendering of objects in specified locations within a space requires the use of defined coordinate systems and a frame of reference. Tsingo [35] highlights that for interactive rendering, such as those found in many IMEs, the position of audio objects are usually represented as allocentric Cartesian coordinates. For IMEs where a single perspective needs to be rendered, which is often the case for first-person experiences, the coordinates of the objects can be converted into user-centric spherical/head-related coordinates. The reason for taking an allocentric first approach is that for experiences where sound is reproduced in a real space, such as in cinemas, theatres, and exhibition spaces, the position of audio objects can be described for every listening position and any room size in a way that allows optimal reproduction for a range of room sizes and shape. Spatialisation from a user-centric first approach can result in an object being reproduced is an incorrect location due to different room dimensions. Tsingo [35] presents an example of an object on a side wall within an elongated room being reproduced on the back wall when reproduced in a smaller room.

There are a number of different object-based rendering algorithms that can be used to render the position of audio objects according to their associated location and trajectory metadata. The purpose of an audio rendering algorithm is to decode the audio data associated with an audio object, according to the location and/or trajectory metadata, to a set of loudspeakers such that the object is perceived to be originating from that location. In contrast to CBA, where the encoding and decoding of content are intrinsically linked as channel-based content will be encoded directly to loudspeaker signals for the target loudspeaker configuration, OBA separates the encoding and decoding process by encoding

the spatial information on a per-object basis using coordinate data that is then interpreted by the chosen rendering algorithm, which decodes the signal to the appropriate loudspeakers.

Extending the amplitude panned principle, rendering methods such as vector base amplitude panning (VBAP) [135] and distance-based amplitude panning (DBAP) [136] allow for the use of known arbitrary 2D and 3D configurations. VBAP utilises triplet-wise panning to render an audio object at the desired location, with the speaker triplets being obtained via triangulation of the convex hull of the loudspeaker array [137]. The position of an audio object $\mathbf{p}$ can be described by a linear combination of three loudspeaker vectors as defined in [135]:

$$\mathbf{p} = g_1\mathbf{s_1} + g_2\mathbf{s_2} + g_3\mathbf{s_3} \tag{2.55}$$

where $\mathbf{s_1}, \mathbf{s_2}, \text{and } \mathbf{s_3}$ are loudspeaker vectors and where g_1, g_2 and g_3 are gain scaling factors. This can be expressed in matrix form as:

$$\mathbf{p^T} = \mathbf{gS_{123}} \tag{2.56}$$

where $\mathbf{g} = [g_1\ g_2\ g_3]$ and $\mathbf{S_{123}} = [\mathbf{s_1}\ \mathbf{s_2}\ \mathbf{s_3}]^T$. To solve for the gain vector, $\mathbf{g}$, the expression can be reformulated as:

$$\mathbf{g} = \mathbf{p^T S_{123}^{-1}} = [p_1\ p_2\ p_3] \begin{bmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} & s_{23} \\ s_{31} & s_{32} & s_{33} \end{bmatrix}^{-1} \tag{2.57}$$

with the gain vector requiring normalisation to ensure constant loudness

$$g_{s^{scaled}} = \frac{g_s}{\sqrt{\sum_{s=1}^{s} g_{s^2}}} \tag{2.58}$$

Zotter and Frank [137] also highlight that gains in $\mathbf{g}$ should always remain positive as to avoid in-head localisation and other artefacts.

However, as VBAP (and other directional vector based panning methods) only use the direction of the audio object relative to the reference position (the sweet spot), it is unable to effectively deal with object source locations not on the surface of the unit sphere e.g. an object positioned or moving towards the

centre of the space. In some cases this may result in the object appearing to sharply go up and over the centre of the configuration [35].

Distance based panning methods, such as DBAP [136, 138], seek to address this issue by calculating the loudspeaker gains based on the relative distance from each loudspeaker L_i to the virtual source $\mathbf{p}$. This can be expressed as in [35]:

$$G_i(\mathbf{p}) = \frac{1}{\epsilon + (||L_i - \mathbf{p}||)^\alpha} \tag{2.59}$$

where α is the distance exponent, usually assigned a value of 1 or 2, and ϵ is a coefficient that controls how much an object can be rendered by a single loudspeaker only.

This generally means all available loudspeakers are used to render the position of an object, which leads to smoother object panning trajectories. The benefits of this being that the number of loudspeakers is not restricted and the loudspeakers may be placed in any configuration [138]. This lends itself to irregular configurations, such as those required by concerts and outdoor events, where pre-defined geometric configurations may not be suitable [136]. Additionally, deriving gains from distance values rather than directional vectors means no assumptions are made about the position of the listener [138], although if listener position is known, further optimisations can be made through the addition of loudspeaker signal delays to ensure the sound from each speaker will arrive at the listener at the same time.

One potential disadvantage, however, of utilising all speakers for all objects is that as the number of objects increase the leakage to all speakers can result in the reproduction sounding less discrete. Furthermore, although Kostadinov, Reiss and Mladenov [138] found that localisation performance between VBAP and DBAP was comparable, they did not compare timbral quality. Given the case where the listener position is not known, and loudspeaker signal delays are not applied, coloration may occur due to the differences in time of arrival.

Wavefield Synthesis (WFS), as proposed by Berkhout [139], also lends itself to OBA production as it itself is based on a sound object paradigm [140]. Given the associated position of an audio object, a WFS rendering is able to calculate

information for the direct sound, early reflections, and diffuse reverberation, rendering the result to any given loudspeaker configuration. The initial object-based mix itself does not require prior knowledge of any intended loudspeaker configuration. However, unlike the previously detailed rendering algorithms, WFS can be considered a physically motivated spatial audio technique [124] since it aims to reproduce a physically accurate approximation of the target soundfield rather than render only the perceptually relevant auditory cues such as ILDs and ITDs. Additionally, it is also unlike most other spatial sound methodologies as it is a *volume solution*, meaning that it aims to accurately recreate the soundfield throughout the entire listening area, as opposed to just at a single listening position (the sweet-spot).

WFS is based on a combination of Huygens' Principle and Kirchhoff-Hemholtz integral, which together state that a propagating wave front of a primary source can be synthesised by an infinite number of secondary sources (loudspeakers) that are placed on the primary source's wave front of an enclosed volume (listening area) [140]. The superposition of all the secondary source signals combine to reproduce an accurate representation of the target wave front. Given the principle aim is the reproduction the wave front rather than the source itself, WFS is well suited to emulating distance with respect to sources appearing to originate from both behind and in front of the loudspeaker array. Whilst outside the scope of the thesis, the reader is referred to [140] for detail on the mathematical underpinnings of WFS.

WFS theory is based on an infinite number of infinitely small secondary sources, which is not possible given that all loudspeaker arrays will consist of a finite number of spaced non-infinitely small loudspeakers. This is analogous to the continuous to discrete transformation that occurs as part of time domain sampling as described in Section 2.5.1. In this context the spacing of loudspeakers results in a discrete spatial sampling and infers a *spatial aliasing frequency*, above which the sound field will not accurately being reproduced. The following equation, presented in [140], can be used to derive the spatial aliasing frequency f_A, assuming a plane wave:

$$f_A = \frac{c}{2\Delta s \sin \Delta\theta_S^w} \tag{2.60}$$

where Δs is the distance between loudspeakers and $\Delta\theta_S^w$ is the angle between the loudspeaker array and the wave front.

The spectral errors that result from spatial aliasing can cause a decrease in the localisation accuracy of the rendered soundfield, which has led to a number of attempts to reduce the problems caused by it. Wittek [141] proposed Optimized Phantom Source Imaging (OPSI), which uses a combination of WFS at frequencies $f < f_A$ and VBAP for frequencies $f > f_A$. In cases where $f > f_A$, only the two loudspeakers closest to the source position are used with amplitude panning. Although Wittek [141] showed that the coloration from the OPSI approach is less than audible than for pure WFS, it does result in objects with dominant high frequencies being perceived as between loudspeaker positions, along with the distance and size of objects not being reproduced. Corteel et al [142] proposed the use of diffusion filters to reduce comb-filtering effects and thus lessen the coloration of the sound. There have also been numerous optimisation methods where it is assumed that listeners only occupy a portion of the larger listening area and thus optimise for that particular area [143–146].

OBA has become an integral part of cinematic sound, being incorporated into systems such as Dolby Atmos, DTS:X, and Auro3D as well a significant component of current codecs such as Dolby AC-4 [147] and MPEG-H [120], which also allow the application of OBA to consumer devices [35] and vastly increases the flexibility with which personalised content can be delivered [148]. However, the increase in flexibility comes with an increase in complexity for both the encoding and rendering of OBA content given that a scene will often consist of a far greater number of objects that eventual loudspeaker signals. This results in an amount of data required to be transmitted and processed that exceeds that of simpler channel-based approaches.

2.6.5 Scene-based Audio

Scene-based audio (SBA) shares similarities with both CBA and OBA and can be viewed as the midpoint on a continuum with respect to how it represents a given sound scene. CBA, at one end, representing a scene by target loudspeaker gains and OBA, at the other end, representing a scene as a collection of individual objects with associated metadata. SBA, however, is neither focused on individual objects nor makes any assumptions about any loudspeaker configuration it may eventually be decoded to. SBA instead, spatially encodes the scene into a number of specified channels, which collectively describe the spatial characteristics of the scene and can later be decoded to a chosen loudspeaker configuration. The term SBA is often most associated with Ambisonics and Higher Order Ambisonics (HOA), however can also encompass spatial audio coding methods such as Directional Audio Coding (DirAC) [149–151]. The latter of which is covered in more detail in Chapter 6.

Ambisonics was developed by Gerzon [152–154], Fellgett [155], and Craven throughout the 1970s and is a scalable approach to sound field reproduction. Unlike many traditional surround sound methods, Ambisonics does not require prior knowledge of loudspeaker positions during the recording or encoding process. Alongside research into the reproduction of Ambisonics, early research was also conducted into appropriate microphone configurations which could be used to capture and encode into Ambisonic format [156, 157].

Whereas traditional surround sound tends to encode directly into discrete speaker feeds, Ambisonics provides a generic representation of the sound field that can later be decoded according to the given loudspeaker arrangement. Ambisonics is usually described as the decomposition of a sound field into spherical harmonics, which are a set of orthogonal basis functions able to describe any function on the surface of a sphere. Within the context of sound field capture and encoding, it is more intuitive to think of spherical harmonics in a similar fashion to microphone polar patterns that form part of a coincident recording array. When encoding into Ambisonics, the sound field is decomposed into these orthogonal functions

(polar patterns), and weighted combinations of these functions can produce an equivalent order function pointing in any direction.

2.6.5.1

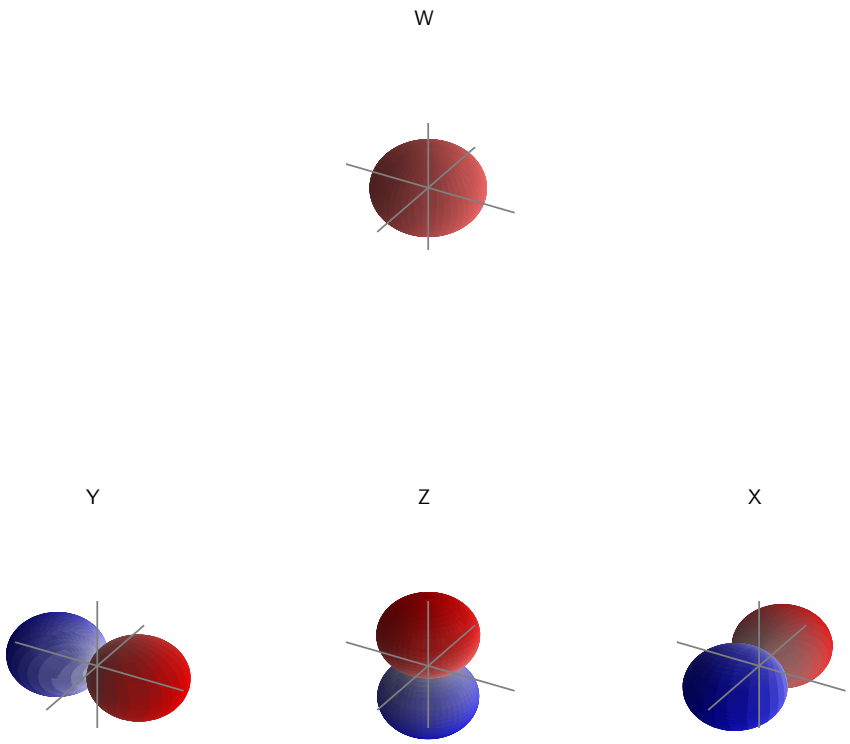


Figure 2.26: B-format spherical harmonics termed W, X, Y, Z. Red denotes positive polarity and blue denotes negative polarity.

First Order Ambisonics (FOA) represents a sound field using four spherical harmonic functions, hereafter referred to as channels (analogous to microphone channels). These first four channels are collectively referred to as B-Format. The W channel is an omnidirectional pressure signal that describes the 0th order component of the sound field and X, Y, and Z are figure of eight patterns facing

in the x, y, z Cartesian directions and collectively encode the three-dimensional particle velocity. Figure 2.26 shows a graphical representation of the B-Format channels. To encode a mono source s at a desired location, the mono signal can be multiplied by the gains for each B-Format channel that correspond to a point on a unit sphere. The B-Format gains can be derived as follows:

$$W = \frac{1}{\sqrt{2}} \tag{2.61}$$

$$X = \cos(\theta)\cos(\phi) \tag{2.62}$$

$$Y = \sin(\theta)\cos(\phi) \tag{2.63}$$

$$Z = \sin(\phi) \tag{2.64}$$

Another advantage of the Ambisonic format is the ease at which the sound field can be rotated about all three axes, a technique utilised for dynamic binaural rendering of Ambisonics to counter head movements and ensure a static absolute source position. This is opposed to the source being head-locked and fixed to a certain position e.g. always being 45° to the listener, irrelevant of the listener's head movement. Rotation about the Z-axis can be defined as [59]:

$$W' = W \tag{2.65}$$

$$X' = X\cos(\theta) + Y\sin(\theta) \tag{2.66}$$

$$Y' = Y\cos(\theta) - X\sin(\theta) \tag{2.67}$$

$$Z' = Z \tag{2.68}$$

Rotation about the X-axis (tilt) defined as:

$$W' = W \tag{2.69}$$

$$X' = X' \tag{2.70}$$

$$Y' = Y\cos(\theta) - Z\sin(\theta) \tag{2.71}$$

$$Z' = Y\sin(\theta) + Z\cos(\theta) \tag{2.72}$$

And rotation about the Y-axis (tumble) as:

$$W' = W \tag{2.73}$$

$$X' = X\cos(\theta) - Z\sin(\theta) \tag{2.74}$$

$$Y' = Y' \tag{2.75}$$

$$Z' = Z\cos(\theta) + X\sin(\theta) \tag{2.76}$$

Gerzon and Craven [156–158] also developed a microphone for recording Ambisonic signals. In this context, Equations 2.61 - 2.64 can be thought of as simulating a B-Format microphone [42]. Gerzon and Craven proposed a FOA microphone that consisted of four sub-cardioid microphone capsules mounted in a tetrahedral array, as shown in the example in Figure 2.27. Though the capsules are not coincident, they are equally non-coincident in each direction, simplifying the process of correcting for a non-coincident array response. The output of the tetrahedral array, known as A-format, are the four channels each captured by their respective microphone capsule. The orientation of the capsules are usually defined as left-front (LF), right-front (RF), left-back (LB), and right-back (RB). The conversion from A-format to B-format is as follows [157]:

$$W = 0.5 \times (LF + LB + RF + RB) \tag{2.77}$$

$$X = (LF + RF) - (LB + RB) \tag{2.78}$$

$$Y = (LF + LB) - (RF + RB) \tag{2.79}$$

$$Z = (LF + RB) - (LB + RF) \tag{2.80}$$

Though FOA allows for a full spherical representation and reproduction of a sound field, it does so using a finite number of sampling points. Furthermore, the wide main lobes of the first order directional patterns also contribute to poor spatial accuracy and an amount of spatial blurring of point sources when decoded over multiple loudspeakers [160]. The wide frontal lobes specifically result in any single panned source being reproduced over a group of neighbouring loudspeakers. This not only has an effect on source width but also results in comb filtering when the paths between each loudspeaker and the listening position differ.

Figure 2.27: Soundfield SPS200 1st order microphone [159]

2.6.5.2 Higher Order Ambisonics

It is also possible to decompose a sound field into a higher number of more directional spherical harmonics alongside the signals already captured in the form of B-format. Higher Order Ambisonics (HOA) refers to the use of higher order spherical harmonics, and thus requires a greater number of channels. This results in a higher spatial sampling density and offers greater accuracy in the reproduced sound field and lends itself to more accurate localisation in both real and virtual environments [69]. The spherical harmonics up to 4th order are shown in Figure 2.28.

Whilst Equations 2.77 to 2.80 show the specific equations for deriving the weights (gains) for B-format, a more general definition for encoding a mono signal s into Ambisonic format β for a given direction can be defined from [161]:

$$\beta = sY_{mn}^{\sigma}(\theta, \phi) \tag{2.81}$$

where Y_{mn}^{σ} are the three dimensional full normalised (N3D) spherical harmonic functions of order m and degree n further defined as:

$$Y_{mn}^{\sigma}(\theta, \phi) = N_{mn}P_{mn}(\sin(\phi)) \times \begin{cases} \cos(n\theta), & \text{if } \sigma = +1 \\ \sin(n\theta), & \text{if } \sigma = -1 \end{cases} \tag{2.82}$$

where $\sigma = \pm 1$, $P_{mn}(\sin(\phi))$ are the associated Legendre functions [162], and N_{mn} is the normalisation strategy for the amplitudes of different spherical harmonic

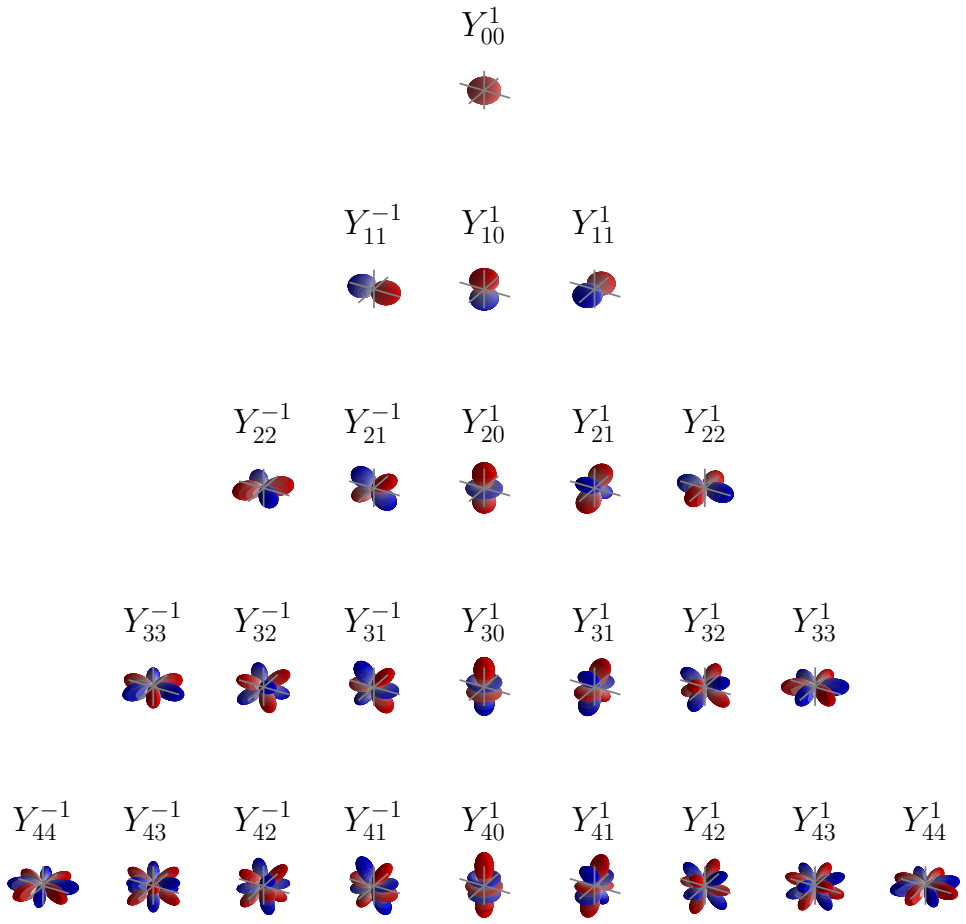


Figure 2.28: Spherical harmonics up to 4th order following Y_{nm}^{σ}

orders. The two most widely used normalisation strategies are three-dimensional normalised (N3D) and Schmidt semi-normalised (SN3D):

$$N_{mn}^{N3D} = \sqrt{(2-\delta_{n,0})(2_m+1)\frac{(m-n)!}{4\pi(m+n)!}} \tag{2.83}$$

$$N_{mn}^{SN3D} = \sqrt{(2-\delta_{n,0})\frac{(m-n)!}{4\pi(m+n)!}} \tag{2.84}$$

where δ is the unit impulse function described in Section 2.5.2 and defined in Equation 2.31.

Alphabetic channel ordering of Ambisonic channels only exists up to 3rd order [163] and the number of channels exceeds the number of letters in the English

alphabet for orders than higher than 4th order. Ambisonic channel numbering (ACN) is now considered the standard method for labelling spherical harmonic channels [164]. ACN can be calculated as:

$$ACN = m^2 + m + n\sigma \tag{2.85}$$

The microphone used later in this work is the MH Acoustics Eigenmike [165], which consists of 32 capsules in a near-uniform arrangement flush mounted on a rigid sphere, and that can output spherical harmonics up to fourth-order. However, unlike the simple directivity patterns associated with B-format, the complex directivity of higher order spherical harmonics can not easily be related to existing microphone polar patterns. Consequently, this require more advanced processing methods to derive them from the given microphone signals. Higher order microphones, such as the Eigenmike, utilise beamforming to approximate the correct directivity patterns from an array of microphone capsules.

2.6.5.3 Decoding

Ambisonic decoding can be seen as the process of converting data stored in Ambisonic format into a set of loudspeaker signals [160]. The conversion from Ambisonic format to loudspeakers is achieved using a decoding matrix $\mathbf{D}$, which results in each loudspeaker signal being a weighted sum of each Ambisonic channel dependent on the position of the loudspeakers [166]. A sampling decoder [42] for an arbitrary number of loudspeakers can be derived by calculating virtual microphone responses for each loudspeaker position θ_l and ϕ_l as:

$$g_w = \frac{1}{\sqrt{2}} \tag{2.86}$$

$$g_x = \cos(\theta_l)\cos(\phi_l) \tag{2.87}$$

$$g_y = \sin(\theta_l)\cos(\phi_l) \tag{2.88}$$

$$g_z = \sin(\phi_l) \tag{2.89}$$

The resulting signal that is then fed to the loudspeaker s_l is given as [167]:

$$s_l = \frac{(2-d)g_w W + d(g_x X + g_y Y + g_z Z)}{2} \tag{2.90}$$

where d is the directivity factor of the virtual microphone response in the range $0 \leq d \leq 2$ such that $d = 0$ results in an omnidirectional response pattern and $d = 2$ results in a bidirectional response pattern [42]. As this method utilises first order virtual microphone patterns, it is only suitable for simple frequency independent decoding of B-format signals. Additionally, this assumes a regularly spaced loudspeaker array, so when applied to irregular arrays, such as the ITU 5.0 array, localisation errors due to a non-uniform sound field are to be expected. Furthermore, FOA can only accurately reconstruct the sound field for a very small area at the centre of the array, known as the *sweet spot*.

Decoding through pseudo-inverse, or mode-matching, [160] is another common method for deriving loudspeaker signals for an arbitrary number of loudspeaker channels L for an arbitrary Ambisonic order with number of Ambisonic channels K. A $K \times L$ re-encoding matrix $\mathbf{C}$ is calculated by encoding the position of each loudspeaker into spherical harmonic coefficients using equation 2.82 resulting in:

$$\mathbf{C} = \begin{bmatrix} Y_{00}^{1}(\theta_1, \phi_1) & Y_{00}^{1}(\theta_l, \phi_l) & \dots & Y_{00}^{1}(\theta_L, \phi_L) \\ Y_{mn}^{\sigma}(\theta_1, \phi_1) & Y_{mn}^{\sigma}(\theta_l, \phi_l) & \dots & Y_{mn}^{\sigma}(\theta_L, \phi_L) \\ \vdots & \vdots & \ddots & \vdots \\ Y_{Mn}^{\sigma}(\theta_1, \phi_1) & Y_{Mn}^{\sigma}(\theta_l, \phi_l) & \dots & Y_{Mn}^{\sigma}(\theta_L, \phi_L) \end{bmatrix} \tag{2.91}$$

When multiplied by the given loudspeaker gains g this would yield the reconstruction of the Ambisonic format signal β as:

$$\beta = \mathbf{C}\mathbf{g} \tag{2.92}$$

To derive the loudspeaker gains needed we need to rearrange for $\mathbf{g}$ which results in:

$$\mathbf{g} = \mathbf{C}^{-1}\beta \tag{2.93}$$

Where $\mathbf{C}^{-1}$ is the inverse of C and known as the decoding matrix $\mathbf{D}$. Usually, as the number of Ambisonic channels is greater than the number of loudspeakers, $\mathbf{C}$ is not a square matrix and it is not possible to obtain the true inverse so instead the pseudo-inverse is obtained where:

$$\mathbf{D} = \text{pinv}(\mathbf{C}) = \mathbf{C}^{\mathbf{T}}(\mathbf{L}\mathbf{L}^{T})^{-1} \tag{2.94}$$

This results in the signal of each speaker being calculated as:

$$s_l = \sum_{k=1}^{K} \beta_k \mathbf{D}_{kl} \tag{2.95}$$

Both decoding methods detailed have been described as frequency-independent and assume a regularly spaced loudspeaker array. Though outside the scope of this thesis, the reader is directed to [72, 168] for a detailed review of decoding methodologies and [42, 169, 170] for details on optimising for irregular arrays.

Additionally, using specialist microphones, such as the ones detailed in this section, IRs can be captured using a similar methodology to that described in Section 2.5.2 and encoded into the SH domain with the spatial resolution being dependent upon the order of SH used. It is also important to note that for many spatial IR techniques, the IRs must ideally be captured at each desired source location using the same microphone array. When this is not possible, and assuming the IRs are captured from a sufficient minimum number of positions, IRs can be interpolated to synthesise any number of locations between two existing IRs for a position in 2D space and three IRs for a position in 3D space [171, 172].

2.6.6 Impulse Response Measurements

The impulse response defined in section 2.5.2, while not a method of capturing a given sound scene, can be used as a method of encoding the acoustic characteristics of a given space with varying degrees of spatial resolution depending on the microphone array used to capture the IR. A convolution operation, as described in section 2.5.3, can then be utilised in order to impart the acoustic properties of a given space onto a recording that is preferably anechoic.

Capturing an IR with a single microphone enables reproduction of an acoustic space but without the inclusion of any spatial information. This results in the same level of spatial resolution as recording a sound scene with a single microphone. Utilising multichannel recording techniques further allows the capture of directional information in the form of the direction of arrival of the direct sound, as well as the early reflections. IR captured through the use of multichannel recording techniques are often referred to spatial impulse responses

[59, 173] as they not only encode the acoustic characteristics of a space, but also the spatial characteristics such as direction of arrival for direct sound and early reflections. When spatial IRs are convolved with an anechoic source, and then reproduced using an appropriate loudspeaker or headphone configuration, they will exhibit the spatial characteristics of the captured space. Applications of spatial IRs include auralization of existing spaces [174], soundscape measurement, modelling, and evaluation [55], and architectural acoustic design [175].

2.6.7 Binaural-based Audio

Binaural-based audio (BBA) refers to a variety of techniques which provide a spatial audio experience over two channels. These two-channels aim to control the sound pressure at the two ear drums and thus reproduce the interaural and spectral cues, detailed in Section 2.4, of the target sound scene [176]. Binaural can be considered a perceptually motivated technique, one that is most commonly reproduced over headphones, although loudspeaker reproduction is possible utilising cross-talk cancellation [177, 178]. Comprehensive reviews of binaural technology are presented by Pike [75] and Rafaely et al. [179].

Interest and research in headphone-based binaural dates back to the 19th Century [179] but has seen a surge in popularity in recent decades given the increased availability of personal headphones and even more recently due to the applicability of the format to IME experiences. The cues required for BBA can either be recorded from human listeners, using the methods briefly outlined in Section 2.4.4, captured from dummy head microphones and head and torso simulators (HATS), or synthesised using signal processing methods. Binaural recordings require that microphones are placed at the ears of a dummy head, HATS or a human listener, which then capture the sound pressure at the ears at the given location. However, as Pike [75] highlights, the limitation of binaural, and all other types of spatial recording, is that it is only able to capture naturally occurring scenes. This introduces complexities for IMEs where the desired sound scenes do not, and in some cases cannot, exist. Rafaely et al. [179], also highlight two additional issues that stem from the HRTF being embedded into the recording

itself. Firstly, as head position is captured during the recording process, head-tracking is generally not possible and secondly, individualised HRTFs cannot be used as the recorded signal already has the HRTF of the associated device/person from which the recording was captured. There are methods for binaural cue adaption, but these have been reported to lack in accuracy and flexibility [180].

Binaural synthesis, also referred to as binaural rendering, is the processing of an audio signal with the aim of simulating the binaural cues required for the original signal to appear at the desired spatial location. Binaural synthesis requires knowledge of the acoustic transfer path between the source and each of the two ear drums, which can be characterised by their impulse responses, referred to as the HRIR and HRTF in the frequency domain. When using direct convolution, as detailed in Section 2.5.3, each source position requires a pair of HRIRs/HRTFs, one for each ear, and when convolved with an anechoic source results in the superposition of the relevant interaural and spectral cues for the given source position. In a free-field environment the HRTF can be seen as representing the anechoic transfer function from a source to the listener's ear drums [75] for a given position and will typically result in a HRIR of around 512 samples in length at a sampling rate of 44.1 kHz [124], which equates to around 12 ms. Depending on the environment within which the HRIRs are captured, there may be some room reflections also captured, but ideally a HRIR should represent solely the effect the morphology of the listener has on the wave fronts impinging on the ears.

A measured HRTF will usually require equalisation to remove the transfer functions of the measurement loudspeaker and microphones [176]. The type of equalisation required will depend on the intended use of the HRTFs. For analysis only, a simple inverse filter can be derived for the free-field transfer function of the loudspeaker and measurement microphones [181]. For rendering over headphones, the headphone-to-ear transfer function must also be corrected for [182–184].

Given that a pair of HRIRs are required for each desired location and that the spatial resolution of human hearing can be as low as 1° in azimuth and 4° in elevation [161], it would require a large number of measurements to be able

to synthesise sounds from all possible directions. It can, therefore, be a time consuming task to collect a dataset of HRIRs with a high spatial resolution. One method to reduce the time taken is the *overlapped swept sine* technique [185], where multiple overlapping sweeps are played through multiple loudspeakers at once, with an offset equal to the reverberation time of the environment.

If measurements are undertaken in an environment other than an anechoic one, there will likely be some influence from the measurement environment in the form of reflections, whether it be from the room itself or the measurement equipment (such as any additional loudspeakers). In many cases this can be unwanted and although it is often possible extract the pure HRIR by truncating the measured signal at a time that would exclude all but the direct sound [74] this will, however, change the accuracy of the low frequency reproduction. However, given the limited low frequency reproduction capabilities of most loudspeakers used for HRIR measurement, low frequencies are often modelled to compensate for this [74, 186], and as such it may not cause any noticeable perceptual issues. In circumstances where the measured HRIR also contains the Room Impulse Response (RIR) (which contains the early reflections and reverberation), these two components collectively are known the Binaural Room Impulse Response (BRIR) [59, 75, 161] and can be used to binaurally render a signal in a given location in the given environment.

As detailed in Section 2.4.4, each person has a unique set of HRTFs and it has often been shown that using ones own HRTFs offers a range of improvements to the listening experience when compared to using non-individualised HRTFs. Some of the improvements include greater externalisation, better localisation and timbral accuracy, and a generally more natural and believable binaural experience [63, 73, 187]. However, for mass-market consumer applications of binaural technology, the wide-spread use of personalised measurements is impractical as the measurement of HRTFs are often time consuming and requires specialist equipment. This has lead to a lot of research investigating whether individualised HRTFs are needed, what is the subjective/objective difference (if any) of experiences using individualised and non-individualised HRTFs, and devising

means of individualising the technology, but without requiring the capture of the individual's HRTF through traditional methods.

Whilst there are significant differences between the HRTFs of different individuals, there are certain interaural differences that are much more consistent over the population. ITDs and ILDs from non-individualised HRTFs still provide relatively robust horizontal localisation cues given that the differences in head size and ear spacing of different individuals is relatively small [188]. The greater differences in spectral cues, however, give rise to higher rates of front-back confusions and up-down confusions [189]. However, a later study by Begault [190], showed little benefit of individualised HRTFs on localisation, front-back reversals, or externalisation. The results did, however, show that the introduction of head-tracking gave lower azimuth errors due to the ability of the listeners to utilise the dynamic localisation cues described in Section 2.4.3. It has been shown frequently across multiple studies that head-tracking reduces front-back reversals [190], aids distance localisation [68] and improves externalisation [100].

Binaural rendering can also be used in conjunction with other spatial sound methodologies, such as VBAP and Ambisonics. The binaural rendering of Ambisonics was first proposed by McKeag and McGrath [191], with Jot et al. [166] labelling the methodology as the *virtual loudspeaker* approach. Further developments to the binaural rendering of Ambisonics have been proposed by Noisternig [70], McKenzie[161], and Armstrong [192] amongst others. The interest and advancement in the area of headphone-reproduced Ambisonics has progressed significantly in recent years as new applications have emerged, such as the use of binaural technology in headphone-based media such as video games, virtual reality (VR), augmented reality, and mixed reality, all of which are discussed in detail in Chapter 3.

One of the reasons for the popularity of binaural Ambisonic rendering is the ease with which a spherical harmonic scene can be rotated about all 3 axes, as demonstrated by the rotation matrices presented by Equations 2.65 - 2.76, which is useful for head-tracking [168]. Additionally, as noted by Wenzel and Foster [193], binaural Ambisonic rendering also negates the need for HRTF interpolation,

which can be computationally expensive. Binaural rendering using the *virtual loudspeaker* method also reduces the number of convolutions required to render the scene. Standard binaural rendering requires each individual source to be convolved with a HRTF pair, however, rendering binaural Ambisonics using virtual loudspeaker reduces the number of convolutions to twice the number of virtual loudspeakers used in the decoding process. The binaural rendering of Ambisonics signals using the *virtual loudspeaker* approach can be mathematically expressed as follows:

$$B = \sum_{l=1}^{L} H_l * s_l \tag{2.96}$$

where B denotes the binaural signals and s_l denotes the loudspeaker signals as calculated in Equation 2.95. The same process would be followed to render binaural VBAP signals as it is simply the convolution of HRTFs with chosen loudspeaker signals and is not domain specific to Ambisonics.

It is possible to reduce the number of required convolutions further by moving the convolution into the spherical harmonic domain as proposed by Avni et al. [194]. By encoding the HRTFs into the spherical harmonic domain the number of convolutions required is now equal to the number of Ambisonic channels as opposed to the number of loudspeakers, and as generally the number of loudspeakers will be greater than the number of Ambisonic channels this reduces the number of convolutions required and therefore lessens the computational complexity. This method differs from the virtual loudspeaker method in that the HRTFs are first encoded into the spherical harmonic domain using a transposed decoding matrix. The resulting spherical harmonic components are then convolved with the corresponding Ambisonic channels and summed to get the resulting binaural signal. This can be expressed as in [192]:

$$\sum_{k=1}^{K}\left(\left(\sum_{l=1}^{L} Y_k(\vec{v}_l)h_l\right) * \beta_k\right) \tag{2.97}$$

for each binaural channel where K is the number of Ambisonic channels, L is the number of loudspeakers, $Y_k(\vec{v}_l)$ is the decoding matrix coefficient representative

of the Ambisonic channel, k, for the loudspeaker l, h_l is the HRTF measured from the position of loudspeaker l, and β are the Ambisonic input channels.

2.7 Machine Learning for Audio Production

The application of ML within the context of audio production often relates to two problem spaces. One is the application of *intelligent* adaptive signal processing algorithms based on analysis of the input signal and the use of a set of input/output mapping functions learnt from the training data and the second is the synthesis of music, speech, or sound effects to be used as content. Whilst both adaptive digital audio processing and audio synthesis have both long been areas of interest, the use of ML, and specifically neural networks, has seen rapid advancement during the last decade, particularly in the last half decade. This is largely due to an increase in the availability of the software tools and computational resources (such as Graphical Processor Units (GPUs)) to make the optimisation of ML systems tractable, along with, and equally as important, the quantity and quality of suitable training data with which to optimise such systems. Whilst this section focuses on the application of ML to audio production tasks, the reader is directed is to [195] for an in-depth theoretical review of Deep Learning and [196] for a more hands-on approach to understanding and training different neural network architectures.

2.7.1 Digital Audio Effects

Digital audio effects (DAFx), as defined by Verfaille [197], "*are boxes or software tools with input audio signals or sounds that are modified according to some sound control parameters and provide output signals or sounds*". Wilmering et al [198], provide a somewhat more focused definition of digital audio effects, one which is adopted in this thesis, where DAFx are viewed from the perspective of an audio engineer and refer to those processes commonly used in a music/post-production studio. Furthermore, the use of ML applied to DAFx can broadly be categorised as either parameter estimation or end-to-end transformation [199].

Using ML algorithms, such as neural networks, as parameter estimators involves using the networks to predict a set of parameters for the chosen audio processor or group of audio processors. In 2000, Reed [200], proposed a system for automatic frequency equalisation (EQ) using a Nearest Neighbour based approach that allowed a user to choose from 3 settings corresponding to a high frequency boost, a low frequency boost, and a flattening of frequency response. Frequency band parameter data for each condition was collected from participants and then used to train a Nearest Neighbour pattern matcher. Results showed that the Nearest Neighbour approach performed better than simply taking the linear average of the user provided parameter values.

Kolasinski [201], later proposed a framework for automatic level mixing based on representing the distance between a mix and a target mix as the Euclidean distance between the respective Spectral Histograms. A genetic optimisation algorithm was then used to approximate the required gain coefficients. Jillings and Stables [202] also utilised a genetic optimisation algorithm to predict suitable gain coefficients to balance a mix based on optimising for minimal auditory masking of tracks. They used the Masked-Unmasked Ratio (MUR), presented in [203], as the metric for the cost function as listening tests showed a strong correlation between a high subjective rating and a lower mean amount of masking. To ensure the training was not biased by differences in the relative levels of the mixes within the training set, each mix was normalised to 70 dBSPL (RMS). Results showed that whilst this method can be used to successfully create a balance mix, there were instances where the genetic algorithm would apply large reductions (between -47.27 dB - -53.05 dB) to a single track, resulting in the complete masking of those tracks within the mix. This was a result of the genetic algorithm exploiting a shortcoming in the definition of the masking metric such that whilst the MUR of the track that is heavily reduced would be increased, the MUR of the remaining tracks decrease, thus resulting in a better overall score. An improved cost function was then proposed which utilised MUR of the track with the most amount of masking applied, which penalises the heavy level reduction of a single track.

Other work has utilised estimated fader values as target data to train a linear dynamical system, which estimates the required gain coefficients for each track using a set of both spectral and time-domain acoustic features extracted from the input audio as input into the model [204, 205].

Chourdakis et al. [206], used a Convolutional Neural Network (CNN) as part of an investigation into modelling expert's decisions when assigning narrative importance of objects in an OBA radio drama. The aim of the study was to automate the assignment of narrative importance values to objects within object-based mixes, which then allows the user to attenuate parts of a mix by using a simple complexity parameter. These narrative importance values are traditionally assigned by mixing engineers. The CNN utilised was VGGish [207], a CNN trained to classify 632 classes found in AudioSet [208]. Transfer learning [195] was then used to leverage the prior learning contained within VGGish, with the final two layers being retrained to classify sounds as either speech, music, or sound effects. The class label determines the importance value assigned to the object and this is then used as input into a decision model, which is derived from data collected from audio production professionals and determines the amount of gain applied to a particular object.

ML has also been used to automate the parameter selection for artificial reverb algorithms. In [18], the application of specific reverb parameter values based on input audio features is approached as a classification task. The system is trained using audio features and desired IR characteristics as input, with the target output being reverb algorithm parameters. Benito and Reiss [209] used hinge-loss Markov random fields (HL-MRFs) with a set of Probabilistic Soft Logic (PSL) rules based on best practices as recommended by experts. The rules were then weighted based on the associated level of confidence derived from existing literature. This can be seen as a combination of a Knowledge-based system, as the PSL rules are based on gathered expert knowledge, and ML, as MRFs allow for the defining of probabilities based on logical relationships.

Sheng and Fazekas [210] proposed a Siamese DNN for learning a feature embedding from which Dynamic Range Compressor (DRC) control parameters can

be estimated given an unprocessed (uncompressed) input signal and a processed (compressed) reference signal. For the training phase the reference signal is the processed version of the unprocessed signal. For inference, an arbitrary DRC processed signal can be used. The study investigated the prediction of attack time, release time, ratio, and threshold. A Siamese neural network structure consists of two identical sub-networks and can be suitable when a model needs more than one input or branch and when all inputs are from the same domain [211]. This proposed system employed a CNN structure with one branch receiving the unprocessed audio, and the other receiving the processed reference. The subsequent feature embedding is formed by the difference between the outputs of the two branches. Once the feature embedding has been trained, the embedding is then used as the input feature vector to train a random forest regression model [196] to predict the final parameter values. As noted by Ramirez [212], parameter estimations tend to lack wider generalisation as they are often based on fixed audio processing architectures.

End-to-end methods, however, describe systems where raw audio is both the input and output of the system and is predicated on the idea that the complete mapping from input to output signal can be represented within the latent space of the network used [199]. Long-Short-Term-Memory (LSTMs) were investigated in both [213] and [214] to model static configurations of tube amplifiers. In [213], a real-time emulator is proposed that utilises a LSTM with a CNN added to the input, the addition of the CNN enables a reduction in the length of the required LSTM, whilst maintaining the accuracy of the model. The model proposed by Zhang [214], instead utilises LSTMs with many layers but a small hidden size in each layer, although this did result in audible differences being reported between the resulting model and the target device. A feedforward variant of WaveNet [21], was presented in [215], which unlike the models proposed by [213, 214], allowed the model to be conditioned with user control settings enabling the model to represent various control setting configurations. Damskägg [215] notes, that in cases where a static configuration is modelled, a separate model must be estimated for each different control configuration. This work was then built upon

in [216] to include a number of different popular audio distortion pedals and valve amplifiers. Results showed that larger models are required when emulating amplifiers with a higher amount of distortion.

A time-domain approach to arbitrary EQ matching is proposed in [217], which utilises an encoder-decoder CNN configuration and maps the waveform into the latent space via time-domain convolutions. The latent space is then modified by a DNN consisting of fully-connected dense layers before being upsampled by the decoder. This general architectures requires no knowledge of the filter banks, frequency bands, or filter types of the target signal and therefore makes it suitable for matching to arbitrary frequency responses. This work was expanded in [218] and applied to the modelling of distortion effects in order to develop a general purpose end-to-end DNN that can be used to model numerous non-linear effects.

Hawley [219], proposed an end-to-end method of modelling non-linear effects with longer temporal dependencies, such as DRCs. The architecture is a combination of U-Net [220] and Time-Frequency [221] net, and utilises input-output measurements in conjunction with conditioning the network on given input-output pair parameter values. The desired mapping function from unprocessed to processed is therefore learnt/optimised for without any explicit internal compressor/effect model. This methodology can be considered as effect-agnostic and has the potential to be applied to a number of other non-linear effects.

2.7.2 Audio Synthesis

Much like the application of ML to DAFx, audio synthesis using machine learning is still a relatively new area of study, one that has seen rapid progress during the last 5-10 years from within both academia and industry. Within the context of audio production for IMEs, audio synthesis can be applied to speech synthesis [222, 223], sound effect synthesis [224, 225], and musical sound synthesis [222, 226].

WaveNet, presented in [21] and based on the prior work in [227, 228], is a fully autoregressive model for generating raw audio, where each predicted sample

is conditioned on (influenced by) all previous samples. It is a fully convolutional network that leverages dilated convolutions [229] to allow a much larger receptive field than would be possible with non-diluted convolutions given the same number of parameters. This allows WaveNet to capture and model longer-term temporal dependencies within audio signals. SampleRNN, proposed in [230], is another end-to-end time-domain audio synthesis model, but one that utilises RNNs to model temporal dependency. It utilises a hierarchy of modules, each operating at different temporal resolutions. Lower modules operate at a higher temporal resolution, with the lowest module processing individual samples and higher modules operating over longer timescales and thus at a lower temporal resolution.

Even when utilising methods such as RNNs and dilated convolutions, long-term dependencies can be difficult to model in the time domain. Vasquez and Lewis [222] address this in their proposed MelNet, which models 2D time-frequency domain representations, such as spectrograms, rather than 1D time-domain waveforms. This results in a temporal axis in the time-frequency domain which can be much more compact than that of raw waveforms, meaning that dependencies that span tens of thousands of timesteps in the time domain only span hundreds in the time-frequency domain [222]. Results showed that MelNet could be applied to a variety of end-to-end audio generation tasks including unconditional speech generation, music generation, and text-to-speech synthesis.

Generative adversarial networks (GANs) [231], have been used to model both waveform [224] and spectral representations [226]. Donahue, McAuley, and Puckette [224], proposed WaveGan as a model which utilised GANs for the unsupervised synthesis of raw-audio waveforms. It was based on the deep convolutional GAN (DCGAN) [232], but modified to have a flattened architecture to operate in one dimension, thus making it suitable for time-domain audio synthesis. Given that the output of DCGAN is a 64x64 pixel image, which equates to 4096 samples when flattened, an additional layer was added to the model resulting in an output length of 16384 samples, which equates to 1.024 seconds at the 16 kHz sampling rate used in the study . The conditioning of WaveGan architectures on class labels was then investigated in [233], which built

upon previous work in conditioning GANs for image synthesis [234]. Although it was noted by the authors of [233], that the synthesised audio resulting from a model conditioned on speech did contain a noticeable amount of unwanted noise. A conditional WaveGan was also used by Barahona and Pauletto [225], to synthesis knocking sound effects with emotional intention. Results showed that for persons without sound design experience the model was close to synthesising samples that were indistinguishable from their recorded counterpart. However, those with sound design experience could easily identify the synthesised samples from the recorded samples. Emotional intent of both the recorded and synthesised samples were, on average, correctly identified, although both those with and without sound design experience confused *fear* with *anger*. This highlights that, even with distinguishable discrepancies between the recorded and synthesised samples, emotional characteristics are successfully encoded within the latent space of the model.

In the previous two years, diffusion models have surpassed GANs to become the state-of-the-art in generative modelling. They are now applied to a range of generative modelling tasks including multi-model generation e.g. text-to-image, text-to-video, and text-to-audio, natural language generation, and audio waveform generation and processing [235]. A key aspect of all diffusion approaches is the progressive addition of random noise to the data, after which the noise is then iteratively removed to generate new data samples. Whilst it is outside the scope of this thesis to outline in detail the mechanisms involved in diffusion based models, those interested are referred to [235], which provides a detailed review of their methods and applications.

There are a number of recent diffusion approaches to the task of text-to-audio generation, which includes both text-to-speech and text-to-nonspeech sounds. Popov et al. [236] proposed Grad-TTS, a diffusion model for text-to-speech generation, whereby noise predicted by the encoder is gradually transformed and aligned with the text input on which it was conditioned. This was improved upon in Grad-TTS2 [237], by the addition of a speaker-dependent phoneme classifier providing an adaptive text-to-speech system. Text-to-nonspeech sounds hav been

cited a more challenging problem, as unlike text-to-speech there is not necessarily a direct correspondence between the written text and the resulting sound [238].

Diffsound, presented in [238], is a non-autoregressive decoder which transfers features extracted from a text encoder directly to a mel-spectrgram. A vocoder is then used to transform the generated mel-spectrogram into a waveform. Additionally, Diffsound predicts all the mel-spectrogram tokens in a single step and then refines the predicted tokens over subsequent steps. Results showed improved generation results when compared to autoregressive decoders as well as an increase in generation speed. Haohe et al. [239], proposed AudioLDM, a text-to-audio system that learns continuous audio representations through contrastive language-audio pretraining (CLAP) latents [240]. Pre-training with CLAP enabled AudioLDM to train with audio embeddings, whilst text embedding could be used to condition the model during inference. Similar to Diffsound, AudioLDM generates mel-spectrograms and then employs a vocoder, in this case HiFi-GAN [241], to generate audio samples from the reconstructed spectrogram. Results show good performance in the generated text conditioned sound effects, speech, and music. The use of text conditioning also enables text-guided audio manipulations, such as style transfer.

2.8 Summary

This chapter has provided a foundation in areas relating to sound and audio signals, on which the rest of this thesis is based. This includes the physical properties of sound waves, their propagation through space, and their perception via the human auditory system. It was discussed how sound is represented in the digital domain as discrete measurements and how, through the use of the FFT and STFT, a signal can be transformed into the frequency-domain and time-frequency-domain, respectively, for analysis and/or processing. This chapter then presented a simple definition of what constitutes a soundfield alongside a brief summary of the basic mechanisms of soundfield recording and encoding, with a focus on the main representation formats used for IME production. This chapter

then concluded with a brief overview of the applications of ML with respect to audio production with a focus on DAFx and audio synthesis. Whilst this chapter focuses on the technical aspects of sound and digital audio, the next chapter will introduce sound and sound design within the context of IMEs including providing a definition for what, within the context of this thesis, constitutes immersion and by extension an immersive media experience, the different types of IMEs, and how spatial audio can be utilised within them.

Chapter 3

Sound Design for Immersive Media Experiences

3.1 Introduction

As this thesis is interested in developing new methods to assist in the sound design for IMEs, it is necessary to provide an introduction to how sound is designed and utilised within IMEs complementing the technical introduction to sound and audio signals given in the previous chapter. The term *immersion* will be defined within the context of this thesis. By extension, it will also explain what constitutes an IME and will provide an introduction to how sound design and the utilisation of spatial audio is approached for IMEs, as well as briefly outlining the main types of IMEs that are commonly encountered.

3.2 Defining Immersion

The term *immersive* is often used vaguely and interchangeably with related terms such as *realism*, *naturalness*, *involvement*, absorption, and *presence* [242, 243]. This inconsistency within the terminology can cause confusion, both for consumers and for those undertaking research in the area [244]. This can be further complicated when taking into account the multi-sensory nature of many IMEs.

Although there is, as yet, no standard definition of immersion, current literature supports the idea that immersion is a multi-faceted concept. A recent study by Eaton and Lee [245] identifies two overarching categories of immersion; *passive immersion* and *active immersion*. Passive immersion is defined as being related to a feeling of presence or being in an environment [245], and encompasses previously defined notions of sensory immersion [246] and perceptual immersion [247]. Both of which requires the user's perceptual systems to be submersed in the environment, but have no prerequisite for the user to play an active role in the experience. Examples of such experiences would be non-interactive VR, music, and soundscape recordings utilising 360° video and/or audio systems. The intent of these 360° audio-visual systems is to provide such perceptual and sensory submersion through surround sound (if audio only), or multi-sensory audio-visual stimuli. McMahan [248] describes this as constraining the user's perception to the presented stimulus and as such blocks out the external world. Active immersion relates to immersive media with an interactive (task-based) element [245], for instance video games, where the user needs to make choices or be constantly attentive due to the task at hand [249].

Another cause of immersion pertinent when discussing IMEs is the narrative presented to the user, in which they may or may not have direct involvement, and results in an attention shift towards the story and away from the physical environment [250]. This is related to the concept of *imaginative immersion* [246] where users relate to or are emotionally invested in the characters and events within the experience itself. This dimension of immersion, though defined within the context of video games, can also be associated with the immersion experienced when reading an engaging novel or listening to a radio drama.

Other forms of immersion related to the narrative are that of *temporal immersion*: focused attention on an unfolding story [250], where the user is interested in, or in a state of anticipation for, what comes next, and *emotional immersion*: where an attachment is formed with the characters in the story [251]. Immersive content will often combine or aim to elicit several dimensions of immersion; for instance the binaural version of the Doctor Who episode 'Knock

Knock' [252] combines aspects of perceptual, narrative, and emotional immersion. The binaural audio serves to elicit perceptual immersion by providing a spatial sound scene, whilst traditional storytelling devices aim to provide both narrative and emotional immersion, i.e. cliffhangers to create tension and anticipation, and character development.

Contrary to those who regard immersion as a cognitive phenomenon, there are those who regard immersion as being an intrinsic, objective property of a system. In other words, the more advanced the system is at replicating the relevant perceptual stimuli, the more immersive it is considered to be. Slater [253] argues that the term *immersion* should be reserved, "to stand simply for what the system delivers", and does not see immersion as a subjective experience. Rucella [254], also suggests a distinction between *immersion*, which is related to the ability of a system to produce sensory stimuli and can be quantified by the number and types of stimuli, and *presence*, argued as the cognitive result of the immersive capability of a system which is difficult to quantify. However, this idea has been rejected by others in favour of regarding immersion as a psychological or cognitive experience that can be caused by both technological and non-technological processes [244, 246, 250, 251].

Given the many definitions of *immersion*, some of which overlap and some of which conflict, it is important to have a clear definition of *immersion* within the context of this thesis. For the purposes of this research, it was deemed appropriate to use a definition which is broad in scope but also captures the multidimensional nature of *immersion*. The following definition, as defined by Agrawal et al. [244], is therefore adopted:

> Immersion is a phenomenon experienced by an individual when they are in a state of deep mental involvement in which their cognitive processes (with or without sensory stimulation) cause a shift in their attentional state such that one may experience disassociation from the awareness of the physical world.

With respect to potential technological innovation, it is important to identify

the spatial audio technologies used to target the elicitation of perceptual and sensory immersion and, as such, focus on IMEs that utilise these and related XR technologies. Within a broader exploration of sound design practice it is also interesting to understand how creators of IMEs utilise and target the other dimensions of immersion. Using this definition allows both areas of interest within the scope of the investigation to be considered. Though not a requirement, many immersive media experiences utilise technologies such as 360° audio and video which provides a subjective sense of being surrounded and the multi-sensory stimulation associated with previously discussed forms of immersion [246][247]. The rise in popularity of this content has resulted in companies such as the BBC, Facebook, and Google, releasing tools [255–257] and producing content for IMEs. However, a question could be asked as to whether the current tools cater to the needs, and wants, of content creators within the wider sector.

3.3 Immersive Media Experiences

Immersive experiences are therefore experiences which should aim and succeed in eliciting a state of immersion through either sensory (auditory or visual stimuli) or cognitive (investment in characters/narrative) processes. As this thesis focuses on developing novel machine learning applications for the sound design process, it is the technology-driven immersive experiences that are most relevant to this research. That is not to say that aspects of narrative and user involvement are not important, but that those techniques are often used in conjunction with and augmented by a variety of emerging technologies. This thesis therefore uses the term Immersive Media Experiences (IMEs) to refer to immersive experiences that are delivered or facilitated by technology. It should be noted that in much of the recent literature the term *immersive experience* is often used synonymously for experiences that utilise, at least in part, technology to facilitate immersion.

IMEs are often delivered via Extended Reality (XR) technologies, a term that encompasses Virtual Reality (VR), Augmented Reality (AR), Mixed Reality (MR) and associated technologies [258], and includes training simulators [1],

Level	0	1	2	3	4
Interactivity	Passive	Participatory	Physicalized	Problem Solving	Interpersonal
Embodiment	Detached	Watcher	First-person POV	Movement	Human2Human Interaction
Co-Participation	Single-Player	One-on-One	Secondary Perspective	Group	MMO
Story	No Story	Setting	Pre-Created	Choose Your Own	Interactive Story
Dynamics	Pre-determined	Choice	Multi-Thread	Free Will	Convo Reality
Gamification	Ungamified	Instruction	Reinforcement	External Process	Reward System
Immersive Tech	**None**	**AR**	**360 Media**	**VR**	**XR**
Meta Control	No Meta Control	Journey	Character	World Builder	World Master
Didactic Capacity	Elemental	Explicit	Implicit	Recall	Synthesis

Table 3.1: Taxonomy for elements/dimensions of an immersive experience. The level for the corresponding element and indicates its range of depth. The levels associated with the Immersive Technology element (in bold) can be considered the broad categories of IMEs. The table is adapted from [254].

multi-channel music mixes/soundscape recordings [2], 360° video, and video games utilising VR and/or spatial audio [3]. Ruscella [254] offers a taxonomy of immersive experience design (reproduced in Table 3.1), which details nine dimensions of immersive experiences, each of which has five elements scored from (0-4) to indicate the level of immersion the authors feel is equated with that particular element. The overall score of the experience is then determined by the summation of its element scores. As can be seen from Table 3.1, the elements within the dimension referred to as immersive technology outline what might be considered as the broad categories of IMEs. However, the assertion by Ruscella [254], that XR activities mix virtual components into a real-world environment, could be argued to be more accurate when used to describe Mixed Reality (MR) experiences, given that XR is often used as a collective term to collectively describe AR, VR, and MR. Therefore, this thesis suggests and the main categories of IMEs are Augmented Reality, Virtual Reality, Mixed Reality, and 360° Media. Although this thesis is concerned with the production of audio for IMEs, as opposed to the complete XR technology pipeline, a brief overview of each type of experience is now provided for completeness alongside selected examples to provide context to the reader who may be unfamiliar with these types of experiences.

3.3.1 Augmented Reality

The term AR often refers to systems which superimpose digital assets over a real-world view [259], but those digital assets not necessarily interacting with physical objects within that space. The term was reportedly first used in relation to a project at Boeing which superimposed information on the visual field to aid workers laying aircraft cables [260]. In recent years, AR has been applied to both the entertainment and education of consumers, alongside industrial training applications. One of the most well known AR applications of recent years has been Pokémon Go [261], which is a location-based AR games where users can battle, capture, and train virtual Pokémon that superimposed onto the user's world and viewed through a mobile device, such as a smart phone.

As XR technology has developed, varying definitions have been given for what constitutes an AR system. A widely accepted description of AR, which we will map onto Pokémon Go experience, is that given by Azuma [259], who proposes that for a system to be considered AR it must possess three characteristics; it must combine the real and virtual, be interactive in real time, and registered in three dimensions. Related to Azuma's first characteristic, Drascic and Milgram [262] focus on the visual domain defining an AR system as one that displays an image which is predominately the real environment, but is enhanced or augmented with digital assets. More recently, descriptions of AR have made specific references to the devices through which such experiences are deployed, most notably smart and wearable devices such as smartphones and smart glasses [263]. It could be suggested that these more recent descriptions, at least in part, are a result of the ubiquity of smartphones capable of delivering AR experiences such as Pokémon Go, whereas in previous years more expensive and larger Head-Mounted-Displays (HMD) would have been needed. Doerner et al. [264] offers a general definition of AR, which encompasses not only the augmentation of visual perception but also the broader perceptual perspective of AR which can include the augmentation of any sensory experience:

> Augmented Reality (AR) refers to the immediate and seamless percep-

> tion of the real environment enriched by virtual content in real-time, the latter resembling reality to the largest extent possible regarding its characteristics, appearance, and behaviour, so that (if desired) sensory impressions from reality and virtuality may become indistinguishable (for any senses). [264, p. 19]

As the majority of AR is now experienced through lightweight portable devices, they are not required to be tied or even associated to a specific location, which has in turn given rise to a variety of different AR applications. A non-exhaustive range of applications include augmented retail experiences where users are able to view digital representations of items on themselves [265], furniture in their home [266], through to appearance modification by way of facial augmentation [267], and practical applications such as superimposing signs and directional information to aid user navigation [268]. Pokémon Go [261], takes advantage of this portability and utilises the GPS functionality of many mobile devices to create a location-based experience where digital assets can be anchored to, and persist in, physical locations. This also allows the world map to be based on the geographical location of the user, as shown in Figure 3.1, which depicts the world map of a user located at Wellingborough train station, London, UK. Using the location data from the user's device, digital assets associated with given location can be overlaid at the appropriate positions relative to the user's device. Whilst Pokémon Go uses geolocation to create a worldwide AR experience, it can also be used to create multiple location-specific experiences, such as StoryTrails which has developed site-specific AR experiences for 15 locations across the UK [269].

Although AR is usually associated with the augmentation of visual content, AR-audio has also long been an area interest and aims to augment the real auditory environment with virtual audio objects [270]. Darkfield Radio [271], an application developed by DARKFIELD [272], delivers AR experiences with 360° immersive audio to at-home audiences through their smart phones and a set of headphones. One such experience, *Visitors* [273], instructs two participants to sit opposite each other in a living room with each participant receiving their own audio track. There are also location-based AR-audio experiences, such as Ghost

Figure 3.1: Wellingborough train station as depicted on the Pokémon Go app.

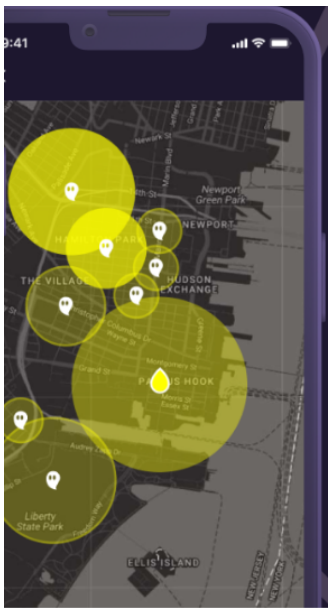


Figure 3.2: Illustration of the map presented during Ghost Walk to guide users (yellow icon) to different points of interest (ghost icons). Taken from [274]

Walk [274], which is a GPS facilitated audio experience through the Broadgate and Finsbury Circus areas of London, UK. Users can walk around the area, guided by the map shown in Figure 3.2, with different pieces of audio content fading in and out depending on the location of the user.

Whilst a large proportion of AR experiences are targeted for mobile devices, such as smart phones, there is a history of technology companies developing other devices to facilitate AR. Most of these fall into the category of Smart Glasses, being designed to present visual and/or auditory information alongside what the user already sees/hears. Bose Frames [275], were Bluetooth enabled audio sunglasses that could be paired with a user's smartphone to deliver AR audio experiences. Consequences [276], was a location-based narrative delivered through Bose Frames, where the audience is able to move freely around the physical environment and interact with other performers. The audio was delivered entirely though the Bose Frames and enabled the audience to choose their path through the choose-your-own-adventure-style narratives. Whilst Bose discontinued the Bose Frames and its associated AR projects, other companies have continued

to develop within the this particular area. Meta have released the Ray-Ban | Meta [277], a collaboration with sunglasses manufacturer Ray-Ban, which do not include any visual AR, but do provide audio through open-ear speakers, similar to that in the Bose Frames, alongside additionally functionality to capture and stream images and video to linked Meta social media accounts. Maverick Smart Glasses by Everysight [278], offer a set of glasses that come with either a tinted or clear coated visor on which apps and information can be projected, however does not offer any audio functionality. Google presented a demo of AR glasses being used to provide real-time speech-to-text language translation [279], which builds on their use of AR translation within their Google Lens application [280], which itself serves to provide visual input into their search engine. AR glasses, however, should not be confused with MR headsets, examples of which are detailed in Section 3.3.3, which are often bulkier and designed to deliver higher fidelity experiences, as opposed to simply displaying additional information.

3.3.2 Virtual Reality

In contrast to AR, VR aims to replace the user's sensory perceptions of the real world with that of a wholly computer-generated virtual world [281], which will often target multiple modalities including visual, auditory, and, occasionally, haptic. To block user perception of the real world, VR experiences are delivered by HMDs equipped with a stereoscopic display (one screen for each eye) with spatial audio being delivered through headphones or loudspeakers built into the headset. VR experiences can also afford a greater degree of agency within the environment alongside a higher level of interaction between the user and the environment; this is achieved through 3D tracking. The extent to which the user is tracked within the virtual environment is dependent on the specific hardware that makes up the VR system. Flood by Megaverse [282], is a multi-user location-based VR interactive theatre experience that utilises a Vicon tracking system [283]. This not only enables the tracking of users within the performance area but also enables motion capture of the participants' limbs, which can then be mapped to avatars in the virtual world. Figure 3.3b, shows an example of the

types of body trackers used in such experiences. As all VR experiences require the use of a HMD, head movements of the user are used to facilitate the tracking of their orientation and position within the environment. This enables what Doerner [264] refers to as "viewer-dependent image generation" and operates on both visual and auditory stimuli by rotating both the visual and auditory scene scene to counter the head movement of the user. The aim of movement tracking in VR, whether it be solely head-tracking, or full-body-tracking, is to take the movement of the user within the physical space and map them to corresponding movements in the virtual space [284]. At the time of writing, Flood utilised the HTC Vive Focus 3 headset [285], shown in Figure 3.3a, which supports a performance area up to a recommend maximum of 10m X 10m and includes hand tracking functionality. Although multi-user experiences, such as Flood, may employ a dedicated tracking system, many headsets come with built in standalone tracking functionality, often referred to as inside-out tracking [286]. This type of tracking utilises sensors or cameras are mounted on the device itself, which look outward into the environment to track the position of the user. Systems, such as those provided by Vicon, are often referred to as outside-in tracking, as the sensors or cameras are placed in static locations around the performance area and they track markers, such as the Vicon pulsar markers [283], that are placed on the objects to be tracked.

Many HMDs now support hand tracking [287–291] as this allows the user to interact with the environment in a way that is more intuitive and natural when compared to traditional computer-human interfaces such as games controllers, and mice and keyboards. In the case of Flood it also allows for physical contact between participants. Due to the additional hardware and set up required, VR experiences are less portable than AR experiences and are normally experienced within a defined play area at a specific location.

In addition to tracking users within the play area, the same hardware can be used to track objects within the physical world and map them to objects within the virtual world. Figure 3.4, shows a split-screen view of Flood, where trackers have been placed on light-weight physical props and mapped to virtual objects.

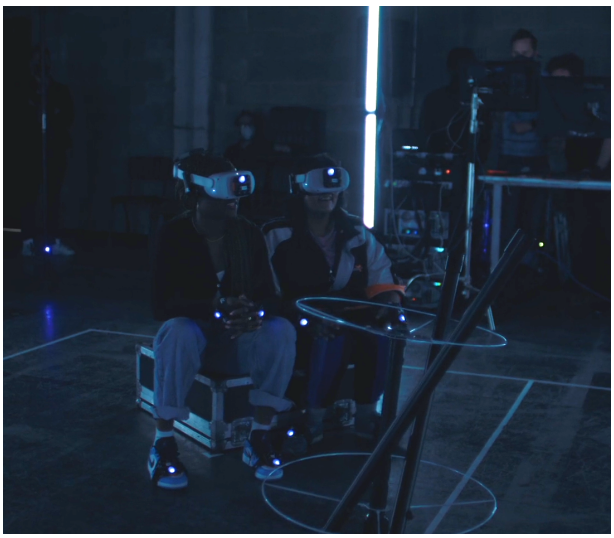
(a)

(b)

Figure 3.3: (a) The HTC VIVE Focus 3 Headset. Image taken from [292]. (b) Pulsar gloves as part of the vico tracking system used for Flood. Image taken from [293]

This aims to reduce the friction between the physical and virtual body of the user by not only providing the user with visual feedback when they use their hands to manipulate virtual objects but also by integrating haptic feedback from the physical counterpart of the virtual object. This serves to further align the user's physical self with their virtual self by increasing the number of sensory pathways integrated into the the virtual world.

Whilst Flood can be considered a commercial location-based experienced, there are also experiences designed as at-home experiences, which tend to require less equipment, a smaller play area, and are able to run on cheaper, lower specification hardware. At the time of writing, HMDs such as the HTC Vive Cosmos [288] and the Meta Quest2 [287] are marketed as at-home VR headsets. The differences between commercial and at-home experiences and devices are usually scale, with respect to the the parts of the user that are able to be tracked, and the processing capabilities of the devices, and the required size of the play area. Table 3.2, presents a selection of at-home and commercial devices specifically marketed for VR, as well some of their key specification and features. The target



(a) View of Flood from the real-world

(b) VR view of image (a)

Figure 3.4: Two views of the same scene from Flood, taken from [282]. (a) shows the view of the real-world with participants sat in front of an object with stick like objects protruding out of it; (b) shows the same scene from the virtual world and as can be seen, the stick like objects present in (a) are mapped to their virtual counterparts.

market indicates whether the device is for consumer at home use, productivity where it may be used at home or in an office/workplace setting, professional, where the device would be used for commercial location based experiences, such as Flood. As VR headsets are often sold with the option to purchase additional extra

hardware, standardisation of prices are not straightforward. The prices listed are the cost of the standard headset with a set of controllers (if required), and any additional hardware required to provide 6DOF tracking with an environment. In the case of the Varjo-Aero [294], controllers and tracking hardware must be bought separately from a different vendor.

Given that VR require users to purchase additional hardware, such as the headsets, the barrier for entry is higher than that of AR. There are also, at the time of writing, a greater number of consumer level headsets catering to VR than there are wearable AR devices given that most AR can be deployed on a smart phone, which at present, have a greater reach in terms of audience. As briefly mentioned, there is however a divide between at-home VR devices and commercial/industrial VR devices.

3.3.3 Mixed Reality

Mixed Reality (MR), like immersion, is another term where a lack of consistency in its use is observed within both academic and professional literature [302]. Whilst MR shares similarities with AR, MR often refers to the combination of real and virtual content. Milgram et al. [303] proposed MR as a continuum, illustrated in Figure 3.5, that spans from reality to virtuality, whereby the amount of reality present decreases whilst the amount of virtuality increases. More recently, Rokhsaritalemi [304] describes MR as the merging of real and virtual worlds that results in real-world objects interacting with virtual objects. They continue to propose three features important to any MR system; combining the real-world object and the virtual object; real-time interaction; and mapping between the virtual object and the real object to create interactions between them. Flavian [302] also argues that MR should no longer be considered as a broad part of the continuum that includes AR, but should instead be regarded as a specific point on the continuum: this point Flavian [302] refers to as Pure Mixed Reality, but will be referred to as MR for the purpose of this thesis.

The key difference between MR and AR is the extent to which virtual objects are integrated and interact with the real components of the environment. Taking

Table 3.2: Selection of VR headsets and associated specifications. All specifications and costs correct at time of writing. Prices may vary depending on retailer. *price from Amazon.

Model	Cost	Display resolution (pixel per eye)	Audio	Tracking	Standalone/PC VR	Target Market
Meta Quest2 [287]	£249.99	1832 x 1920	Headphone & built in loudspeakers	Inside-out	Both	Consumer
Pico Neo3 Link [291]	£307.99*	1832 x 1920	Built in speakers	Inside-out	Both	Consumer
Valve Index [295]	£919	1440 x 1600	Built in speakers	Outside-in	PC VR	Consumer
HTC Vive Cosmos [288]	£549.99*	1440 x 1700	built in speakers	Inside-out	PC VR	Consumer
Varjo Aero [294]	€990 (Headset) £537 (controllers and two tracking basestations)	2880 x 2720	in-ear headphones	Outside-in	PC VR	Professional
HTC Vive Focus III [285]	£1,372	2448 x 2448	Built in speakers & duel microphones	Inside-out	Both	Professional

Table 3.3: Selection of MR headsets and associated specifications. All specifications and costs correct at time of writing. Prices may vary depending on retailer. *price from Amazon.

Model	Cost	Display resolution (pixel per eye)	Audio	Tracking	Standalone/PC VR	Target Market	Type of MR
Pico 4 [296]	£379.00*	2160 x 2160	Built in speakers	Inside-out	Standalone	Consumer	Video-passthrough
Meta Quest3 [297]	£479.99 (128 GB)	2064 x 2208	Headphone & built in loudspeakers	Inside-out	Both	Consumer	video-passthrough
Meta QuestPro [289]	£999.99	Not Specified	Built in speakers	Inside-out	Standalone	Consumer	Video-passthrough
HTC Vive XR Elite [298]	£1,299	1920 x 1290	Built in speakers & duel microphones	Inside-out	Both	Consumer/Productivity	video-passthrough
Microsoft Hololens 2 [299]	$3,500	see through holographic lenses	Built in speakers	Inside-out	Both	Industrial	optical see-through
Varjo XR-4 [300]	$ 3,990	3840 x 3744	Built in speakers	Both	PC VR	Industrial	video-passthrough
Varjo XR-3 [301]	Price not freely available	1920 x 1920 & 2880 x 2720	3.5mm audio jack	Both	PC VR	Industrial	video-passthrough

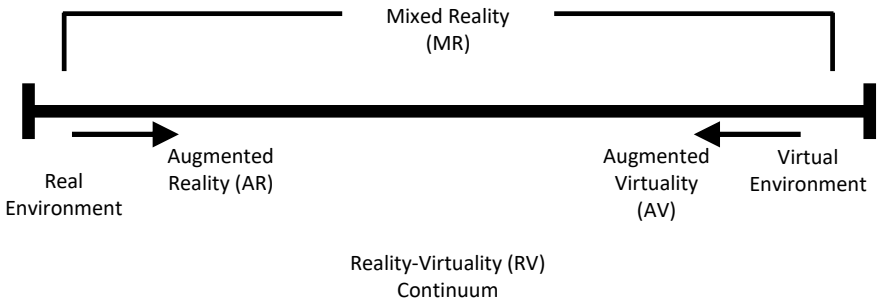


Figure 3.5: Simplified representation of Reality-Virtuality Continuum taken from [303]

the previous AR example of Pokémon Go, the virtual objects are superimposed onto the real world, but there is a lack of any interaction between the two, MR aims to merge virtual content into the real world such that both virtual and real content interact in real-time and appear to share the same space [302].

Interchange by Prox & Reverie [305], is a self-described advanced prototype for a multi-functional, physically anchored, MR portal system and was created as part of the 5G futures programme run as a partnership between XR Stories [306] and Warner Media [307]. It allows users to cross from the real world into a virtual world using a digital arch, which is anchored over a physical structure, as a doorway between the two. This allows the user to not only pass between the two, but also allows virtual objects and avatars to travel between the real and virtual world. Consider Figure 3.6, which shows a frame taken from an Interchange demonstration [308] where a digital avatar has crossed through the archway into the real world. Compare this to Figure 3.7, where Figure 3.7a and Figure 3.7b depict the same Pokémon Go experience but with the device having been moved farther back in Figure 3.7b. Note how in Figure 3.7 the digital objects look as though a digital image has simply been overlaid onto an image of the real world, with a size that does not keep scale with its surroundings as the device is moved closer to or farther away from where the object appears to be positioned. It instead keeps a fixed size relative to the size of the screen. Interchange, arguably, depicts an object more convincingly appearing to share the same space as the

Figure 3.6: A frame taken from [308], showing a digital avatar crossing from the virtual world into the real-world.

real components of the scene with respect to positioning and scale within the real environment. Subsequently, MR systems will often require more advanced hardware and software to facilitate the integration of the real and virtual objects in a way that results in not only the user being able to interact with both sets of objects, but both sets of objects being able to interact with each other. Collins [309] refers to this as *visual coherence* which is a key component of MR.

MR systems also make use of HMDs, which at first may appear similar in appearance to those used within VR experiences. However, MR system HMDs require additional functionality which allows the real world to be displayed to the user. Table 3.3, presents a summary of currently available MR headsets and a selection of relevant specifications for comparison. At the time of writing, Interchange utilised the Varjo-XR3 [301], shown in Figure 3.8, which is, at the time of writing, marketed as an industrial MR device and utilises video-passthrough to achieve an MR scene, which captures the real world using headset mounted cameras. Virtual assets are then superimposed directly onto the image presented to the user with LiDAR ensuring correct depth of field for the passthrough image.

(a) (b)

Figure 3.7: (a) and (b) depict the same Pokémon Go experience but with the device having been moved farther back in (b). This illustrates that in AR objects do not always have capability to keep scale with their environment but instead just have a fixed sized relative to device screen size.

It is worth nothing that any MR headset that uses video-passthrough is also capable of delivering full VR experiences. As shown in Table 3.3, other MR headsets that utilise video-passthrough include the Meta Quest3 [297], the Meta Quest Pro [289], the Varjo-XR4 [300] (which is the successor to the previously mentioned XR-3), Pico 4 [296], and HTC Vive XR Elite [298]. As can be seen from Table 3.3. Whilst products from Varjo are specifically aimed at the industrial use case, those from Meta, HTC, and Pico are, currently, marketed as either entertainment or productivity devices. Other headsets, such as the Microsoft Hololens [299], utilise optical seethrough MR, where, similar to the AR glasses discussed in Section 3.3.1, the assets are projected onto the lenses.

Figure 3.8: The Varjo-XR3 MR headset used in Interchange. Taken from [301].

3.3.4 360° Media

Whilst AR, VR, and MR can be considered 360° experiences, the term 360° Media is used to refer to those experiences that lack the interactivity or freedom of movement of the previous categories, but still provide content which surrounds the user. Ruscella [254] describes 360° media as immersive videos which surround the user in a photo-realistic environment and typically offers three degrees of freedom. A typical example of 360° media would be cinematic VR, which Mateer [310], describes as a type of IME where the user is able look around the virtual world in 360° and which is usually accompanied by spatialised audio. These experiences are usually presented through either a HMD, such as the ones detailed in Tables 3.2 and 3.3, a smart phone/tablet, which then behaves as a window into the virtual world, or a web portal, such as YouTube [311], which would require the user to click and drag around the screen to change their orientation. A good example of 360° Media is BBC's Click 360 episode [312], which was the first TV episode filmed entirely in 360° and allows the user to explore the view around them in a variety of locations. Alongside 360° video, multi-channel/spatial audio content, such as that detailed in section 2.6, is also included in this category as it aims to provide the auditory equivalent of "photo-realism" to the user. Many 360° videos are often captured using traditional filming methods except with cameras containing two or more lenses, resulting in a 360° field of view. Once the video is captured each image captured by the different lenses within the

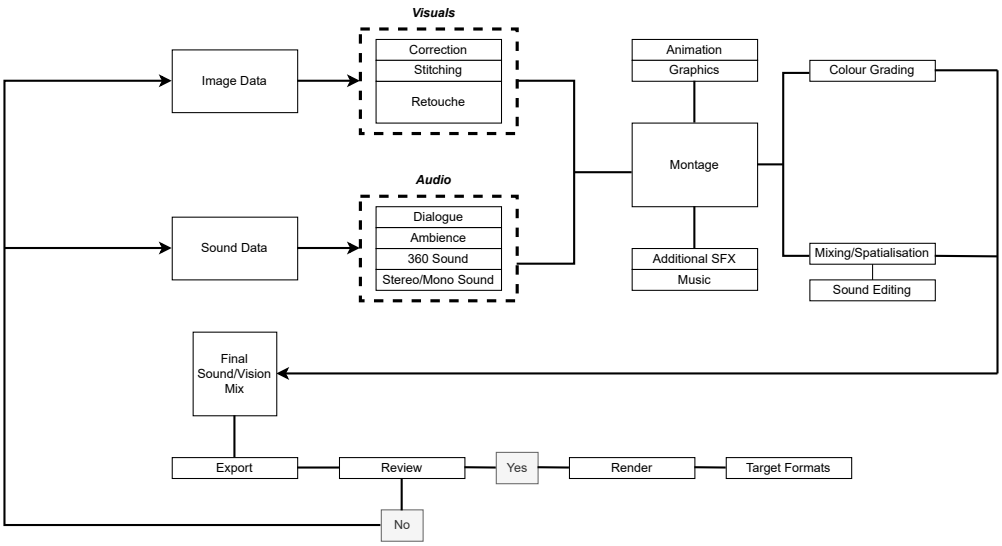


Figure 3.9: Post-production workflow for a 360° film. Adapted from [314]

camera then need to be integrated into a single surround video, through a process referred to as stitching [313], to generate the 360° image. Figure 3.9 shows a generic post-production workflow for a typical 360° film. The process shares many similarities to that of traditional film production, but with the addition of spatially recorded sound and the output from 360° cameras. Some of the challenges associated with this are explored in Chapter 4.

From this perspective, 360° Media can be considered closer to traditional linear media than to the interactive experiences afforded by AR, VR, or MR as scenes are captured, edited together, and the result is then simply replayed to the user in a predetermined order. As such, the user does not have a physical representation within the experience. This is exemplified by Figure 3.10, which shows a frame taken from the BBC's Click 360 [312] with the user orientating their view towards the ground. All that can be seen is the base of what appears to be the camera stand.According to Ruscella's taxonomy [254], this would result in the experience not being able to surpass a score of 1 on the interactivity dimension. Additionally, as the majority of 360° and/or spatial audio experiences are passive in nature, most would score 0 within the interactivity dimension. The exception to this would be interactive films such as Afterlife [315], where the user

has a limited participatory role when choosing between a set of pre-determined options, each with pre-determined outcomes. By observing again Ruscella's proposed taxonomy in Table 3.1, experiences within this category would score poorly across all dimensions of IME design apart from immersive technology, where 360° Media scores higher than AR. It should be noted, however, that this is due to the taxonomy subscribing to the notion that immersion is an objective measure of the ability of a system to produce sensory and multi-sensory stimuli surrounding a listener. Given the definition of immersion used in this thesis as defined in section 3.2, it could be argued that, provided sufficient production quality, AR could be more likely to elicit a state of immersion given the greater degree of interactivity afforded to the user, particularly if these augmented visuals are accompanied by augmented and/or spatial audio.

Figure 3.10: A frame taken from BBC 360 Click [312] where the user has orientated themselves to face the ground. This is an example of how within 360 Media, users often lack a physical representation within the space as all that can be seen in this scenario is the base of the camera stand.

3.4 The Role of Sound in Immersive Experiences

As experiences under the IME umbrella can be both linear and non-linear in nature, the role and approach to the use of sound and sound design takes from practices associated with both traditional linear media e.g. film, TV, and radio and non-linear/interactive media e.g. video games. Both types of media have

been written about extensively [8, 9, 316, 317] and will be explored in this section within the context of IMEs. Sinclair [316] proposes that the multi-faceted use of sound can be described by three main roles: to *inform*, to *entertain*, and to *immerse*. However, as the role of entertainment has arguable overlap with the role to inform and to immerse it will not be considered in isolation but in conjunction with the other two roles. Whilst these terms were used to describe the use of sound for video game and VR environments, they are, as will become apparent, appropriate umbrella terms with which to explore the role of sound within both linear and non-linear IMEs.

3.4.1 Inform

Sound has an important role within various forms of media, providing information about both the narrative [9] and the user's surroundings [316]. Murray [9] explains that sound provides direct narrative through dialogue and exposition, as well as subliminal narrative through the use of music and sound effects to guide and influence the user's emotional response. Sound, therefore, can be used to communicate factual or emotional context [318] which may not be obvious from the visual scene alone. This is used to great effect in factual programming, such as the previously mentioned BBC Click 360, where, without the accompanying dialogue (both onset and overdubbed), much of the on screen action in isolation would not provide an adequate amount of information for the average audience member to fully understand what is taking place. Additionally, the emotion elicited by a scene can be changed by manipulation of the soundscape associated with it. Popp and Murphy [319], describe how their audio-driven VR experience *Planet Xerilia* manipulates oscillatory sounds of two large rotating objects and their associated distance attenuation curves to create a wave-like gesture of sound, which moves across the user from both the left and the right resulting in a reported sensation of tension and dissonance.

Sound can also convey information about the environment the user is located within and be used much in the same way as humans use sound to perceive and make judgements about their physical environment. Acoustic modelling can be

used to create an approximation of how sound behaves in the space, given its geometry and other environmental attributes such as room size, surface materials, and other objects [317]. In an experience where users have agency, the acoustics of the space may provide information that can be used to guide the user through the experience, something that does not have to factored in when designing sound for linear experiences. For example, the level and spectral content of a sound would change dependant on user location relative to the object within the environment. In the case of a sound-producing object being located in a different room to the user, both the direct and reflected sound would have no direct path to the listener and the spectral content would often be modified by the materials of the surfaces between the object and the listener. In this example, a reduction in higher frequencies compared to that of lower frequencies would usually result. Alternatively, if a sound-producing object was located behind a structure in the same area as the user, with the structure obstructing the user's line of sight, there may be a higher proportion of reflected energy reaching the user when compared to direct energy. Information on the distance from a sound-producing object to the user can also be constructed by manipulating the cues described in section 2.4. This information may then be used to aid the user in either finding or avoiding specific sound-producing objects. A study by Grohn and Lokki [320] found improvements in the ability of the user to locate objects within a environment when both auditory and visual cues are present, compared to either cue in isolation. Results showed participants utilised audio cues first to pinpoint the rough location of the target before proceeding to use visual cues to identify the exact location. It was also found that users were able to navigate through an environment by the sole use of auditory waypoints [321]. The use of sound for guiding the user is an implicit way of educating the user about the game or interaction mechanics of the experience [317].

Given that many IMEs provide 360° environments and the field of view of the user often covers between 90° - 120° at any given time [316], audio is also important for providing information about the unseen portion of the environment and can again be used by the user to build up knowledge of the environment and

assist them to navigate through it. In traditional linear media, the attention of the user is dictated by the camera position chosen by content creators; in non-linear interactive experiences, such as IMEs, however, sound is used to influence and inform the user's choices, without making choices on their behalf [317]. Spatial audio can play a vital role in guiding the user in a specific direction within an experience through the spatialisation of sound whilst also providing a greater sense of space, with the sound scene not being limited to the field of view of the user or the lateral plane.

Sound also serves to provide feedback to users in what Sinclair [316] refers to as the *Chime vs Buzzer* principle, which provides feedback to whether a user action was successful or not. This principle stems from the chime sound being interpreted as positive feedback and the buzzer sound interpreted as negative feedback. This is utilised prominently in many television gameshows, where a chime has historically indicated a correct answer, whilst a buzzer indicated an incorrect answer. In situations where users are required to interact with objects, a sound may be triggered when the user successfully makes contact or interacts with an object, i.e. the sound of a key entering a keyhole, with an alternate sound indicating non-completion.

It should also be noted that sound, or lack thereof, can also be used to distract or limit the amount of information available to the user, which can be used for narrative effect. As stated by Zdanowicz [317], it is not uncommon for horror experiences to place sounds typically associated with a threat in areas where no such threat is present in order to heighten the user's state of alertness. Likewise, an unseen danger may be placed in an area with a lack of indicating sound to catch the user off-guard.

3.4.2 Immerse

Some of the ways in which sound can contribute to a state of immersion is by influencing a user physiologically, psychologically, cognitively and/or behaviourally [322–324]. Psychologically, sounds can trigger a range of emotions, something which Roscar [325] refers to as *mood induction*. This is differentiated from the

communication of emotional meaning as it involves the changing of user emotions instead of simply conveying emotional information. This differentiation is based on the premise that one can receive information relating to happy/sad/anxious situations without necessarily feeling those emotions themselves [325]. While mood induction was originally defined within the context of music, it can also be applied to the use of non-musical sound if an emotion can be elicited from the user. This contributes to emotional investment in the experience, which Grau [326] refers to as *emotional involvement*. This in turn increases the likelihood of eliciting a state of immersion, as defined in section 3.2.

The ability to create a plausible sound scene which matches and enhances the visual content not only provides the wealth of information detailed in section 3.4.1 but also plays a significant role in immersing the user within the experience. The previously mentioned use of sound spatialisation and acoustic modelling can facilitate user development of a mental model of the environment by providing a detailed sound scene with a variety of sounds to accompany and complement different aspects of the environment. A scene based inside an inner-city flat may contain common home sounds; for instance the background noise of a television, a refrigerator hum or central heating. Contrastingly, the wider location of a flat located within an inner-city environment might be facilitated by the inclusion of constant external low-level traffic noise and the occasional police siren. Even if these objects are never present within the field of view of the user they help to build up a more detailed mental model of the environment the user is present within. Summers [327] argues that VR environments are created more in the audio space than in the visual space, since audio represents the entire space rather than the portion that exists within the field of view of the user.

Sound can also contribute to the consistency amongst various elements of the environment [316], ensuring congruency between the visual and auditory content. In the example of the inner-city flat the sound of the exterior environment would be expected to differ dependent on whether the windows were open or closed. Small details such as these may not be noticed by the user if executed correctly, but have the potential to break immersion if overlooked. This is largely due to the

behaviour of the environment not being consistent with user expectation, even if these expectations are subliminal. Though consistency is important in eliciting and maintaining a state of immersion, auditory repetition has the potential to break it [328]. For instance, repetitive events like footsteps or an object being repeatedly struck should result in a slight variation of the auditory event that is triggered by, and associated with, these sounds. As exact repetition is not a natural occurrence and the user may become hyper-aware of the artificial nature of the experience instead of focusing on the experience itself [329]. A study by Vachon [328] identified footsteps as the most problematic repetitive sound, as they occur frequently within a short interval.

3.5 Spatial Audio for Immersive Experiences

Over the last decade, spatial audio has become increasingly used within a variety of IMEs including VR, AR, video games, audio only experiences, installations, and 360° video [330–335], although the use of spatialised sound has been a topic of interest since the 1930s [336]. It is also important to note that the approaches to the design of spatial sound will often be dependent on the type of IME it is being designed for. For example, 360° video and VR share certain similarities where both offer a 360° field of view. This similarity presents some of the same challenges with respect to sound design, but ultimately very different experiences with respect to interactivity and agency are by offered by both. 360° video extends a linear screen-based experience to one where the screen surrounds the user, and as a result, the user can turn their head and experience 360° content, but often with very little agency within the experience. Contrastingly, VR allows the user to move freely within, and interact with, the environment, this enables different visual and auditory perspectives as well as often granting the user greater agency and usually a more involved role in the narrative.

3.5.1 Traditional vs Immersive Media

Spatial audio has long been used in broadcast and film; however, as most events take place on a single screen in front of the viewer with this type of media, sound mixing strategies typically use a frontally-biased approach to sound design. As such, traditional loudspeaker configurations like 5.1 and 7.1 have a higher concentration of loudspeaker channels towards the front to ensure accurate spatial imaging at the screen, whilst rear channels are typically reserved for sounds which provide general ambience as well as any supporting sound effects [337]. This introduces a conceptually defined "front" and "back" to the sound scene which typically results in any dialogue and character-focused sound effects being panned to the centre channel, with secondary sound effects - for instance, other objects within the visual field of view - being panned left and right, and ambient effects panned to the rear [337–339]. As noted by Lopez et. al. [340] the character-focused sound effects and dialogue are often placed in the centre channel to ensure good auditory localisation at the screen and rely on the ventriloquist effect [341] for audiences to perceive the location of the sound as being the position of the associated character. While advances in cinematic spatial audio, such as Dolby Atmos [342], allow for a much higher channel count, the nature of screen-based linear media means there is still a focus on frontally-biased sound design. However, systems such as Dolby Atmos aim, through the use of object based rendering and higher channel counts, to create a more complete sound scene around the listener by enabling more accurate placement of sound objects, particularly at the sides, above, and to the rear. In many IMEs the user is often able to view an environment that extends 360° around them, so unlike traditional single screen media, there is no clearly defined front and back since the user controls their orientation within the environment. It is worth noting that while many screen-based IMEs such as those using headsets, mobile devices, and some non-VR video games may use a 5.1, 7.1, etc. channel configuration, the audio is often adaptive to the position of the user. In a situation where the user rotates clockwise, the sound scene would rotate anti-clockwise to match the rotation of

the visual scene, using equations 2.65 from Chapter 2 in the case of Ambisonic audio. Sounds in front of the user as a result are always reproduced at the 'front' of the loudspeaker configuration, thereby taking advantage of the stable auditory image associated with the higher spatial density of loudspeaker channels. Subsequently, one of the key differences between the use of spatialised sound in traditional and immersive media is not the loudspeaker channel configuration but the facility for the audio to complement the 360° visual environment in a way that supports and is congruent to the user's agency within the environment; this allows the user to experience the environment from different perspectives without bias in any direction.

3.5.2 Use of Spatial Audio

In many cases the aim of using spatial audio is to deliver an auditory experience to the user that replicates or approximates a real-world auditory experience, which when combined with a visual environment results in a more plausible experience overall [343]. Although many IMEs aim to create a plausible environment capable of being explored and viewed from any number of different perspectives, there is usually still an overarching narrative or points of interest that the content creators have pre-determined the user to interact with. However, the agency afforded to the user does introduce a degree of unpredictability to the user experience, which may result in the user missing points of interest or elements of the narrative [331], unless adequately guided through the experience. Nielsen [344] identified three dimensions of cues with which to guide users in VR; these can also be applied to IMEs in general. The first dimension describes to what extent cues are explicit or implicit, the second describes the extent to which the freedom of the user is limited and the third describes whether cues are diegetic or non-diegetic. As noted by Bala [331] although these dimensions were not defined with audio elements in mind, there are existing examples of such cues being used in the context of audio cue design. Gödde et. al. [345] suggests that spatial audio is effective for guiding user attention, while mono sound sources can cause the user to search for the sound source. Additionally, they propose

sound as being especially effective when used in conjunction with visual cues, a notion that is generally supported by the literature [320, 346, 347]. A study by Bala [331] showed mixed results for using spatial audio (diegetic effects and music) to orientate users within 360° videos. They observed that spatialised diegetic cues associated with specific objects, in this case an elephant trumpeting, did not introduce significant benefits in directing attention as users would look around to find the origin of the sound irrespective of whether the sound itself was spatialised. In some cases users would also incorrectly identify a diegetic sound as being spatialised when associated with an object within the scene, a result which can be explained by the previously mentioned ventriloquist effect [341].

In some instances, a combination of spatial and non-spatial sound, such as head-locked stereo, which will remain in a fixed position relative to the user, is used to create environmental ambience, as not all sound needs to be spatialised [327, 333]. When designing the environmental soundscape for Barking Irons, a location-based VR experience designed in Unity with audio being modelled and rendered using Wwise and Two Big Ears (a company that subsequently was bought my Facebook/Meta [348]), Summer [327] explains that certain individual environmental sounds, such as animals, vehicles (in this case a train), and tumbleweed would be spatialised using Two Big Ears, whereas other sounds, such as wind, would be implemented using a looped head-locked stereo track. Although an explicit reason for this decision is not presented beyond personal choice, it may be as a result of certain sounds, such as wind and rain, not necessarily having a defined position in a real listening situation as these types of environmental sounds often envelope a listener. Therefore, having those sounds as stereo mixed in with other environment sounds which are spatialised as point-like sources creates a plausible soundscape. Non-diegetic music is also usually rendered as non-spatialised stereo to provide a sense of separation from the diegetic sounds [333]. Additionally, some spatialised sounds may also be head-locked, to ensure the user does not miss key auditory points of interest [349].

3.6 Summary

This chapter has provided an introduction to relevant areas relating to sound design as applied to IMEs. This includes defining the term *immersion* within the context of this thesis, and therefore what constitutes an immersive experience and by extension an IME. Different types of IMEs were defined and discussed including the differences between common categories such as AR, VR, MR, and 360° media and the role that sound can play within IMEs and how it can be used to not only assist in immersing the user, but also for driving the narrative and guiding the user through the experience. The chapter was concluded by exploring spatial audio within the context of IMEs including how spatial audio differs in its use between traditional and immersive media and how it can be utilised to support the narrative and provide a congruent experience to the user given the addition of agency often afforded within IMEs. The next chapter presents the first piece of original research in this thesis, a qualitative investigation into the defining features and associated challenges of spatial audio production for IMEs from the perspective of sound design practitioners working in the field.

Chapter 4

Immersive Sound Design Practice

4.1 Introduction

In order to aid the development of useful technology and methods for immersive audio production it is important to first ascertain areas of the production process that would most benefit from such interventions. This will assist in reducing the scope of the research and identify areas where the research may have the most impact. To achieve this, a thorough understanding of the problem space, alongside the processes and challenges involved in creating audio content for immersive media experiences is required.

Designing sound to be interactive and/or immersive is not new, and is well established within video game sound design practice [248, 350] with many approaches being potentially suitable for adoption within non-video game IMEs. Sound design for traditional linear media also has well documented practices and workflows [8, 351]. The process of immersive sound design for IMEs utilising spatial audio can, however, still be considered a relatively new area of practice with less well defined methods, and requiring a new, and still emerging, set of skills and tools. Hence, at present, little has been formally documented in the literature with respect to the new challenges introduced by this new type of

content create, the tools used to create it, and what those creating the experiences see as the defining features that differentiate it from traditional media. Due to this, the question is raised of whether the technology and tools being developed align with the needs of content creators.

This chapter presents a qualitative study that was undertaken to gain an understanding of how practitioners working in the field have responded to this new form of content with respect to their working practices and what they perceive as being the challenges in producing immersive audio content. It further looks to explore how current machine learning technologies could be used in an assistive capacity to address some of these challenges.

This study was approved by the University of York's Physical Sciences Ethics Committee (ref: Turner190919) and is available in Appendix A.

4.2 Background

4.2.1 Recent Related Literature

Within the context of non-interactive linear media, Baume [352], presented a study that aimed to both investigate how radio programmes were created and identify opportunities where technological intervention may improve existing processes and workflows. Data was gathered through both ethnographic observations and interviews that were conducted between observation periods. Three varied case studies were examined with observations ranging from half a day to four days. The study found that production teams often relied strongly on scripts and transcripts of audio content and often preferred working with paper. Improvements to existing workflows and tools were suggested that link audio to text, highlight repetitive audio segments, compare takes, and speaker classification and segmentation i.e identify where in the text different people are speaking. Later work by Baume [353], addressed some of the previous studies findings through the development of PaperClip, a digital pen interface for editing speech recordings directly on paper.

Ward et al. [354], investigated the perceptions of production teams with respect to personalisable object based media and its integration into their current

workflows. The results were then used to develop a framework for creating production tools and workflows for new media experiences, which was then evaluated through comparison with case studies which covered three end-to-end productions. The authors concluded that iterative development processes and close collaboration with production teams leads to the development of tools that offer both the key functionality and simplistic and intuitive interfaces required by practitioners. Furthermore, the collaborative process allows the development of tools that suppose storytelling and production. Conclusions drawn from the survey results suggested that, most importantly, new media formats, such as personalisable media or other IMEs, must ideally benefit both the creative process of the producer and the end-user experience.

To further understand how personalised experiences were being created, Cieciura, Glancy, and Jackson [355], present, what they define as, an ethnographic study of six case studies produced by BBC R&D and examined through interviews with the content producers. Analysis of the data was conducted reference to pre-defined areas of interest including, collection and use of metadata, how metadata models were created to facilitate personalisation, and the use of production tools. Results from the interview suggested that production of personalised media, when compared to linear media, involved many additional tasks which resulted in a substantial increase in workload and additional responsibilities. Development of metadata models were highlights as being particularly time consuming as experience required the development of a bespoke model. Additional challenges were encountered during the post-production phase which were attributed to a lack of consistent vocabulary, lack of specialised tools, or tools which did not easily integrate into existing workflows. They suggest the development of cross-compatible metadata models and an overarching technical framework which would remove the need for practitioners to develop their own bespoke systems, unless that is what is desired. Additionally, they suggest the use of AI to automate procedural tasks such as metadata creation.

As part of a wider body of work that explored a number of topics centred around collaborative 360° video production for social change, Baía Reis [356],

conducted 21 epistolary (asynchronous) interviews between May 11th and August 23rd 2020. The results from the interview data analysis were not presented in isolation but as part of project's wider conclusions, which highlighted several challenges that may also be present in other types of IMEs. These include placement of microphones when filming 360° video, camera position i.e placing it at either standing or sitting height, directing the participants attention within the space, which was discussed in Section 3.4.1, and It was also noted that only two out of 21 interviewees made reference to audio or sound when discussing immersive media, one of which was a professional sound engineer and the other an immersive filmmaker. It is worth nothing that full biographical profiles were made available for the interviewees, this has the potential to bias results as interviewees may be more guarded about their responses given that they are able to be attributed back to them. The main output of the work is a framework for collaborative 360 video production.

A technical review of virtual museum, heritage, and tourism experiences is presented in [357], with a particular focus on the challenges associated with realistic asset creation. The authors highlight some of the key challenges by presenting two contrasting examples of object digitisation, one for a very large object, which cannot be moved, and a small but more detailed object and then offer a review of current methods and challenges associated with very large asset creation i.e the digitisation of buildings or historic sites, including the use of aerial 3D mapping.

Candusso [358], through the use of an online survey, explored audience perceptions and awareness of current and emerging cinema technologies, with a focus on 3D imagery and spatial sound. The aim was to provide an insight into whether audiences make decisions based on a particular cinema or a particular technology. It also served to try and ascertain the literacy of an audience with respect to film sound formats and 3D technologies. 201 participants responded to the survey, although the author acknowledges that as 80% of respondents were from Australia and, at that time, Australia only had two immersive sound installations, this did limit the number of participants who had experienced

the formats of interest. As such, the results may not be generalisable to an international audience. Results showed that the majority of respondents (71.4%) would select a cinema based on convenience rather than the technical capabilities, although over half of respondents (54.1%) said they would prioritise sound quality over visual quality. Overall conclusions were, that for a largely Australian audience, good quality sound is not a particularly high priority for cinema audiences.

In [359], Candusso takes a practice-based approach to investigating alternative approaches to traditional cinematic mixing practices and methodologies. By creating alternative mixes in both traditional cinematic formats, such as 5.1, and spatial sound formats, such as binaural and FOA, Candusso demonstrated that techniques such as binaural afford a more homogeneous sound scene through its facilitation of distance when compared to traditional methods where the sound has a fixed minimum distance matching that of the reproduction array's position. It was however noted that accurate spatial position of sound in 3D space is time consuming, especially dynamic sound that requires trajectory data.

4.2.2 Relevant Data Collection Methods

Qualitative research methods refer to non-statistical and non-numerical methods of data collection and analysis [360] and are often used when looking to collect data that can be considered non-quantifiable, such as the history and experiences of the people, societies, and cultures. This work in this chapter aims to explore the experiences of those immersive sound design practitioners producing content for IMEs. Stepney [361], presents six different methods of qualitative data collection: surveys, interviews, observations, focus groups, document research, and archival research.

Observations involve the data collector directly observing participants and allows them to collect data that relates to what the participants actually do, as opposed to what they may report they do. Within the context of workplace studies, Luff et al [362] note that, due to some of the performed activities becoming procedural or 'second nature', there can often be a difference between

what participants believe they do and what they actually do. They also criticise interviews and questionnaire methods on the basis that the researcher may not know in advance the right questions to ask to elicit the desired information [362]. When undertaking observations, the observer can either participant in the activity or scene being observed (participant observation) or be present purely as an observer (non-participant) [363]. One of the clear challenges with observational studies is ensuring adequate access and that the presence of the observer is not perceived as intrusive. Additionally, participants may, knowingly or unknowingly, deviate from their natural behaviour due to the knowledge that they are being observed, a phenomena known as the Hawthorne effect [364]. It should also be noted that observations can be used in both single and group participant settings and, as noted by Ranny et al. [365], may occur both in-person and remotely. Data collected from observations may consist of field notes, photographs, audio and video recordings.

Focus groups are similar to observations, in that they involve the observation of group interaction, but occur in a more controlled setting where, as highlighted by Stepney [361], they are effectively researcher-led conversations amongst a small group of participants. Although the data collected encompasses both individual and group responses, group dynamics may result in a biased dataset due to relative suppression of minority viewpoints or potential hesitancy in discussing certainly topics openly [366].

Surveys focus on an individual participant within the wider group being studied and involve the distribution of a set of structured questions to participants [361]. Braun and Clarke [367], note that one of the benefits of surveys is that they allow the respondent more time to formulate answers and space to express themselves using their own language and ideas, and thus may result in the generation of detailed written evidence relating to the research questions. As surveys are often self-administered without the researcher present they often lack the opportunity for additional input from the researcher and excludes the potential for follow up or probing questions. This may cause interesting and relevant themes to be missed if not identified during the survey design stage.

Interviews also focus on an individual participant but afford greater flexibility with respect to data exploration, when compared to surveys, due to the interaction between interviewer and interviewee. A semi-structured approach is commonly employed, with the researcher utilising a pre-defined interview guide that outlines the main questions to be asked or themes to be covered. Using a guide provides the researcher the flexibility to ask follow-up questions and probe the participant for more detail where appropriate [365]. The conversational nature of semi-structured interviews [365], also allows the researcher to explore interesting themes that arise during the discussion that may not have been pre-defined on the interview guide, but the guide can then serve as a sign post to allow the researcher to maintain the flow of the interview once the discussion around the unplanned topic has concluded. Although interviews are usually done with both interviewer and interviewee present, epistolary methods, such as those used in [356], do not require the interviewer and respondent to be co-present in time. This may be beneficial with respect to the difficulty that can sometimes occur when arranging suitable times for observations or interviews. Additionally, it allows the respondent time to consider the questions and formulate responses, similar to that afforded by surveys, but still provides opportunity for communication and rapport building between the respondent and the interviewer [368]. When responses are given via written text it also removes the need for transcription [368].

Ranney [365], offers guidance when formulating questions to be asked during interview or focus group based studies as to optimise the quality of the data collected. Opened-ended questions are recommended as they encourage the participant to provide a more detailed answer, whereas closed-ended question, ones that can be answered by yes/no, does not encourage the participant to provide detail and may then requiring the researcher to ask follow up questions. Leading questions, which are questions phrased in a way that implies a given answer may result in false or inaccurate statements and therefore bias the results. Table 4.1 provides an example of a single interview question phrased as open-ended, closed-ended, and leading.

Question type	Question
Closed-ended	Can ML tools help IME production?
Opened-ended	Tell me what you think the potential use of ML within IME production?
Leading questions	Tell me how ML tools would help IME production?

Table 4.1: A example question phrased as an opened-ended, closed-ended, and leading question

4.3 Methods

4.3.1 Research Questions

This chapter addresses the following question:

- What are the challenges associated with IME production and how, if at all, could ML be used to address any of the identified challenges that relate to the production of spatial audio for IMEs.

4.3.2 Data Collection

Originally, it had been planned for the study to be a mixture of observational case studies, similar to that used by Baume [369], and face-to-face and/video call interviews dependent on the participant's location and availability. However, due to the COVID-19 pandemic and the restrictions put in place, all data collection was undertaken through interviews and an online survey. The interviews were all conducted by the author via either Zoom, a video conferencing application, or telephone ,with participants taking part either from their home or workspace. This presented some challenges in the form of audio quality and issues arising from internet reliability.

Data were collected between the months of February to September 2020. The online survey contained a combination of multiple choice and open-ended questions. Closed-ended questions were avoided, outside of collecting basic demographic information, in order to avoid yes/no answers and encourage the participants to include as much detail as they saw fit. The interviews followed a semi-structured format using the survey questions as a guide, this ensured

each interview followed a similar structure and that the same general topics were covered with each participant [370]. Additionally, the use of a guide also afforded the flexibility to depart from the guide as needed to follow interesting lines of discussion and, as mentioned in Section 4.2.2, acts as method of signposting for the interviewer to ensure the interview can return to the pre-defined topics when appropriate. During the design of the survey/interview guide, care was taken to ensure only a single question was being asked at any given time, as the use of double barrel questions (two questions contained within a single statement) can potentially confuse participants and will often result in only one of the two questions being answered [371]. Interviews ranged in length between 25 minutes to 1 hour 19 minutes and were captured either using Zoom's built-in recording functionality, or via a Zoom H4n audio recorder on the side of the author. At the start of each interview prior to beginning recording the interviewer reexplained the structure and format of the session, roughly how long it could be expected to last, and reconfirmed that the participants consented to audio and video being recorded. Participants were also asked if they had any questions prior to the interview starting and after it concluded. Throughout the interview the interviewer often used clarificatory questions or requests for additional details as a way of prompting a more in-depth discussion around the topic being addressed [355]. As the interviews were being recorded, limited notes were taken and instead the author noted interesting points that were felt warranted further exploration.

The dual data collection method was used to increase the likelihood of responses, as research has shown questionnaires often suffer from low response rates [372]. This was of particular relevance given the selective approach in choosing participants as discussed in Section 4.3.3.

All interview recordings were transcribed using the NVIVO qualitative analysis software [373], while the online survey was created and administered through Qualtrics [374]. Transcripts were reviewed by the author for accuracy and revised where necessary. Written consent to create and use recordings, analyse pseudonymised transcripts, and publish the subsequent results was obtained from each participant.

The interviewed guide is contained within Appendix B. The raw interview data is stored according to the terms of the ethical approval and in line with the University's research data policy.

4.3.3 Participants

As it was desired to interview those with professional experience in producing IME who would have insight into current industry practices, potential participants were identified through nonprobabilistic, purposive sampling techniques. Participants were required to be working professionally within the industry and have experience working on productions requiring immersive audio. Participants were recruited via targeted emails utilising contacts from within the BBC R&D's audio team and the University of York's AudioLab. The selection criteria for candidate participants was broad in nature as to allow for the inclusion of a broad range of experiences within the IME sector. Candidate participants were required to be active in either IME production that utilised immersive audio or R&D with a focus on immersive audio. This could include sectors such as radio, film & TV, broadcast, music/sound recording, experience design, higher education, video games, and technology/product R&D. Candidate participants also either had to have expertise in audio production/audio engineering or have experience in managing productions that heavily utilised immersive audio. Those undertaking taught programmes of study (students) were not eligible unless they were also working professionally, however those undertaking post-graduate research would be considered depending on their portfolio of work and area of research.

26 people were approached and of those 26, 11 were either employed by audio productions or freelancers specialising in immersive audio production, 9 were employed by national broadcasters in roles relating so radio, sound recording, technology R&D, content R&D, 3 were employed in games industry, and 2 were managers for international technology product companies, and 1 for a private research organisation. All those approached were given information that briefly outlined the purpose of the research study and included the name and contact information of the author. Participants were also given the option of either

completing the online questionnaire or taking part in a semi-structured interview.

Of the 26 approached, 5 of these were interviewed and 2 completed an online survey. Although it is acknowledged that the number of participants is low, the experience in industry for those interviewed ranged from 2.5 - 27 years with the median being 10 years. Of those who accessed the survey, 2 had 5+ years experience, and 1 had 3-5 years experience. Participants included award winning sound recordists/audio producers specialising in immersive audio, an audio director for a AAA video game development studio, experienced R&D engineers, and a senior games audio programmer. This resulted in a small, but highly experienced, pool of participants from which to collect data. Participants came from a variety of sectors within the industry which included video games, broadcast, streaming media, and installations, with some participants operating within multiple sectors.

4.3.4 Thematic analysis

Inductive thematic analysis was performed based on the methodologies and procedures presented in [375] and [376]. An inductive approach, also referred to as a *bottom up* approach, codes the data without trying to fit a pre-existing framework or a pre-determined set of codes and/or themes. This is seen as a data-driven approach to analysis. A code book was created in NVIVO and Microsoft Excel where all identified codes could be added and any new codes could be cross-checked against the rest of the data. The Analysis followed five stages: initial reading of transcripts; identify text segments related to research objectives; identification and definition of themes/subthemes; reduce overlap and redundancy amongst themes/subthemes; and interpretation of themes in relation to original research questions.

4.4 Themes

Three broad themes, each containing a number of subthemes (shown in Table 4.2), were generated through the analysis of the coded data: The Virtual Environment,

Production Practicalities, and End User Experience. These themes reflect the current literature concerned with what constitutes a state of immersion and the psychological and physiological factors that can elicit a state of immersion [247, 250, 251, 377] but are drawn from the perspective of those creating the content, and sit within the context of professional practise.

Themes	Subthemes
The XR Environment	Localisation Capturing/simulating reality Multi-sensory Timbre Spatial aspects of experience
Production Practicalities	Availability of resources Automatic processing Sound quality Tool functionality Working with non-experts
End User Experience	Interactivity Cognition Levels of immersion Novelty

Table 4.2: Themes and subthemes generated from inductive thematic analysis of interview and survey data.

4.4.1 The XR Environment

An area highlighted across all interviews as being an intrinsic, yet challenging, feature of IMEs, was the creation of the XR environment and the ability to have it replicate the sensory signals the user would experience were they physically in that environment.

There was a consensus that one part of simulating a real environment required

the auditory scene dynamically reacting to a change in the user's orientation, as would happen in the real world:

> which again, is all down to trying to model reality and create something, whereas, [...] if I'm turning around, instead of the sound field staying still, [...] the sound field moves with my head and it doesn't lock itself to your head. [Participant 1]

Participant 2 noted that what is considered desirable in IMEs might be at odds with what is generally desired within traditional media content,e.g. room ambience captured as part of a recording.

> In 360 you might actually really want something to sound like it's off mic because then it's capturing more of the room that the sound source is in and actually closer to what the real thing sounds like. [Participant 2]

The ability to emulate distance between the user and an object was associated with creating a sense of presence for the user by enabling the externalisation of the content. This was seen as particularly relevant when using headphone rendering, as this method lacks the natural distance between the user and source inherent in loudspeaker systems.

> ...distance to me is a big thing, externalisation, and this seems to me like they are [...] the two main things for me to create the sense of presence in space....[Participant 4]

> externalisation is a big one, [...] if you're delivering via headphones [...]. 'cause having access to a lot of speakers in an array is much more tricky, so assuming it's headphone delivered, having things sound like they're outside of you and not inside your head, is the hallmark of good immersive audio, because in the real world, that's what sound sounds like. [Participant 2]

Participant 4 noted that when simulating distance within a synthesised environment, the processing required would be dependent on how the object had been recorded and edited, particularly with reference to its loudness. Standards for distance processing are noted as being difficult to establish.

> If an object is one metre away in a virtual world, it doesn't mean it actually sounds one metre away because it depends [...] how loud it is going into the spatialiser and how you edited it. [Participant 4]

For some, this has led to a perceptually driven approach to distance emulation using plausible approximations that make subjective sense to the the content creators, even if the parameter settings used are not objectively accurate.

> [...] you basically use your own approximations, arbitrary figures that initially make sense, but then you tweak it to trick your brain. OK, that sounds believable. OK, that works for me. Even if [...] the figures [on the screen are] not correct. [Participant 3]

It was noted by some that the technology presently available for emulating distance is simply not yet of the desired standard.

> I think this is where everything is falling short, where we actually can say, oh, this guy is three meters away and I feel it. So distance modelling is quite a hard thing to do. [Participant 5]

Although importance is placed upon both being able to place objects accurately within a scene, and faithfully recreating the tonal characteristics of an environment, participants felt compromises must be made with respect to these features depending on the aim of the audio at that particular instance. This was because the tools being used to enable finer control of object placement often did so at the expense of introducing greater tonal coloration.

> two things that I'm always looking for, precision or timbre. [...] when I need localisation, maybe I give up a little bit on the, on the sound

> quality of it, I know that timbre might not be there. But when I want everything to sound really nice and smooth maybe I give up of a bit of the localisation. [Participant 4]

Non-spatial aspects of realism were also noted as being important within interactive media, with Participant 1 noting the perceived realism of characters. This may involve greater efforts to have their behaviour, such as in game dialogue, less repetitive by giving them a wider range of possible responses. This can result in many lines of dialogue across all in-game characters.

> something [...] that we bump up against and it (sic) indirectly to do with immersion in as much, I suppose it's more to do with kind of being believable and not repetitive, is editing and organising dialogue lines as we have more characters. They say more things to try and give the illusion that these are real characters. [Participant 1]

Though all the participants were audio professionals, with none undertaking professional visual production work, all expressed the importance of multi-sensory stimuli as a key feature in immersive media.

> It's [...] including all their attention in many senses as technically possible. [Participant 1]

It also considered that a combination of visual and auditory signal processing can together provide a greater sense of depth and distance to an environment. The importance of visual quality should also not be understated as it is a key aspect of many immersive experiences.

> [...] the video is stereo so you've got [...] a sense of depth of vision and having that additional stuff audio wise enhances what you see. [Participant 2]

Responses from all participants focused on the goal of creating an approximation of reality when creating IMEs. This raises the question of what it means or

what are the requirements for an audio object to sound *real*. Although this, and related, terms were used by participants throughout the interviews, no definition was established as to what they considered constituted a real world experience. It could be argued that everything is a real world experience as, even within an IME, our ears and eyes are responding to physical stimuli. One interpretation could be that content creators do not wish it to be apparent that the scene/object is being generated by some form of loudspeaker, rather than the physical object which it aims to represent. This would then account for the desire to simulate the *real world* timbre/frequency response of a sound and avoid any artefacts that could cause the user to focus on the device producing the sound rather than the sound itself. Absolute accuracy, however, appears to not be required as it was acknowledged that often a compromise is needed between accurate auditory localisation and the tonal accuracy of the associated environment.

4.4.2 Production Practicalities

There were many frustrations and challenges associated with the production of IMEs spanning all aspects of the production process. This usually centred around the view that current processes were lacking, hindering content creators in delivering experiences as easily as they might if IMEs were more commonplace and the tools and processes more developed.

Some participants noted the lack of available material in spatial formats (such as Ambisonic B-format), which meant they often had to record their own material.

> There's not a lot of Ambisonic source material around. A lot of the stuff we've used, we've recorded ourselves. [Participant 1]

When unable to access spatial material for the specific environment they were looking to create, some resorted to layering stereo ambiences of the target location with a spatial ambience of a similar environment to help give the scene cohesion.

> If I've got the ambience in stereo from a London street, then I can put it in the background, some random street ambience [in Ambisonic format] just to fill up the space. [Participant 4]

While in some cases they resorted to just using spatialised stereo material.

> A lot of the time it's actually constructing stuff out of stereo and then spatialising it ourselves. [Participant 1]

Some participants commented that common methods for spatialising objects and rendering spatial soundscenes are still quite difficult to work with and can not always deliver the desired results, specifically when rendering over headphones.

> In fact, for probably all of them [VR users], it's going to be on headphones. So I feel like VR brought kind of binaural into focus and trying to get binaural sounding good, which is I think the big challenge. [Participant 1]

Working with non-experts also poses challenges. Clients commissioning IME content often lack the language to clearly articulate their feedback and may not have the skills to pinpoint what is causing any perceived issues.

> Clients are a challenge. They are able to say, I don't like this. I don't know what's happening, but, if it's wrong, it's wrong. [Participant 5]

There can also be conflicting assumptions in regards to the aesthetic goals in IME production when collaborating with production teams accustomed to creating traditional content. Some concepts of sound quality may differ between collaborators, for example, the desired ratio of direct and reverberant sound on a dialogue track.

> a lot of people talk about things sounding off mic, as sort of bad sounding TV mixes. In 360 you might actually really want something to sound like it's off mic because then it's capturing more of the room that the sound source is in and actually closer to what the real thing sounds like. [Participant 2]

When dealing with immersive content that has both 360° video and spatial audio, placement of the microphone in relation to the camera is also of importance when maintaining the correct perspective between the visual and auditory material.

> I have been given audio recorded fairly close to a camera, but just in the wrong place, and it all sounds completely wrong. [Participant 2]

With these immersive experiences still not yet being widespread within the industry, and game audio workflows already having established platforms and tools, there can be hesitancy in adopting new technologies that require new practices and tools

> not everybody's completely sold on Ambisonics. So people are still quite attached to a world that they feel they've got control over. [Participant 1]

Reliability was also felt to be a contributing factor in the adoption of new technology proposed to assist with immersive workflows. As production timelines are often strict, tools need to work first time and complete the task quicker than the content creator would be able to do manually.

> if something is not reliable. You can't use it because the, the time lines of production are so tight. [Participant 5]

4.4.3 End User Experience

Thoughts on end user experience seemed to be predominately two fold: Firstly, aspects of the user experience directly delivered by the content, such as the interactivity afforded to the user within the environment; and secondly, the psychological aspects that occur within the user's own cognitive processes, usually as a result of the technological processes drawn out in previous themes. Although these two groupings have the commonality of being generally technology driven, they are not dependent on one another, since as noted by Participant 5 , immersion

is not exclusive to content delivered via a specific medium and can in fact be achieved without any technological intervention.

> what we're meaning is that we get people losing themselves inside the experience and that can happen in any kind of medium, of books, especially [...] which are [...] non technological. [Participant 5]

In terms of what separates traditional content from immersive content, it is often considered that the user should have a level of participation within the environment and/or narrative, as opposed to merely being an outside spectator.

> the main differences for immersive experience: you're creating a world for the players to participate in. [Participant 1]

Allowing the user to participate in the narrative is often associated with allowing them to make choices that affect the direction or flow of the story, and this in turn gives them a sense of agency within the experience and causes the user to become invested in the story they are now helping to shape:

> it's basically anything, [...]which enhances the player's investment in the experience and their sense of agency in the experience. [Participant 1]

Alongside participation, aspects of the narrative, such as its ability to compel and engage, were seen to play an important part in a user's potential to become immersed in an experience, and was cited as something that should be considered carefully during the production process.

> a key feature, (is the) story, telling a convincing story. [Participant 5]

When experiencing traditional media content, the user may not have a definitive position within the action. Camera angles change, and the sound scene is not always constructed to be a realistic representation of each object's location in relation to the camera location or viewing perspective. This is especially the case in audio reproduction formats that are horizontal only.

> A definite listening position is another big difference, with immersive audio where in traditional stereo or, I guess even in surround really, there isn't a definite, you're not in a very fixed position as the listener [...] as a viewer, you can see things and you might hear footsteps just there for effect, but they don't have to be rendered in such a way that is true to life. [Participant 2]

Not all immersive experiences require a first person perspective, both first and third person perspective are common in video games with some allowing the user to dynamically change between the two. Some participants felt this created differing levels of immersion, depending on the perspective from which the user was experiencing the world. The user still maintains a sense of agency and active participation in the narrative, but with a third person perspective they can be said to be taking control of a character within the world, rather than being the character, as would be the case in a first person experience. This was viewed more as an interactive cinematic experience.

> I think it's [the video game] a kind of more cinematic experience. I think for we're creating a real world. But I think we were creating a real world in terms of a kind of movie that you can interact with. [...] because you can see the character on screen. So obviously you are not the character. So it doesn't have that level of immersion. But you can control the character. [Participant 1]

An experience being believable, as opposed to *real*, as noted in Section 4.4.1, was also seen as an important part of being able to elicit a state of immersion from users. This is interesting because believable does not always have to correlate with creating something that is exactly true to life. There are certain situations where aspects of the experience need to be overstated to have the desired impact and compensate for the fact that the experience is not a complete sensory one.

> They could be a little bit hyper-real. In as much as sometimes you might want to slightly amp up the experience [...] the goal is still for

> people to believe that, you know, that gun that you're picking up and manipulating is a real gun, and it feels like a real gun. Sound does as much as it can to make that thing feel like a real gun. [Participant 1]

In the real world, each naturally occurring sound is often unique, with even repetitive sound events, such as gunfire, differing in small almost imperceptible ways. The lack of these minute differences, as highlighted by Participant 1, is something that a user could become sensitive to, resulting in the immersion becoming broken.

> one of the things which I think breaks the immersion for games...it's repetition, [...] in a game where you're trying to simulate reality, any kind of repetition people are very, very sensitive to. [Participant 1]

Participant 5 noted that another important factor, in addition to the techniques employed by the content creator, is the user's perception of the uniqueness of the experience. This is something arguably outside the control of content creators.

> [the experience] needs to have a certain standard in order to convince people that they are experiencing something unique and special. [Participant 5]

Participant 5 also commented on the user's preparedness for undertaking an IME. The process and effect of taking the time to prepare oneself for an immersive experience could be just as important as the techniques employed by the content creator to elicit the state of immersion.

> If you go to the cinema, you're not just going to the cinema. You're not just sitting in the cinema and watching the film. [...] Making the decision to go to the cinema, travelling for something that is important to you and then going inside and buy a drink and some popcorn and getting in the mood for this whole thing and to be prepared [...] we're going to take time for this and we're going to

> turn off our phones and everything. We're going to be fully there and there's nothing else that is distracting us... [...]how do we get to the experience in order to be prepared to let ourselves go. [Participant 5]

The theme of End User experience encompasses both the cognitive aspect of immersion and how the attributes of the user experience differ from that of traditional media. Participant responses under this theme, relating to the defining features of IMEs, often focused on aspects of the experience that could be associated with the concept of *involvement*. In the literature involvement is often framed as a psychological state necessary for cognitive immersion [246], but within the interviews the term was arguably used as a synonym for participation and/or agency. This was highlighted by participants making a point of describing how the users should be able to interact with the experience and participate or have agency within the narrative, this is particularly evident in video games that are produced for VR as the user often embodies a main character central to the narrative. This participation can result in the user entering a state of involvement as described in the literature. Even within the IMEs where users are more passive a state of affective involvement can occur which represents the emotions resulting from the design and aesthetics of the experience itself [378].

The amount of agency a user would have varies greatly between experiences, as do the differing perspectives the user could take of any unfolding narrative. Participant examples demonstrating these varying combinations of user perspective and user participation included the user having a first person view of a musical concert but being passive as an audience member; having a third person view but being in control of a character; being able to interact and make decisions within the narrative, as is the case with many video games. Which of these examples is more immersive will be dependent on the individual and their situation, and goes beyond just the nature of the experience. The idea of user perspective could also be interpreted as being related to the importance some place on users being given a defined position within an experience, that would be true to life were they physically present in the environment.

4.5 Discussion

The XR Environment, Production Practicalities, and User Experience themes emerge from the perspectives of those professional practitioners who are creating the content, and also happen to closely reflect the current literature concerned with what constitutes a state of immersion and the psychological and physiological factors that can elicit a state of immersion [246, 247, 250, 251]. Though the participants were all audio practitioners, it is interesting that much of the interview data presents a holistic view of IME production, and while spatial audio production plays a key role in defining this new form of content, it is inextricably interconnected with other aspects of the experience such as user interaction, quality of narrative, and visual content.

4.5.1 Distance Perception

There are various well established methods for placing audio objects around the listener, however, placing sounds at a distance from the listener is a commonly expressed area of difficulty, and is therefore related to the second research question. When using headphone based audio systems a prerequisite to creating auditory distance is the ability for the system to externalise the sound so it is perceived as being located outside of the listener's head, and this was seen as a defining aspect of immersive content. If this prerequisite of externalisation is not achieved then it is very difficult to create a sense of auditory distance comparable to a real world experience. Head movement tracking, another technology highlighted as being key to producing immersive content, has been shown to play a significant role in providing externalisation due to facilitating the simulation of dynamic spectral cue changes and can be effective even in the presence of degraded binaural information [100]. A later study by Kearney [68] also concluded dynamic binaural rendering assisted in distance estimation in VR, but only due to the reduction of front-back reversals. Even with externalisation achieved it was still seen as a challenge to simulate objects at specified distances, and often participants relied more on their own subjective approximation of distance and less on whether the

parameter values applied using auditory software reflected accurate values. A possible reason is that our understanding of the mechanisms involved in auditory distance perception are lacking when compared to azimuthal localisation, and the reliability of distance cues can vary with stimuli, environment, and source distance [40]. This introduces an added complexity for content creators to deal with. Some of these cues, such as direct-to-reverberant energy ratio and the overall level of a source, are signal attributes commonly manipulated via software in order to imply an approximation of distance. Given the extra psychophysical complexity involved in auditory distance perception, and the degree to which these estimates vary in accuracy depending on the individual, stimuli, and the environment, it is maybe not a surprise that as yet, a standardised way to effectively simulate distance has not been found.

4.5.2 Multi-sensory aspects

The multi-sensory aspects of the experience were also deemed vital in order to achieve immersion. Alongside the quality and accuracy of the audio reproduction it was also felt that visual quality was an important factor, with techniques such as stereoscopic video helping to reinforce a sense of distance when combined with audio signal processing. The inclusion of multiple sensory stimuli better replicates what would be experienced in the real world, assuming no sensory impairments, further supporting the idea of perceptual/sensory immersion. It can also, given the well documented ability of our visual system to influence auditory perception, assist in achieving a greater quality of experience than current audio technology alone can deliver. This raises the possibility that it may be harder to achieve the same level of spatial plausibility with audio only content.

An important point to consider briefly is that, by their very nature, video games are designed to be immersive. Interactivity is a base requirement for all video games, but the ways in which video games have progressed in recent years, including the rise in popularity of spatial audio and the increased computational power of technology platforms that host them, means they are now often aiming to offer a multifaceted experience of immersion. Many non-video game IMEs

model the interactivity found in video games through involvement in the narrative, or affording the participant some degree of agency within the environment.

4.5.3 Immersion factors

The ability for a user to become fully immersed within an IME was intrinsically linked by participants to the quality of all aspects of the production, both technological (e.g ability to replicate accurate sensory information) and non-technological (e.g. quality of narrative). It was also said that the experience required a certain standard to convince the user they are experiencing something *unique and special*, although the exact implied meaning of this statement is not clear. The idea of being required to present something that the user finds unique and special could suggest that the user's perception of novelty, and their prior experience with the medium, may have an impact on the level of immersion they experience. For users inexperienced in IME environments there may be a greater inclination to suspend disbelief and engage with the experience [379], and this may cause them to be more likely to ignore/not notice quality issues that may be apparent to those more experienced. If this is the case, it raises the question of how long this "novelty effect" might last for, and once users become more accustomed to the experiences will it become increasingly difficult to elicit the same perceived quality of immersion?

4.5.4 Tools and assets

A lack of available or adequate resources and tools were seen as barriers to the adoption of immersive audio within the wider industry. Though multi-channel microphones are becoming more readily available, making in-house production easier, there is still a lack of sound effects libraries containing spatial 360° audio content when compared to mono and stereo content.

The adoption of new technology can often be a challenge as it requires experimentation and adaption in order to be refined, but often due to the tight production schedules and the inherent risk involved it can be difficult to undertake that experimentation outside of a research and development context.

All participants referred to a commonplace requirement to capture bespoke spatial audio recordings as part of their work. This substantially increases the time taken to complete tasks, such as creating atmospheric audio beds for a scene, due to either the need to record a specific soundscape, or create an artificial soundscape by layering existing mono/stereo material. In the context of video games, which have high levels of interactivity, it can make production vastly more complicated when trying to implement a format such as Ambisonics into workflows that have been built around channel based audio. It was noted that practitioners are often much more comfortable using tried and tested methods given the intense time pressure involved in producing modern games. Those working within 360° video and VR seemed to approach the requirement to create a bespoke project based individual audio archive as part of the process when working in this area.

4.6 Recommendations

4.6.1 Automatic panning

Some of the challenges presented by immersive content production may be addressed by the further development of current production tools and, in some cases, the development of new tools and technologies. Ensuring spatial congruence between visual and auditory objects has been highlighted as time consuming, especially when the objects' locations are not static within a scene. Some tools to automate this process have already been commercially developed, (e.g. the object tracker within the Facebook 360 Spatialiser plug-in [256], added which as of the time of writing has now been discontinued), however, responses from participants suggest general issues with reliability.

There are however a number of recent studies that look to automate the detection and spatial positioning of audio objects within audio-visual scenes using a variety of audio-visual signal processing methods. Izhar et al. [380] proposed an object-based 3D audio-visual tracking system, which is able to track an unknown and variable number of sources, and utilises iterated-corrector probability hypothesis density filtering [381] to fuse 3D positional estimates

from both audio and visual modalities. The proposed system utilises a audio-visual sending array consisting of an 11-element light-field camera-array and a 16-element microphone array. The 3D positional estimates from the visual data are derived using a human pose detector, which assumes the positional value of the detected nose joint in the position of the sound source. Positional estimates from audio data are then obtained via the steered response power with phase transforms of the acoustic signals at the 16-element microphone array. The resulted showed that the audio information successfully compensated for missed detections from visual-only tracking, increasing recall from 91% to 100%.

As part of the same wider project, a visually supervised speaker detection and localisation system is presented in [382] that utilised an audio CNN trained using a teacher-student paradigm [12]. The teacher network was an audio-visual speaker detector with an additional face-tracker and the student network was trained to regress the horizontal position of the speaker using signals from a 16-element microphone array that had been processed using a spatial beamformer. This built on previous work presented in [383], which trained a vehicle tracking using using stereo microphone array signals as input to a student network which was trained to match the output of a visual vehicle tracking model and [384], which employed binaural audio for semantic segmentation of 360° street views.

However, whilst there is indeed on-going research into methodologies which could assist in the automatic positioning of audio objects within a visual scene, many are reliant on input in the form of either 3D audio-visual data or multi-channel audio. In many cases producers of IME content may have to produce sound design for visual scenes which were not captured alongside multi-channel audio. Additionally, to the authors knowledge, there are limited cases of such algorithms being integrated into software packages compatible with commonly used DAWs. The use of additional computer techniques could improve object tracking within a scene, where associated on-set multi-channel audio has not been captured, and, through the use of object classification, may be able to candidate sound effects files from a chosen repository, reducing the time taken to select appropriate sound effects. Chapter 5 details the design and evaluation of a

potential proof-of-concept computer vision driven audio production tool.

4.6.2 Distance emulation

The desire to simulate auditory distance is not new. The manipulation of digital audio signals to simulate the psychoacoustic cues for distance have well established methods within audio production and signal processing [197]. In fact, all the studies discussed in Section 2.4.5 used a variety of approaches with which to simulate auditory distance that involve manipulation of one, or a combination, of level, reverberant energy, and spectral content. However, participants felt that While traditional audio production methods may be enough to approximate the general perception of distance, the accurate simulation of distance with standardised techniques is lacking in current tools. Within the context of distance emulation via use of reverberation, Coleman et al. [385], highlights that although there are current standards, such as the ITU ADM [113] and MPEG-H [120], which contain parameters such as distance, spread, and diffuseness and may be used to render a reverberant signal, and thus aid in the emulation of source distance, they do not support the concept of a standardised reverberation object. As a potential solution to this, Coleman et al. [385] proposed the Reverberant Spatial Audio Object (RSAO) as a framework for standardising the synthesise of reverberation inside an object renderer. The RSAO framework models an RIR as a set of early reflections in combination with a late reverberation filter, with an RSAO object being described according to a set of reverberation parameters that may be estimated from measured RIRs. The parameters values can then be edited to alter the listener's perception of room size, source distance, and envelopment. The RSAO framework was then extended in [386], to include parameterization of B-format RIRs making it compatible with existing spatial reverb libraries.

There have been numerous studies investigating the modelling and synthesis of distance-dependent HRTFs [387–394]. Methods proposed for distance-dependent HRTF synthesis have included, the use of near-field binaural cues, such as ILDs [387] and the application of an auditory parallax model for modelling near-field effects [388, 389]. The Distance Variation Function (DVF), proposed by Kan

[390], was applied as a filter to derive near-field HRTFs from far-field HRTFs. The DVF was then modelled as a low-order filter for use in dynamic head-tracked audio [393] with [392] combining it with an Image Source Model [322] to provide additional DRR cues. It is of note that most HRTF approaches to distance rendering focus on the near-field, as far-field HRTFs can be considered distance-independent as in the far-field the perception of distance is much more tied to cues such as level and DDR rather than the effects of the head, pinnae, and torso. Alongside established spherical head models, such as those proposed in [64], models have also been proposed which additionally include the effects of the neck and shoulders [391]. Other methods proposed have manipulated the inter-channel relationships between loudspeakers such as the inter-channel phase difference [395] and inter-channel coherence [396].

In recent years, the application of ML to both auditory distance estimation and auditory distance rendering has become an active area of research [397–399], although estimating source distance has currently received greater attention than distance emulation. Zhang [394], proposed a method which modelled HRTFs as weighted combinations of spatial principal components [400] with a DNN trained to predict the spatial principal component weights required for different distances. Physics-informed NNs have been explored to reconstruct the early part of RIRs [401], which is relevant to distance rendering as the early portion of RIRs often contain information about room geometry and source position.

A CNN was proposed in [398], for the task of predicting the likelihood of acoustic reflectors at specified distances and directions-of-arrival from stereo IRs convolved with Gaussian white noise. Yiwere and Rhee [397], approached sound distance estimation as a classification task, utilising a convolutional RNN trained on log-mel spectrograms representations of speech signals which were reproduced through a loudspeaker and captured by microphones positioned at three different distances. Results showed high classification accuracy scores when the model is trained and tested on the dataset collected by authors, however failed to general well to unknown environments. The authors acknowledge that to improve generalisability will require the use of a larger dataset containing a more

diverse range of environments. The study does, however, provide evidence that time-frequency spectral representations, in this instance log-mel spectrograms, do contain distance dependent information. Additionally, IPDs have also been shown to be effective for predicting both angular position and distance [402]. It is therefore not unreasonable to think that if a system is able to estimate source distance given a set of inputs, with the system extracting features and approximating the mapping function of signal features to source distance, then this may be expanded to allow a system to take a specified distance as input, alongside other environmental parameters, and apply the learnt signal features relating to source distance to a given raw audio signal. The initial starting point for developing such a model, in theory, would be similar to the end-to-end neural audio effects detailed in Section 2.7.1.

Since a person's ability to estimate distance becomes more accurate in situations where congruent visual and auditory stimuli are present [403], it poses the question of whether applying both auditory and visual data would allow machine learning algorithms to develop a more complex representation of the problem space. If this is possible then there is the potential for them to be used to inform cross-adaptive audio processing [17] within an audio visual context. An example might be a sound producing object within a visual scene having its distance estimated from associated visual information. Based on this prediction, parameters for EQ, reverberation, and level are then set according to features mapped from a prediction in the audio space corresponding to the distance estimated. This is similar to SoundNet [12] that utilised the natural relationship between sound and vision to learn acoustic representations from videos for the purpose of acoustic scene classification. This kind of approach would be similar to that taken the by models discussed in 2.7.1, which act as audio effect parameter estimations, but in this instance parameters for multiple effects would be estimated simultaneously. This provides two potential approaches when framing the problem of distance emulation as an DAFx problem. NNs can either be used to estimate control parameters for existing/novel digital audio effect structures, or it can be modelled as an end-to-end problem, where the model predicts the mapping functions and

applied it to the audio as a transform function.

4.6.3 Upmixing

It may also be of benefit to further explore the possibilities of upmixing mono/stereo content to scene-based formats, such as B-format or HOA, as this would alleviate some of the issues surrounding the lack of spatially recorded sound libraries and also allow for legacy stereo content to be more easily integrated into spatial audio productions. It should be noted that between the commencement of this study and the writing of this thesis, there has been an increase in the available number of spatially recorded sound libraries, particularly in B-format. The development of more advanced upmixing algorithms would, however, still provide the benefit of of enabling the use, and archiving, of legacy format into a playback agnostic format.

Much of the current research into upmixing methods focuses on channel based formats [404–406] and are based on decomposing the original signal into into its primary-ambient components and then applying processing to generate additional signals specific to the target loudspeaker configuration [407]. Latitnen proposed methods for converting two-channel stereo [408] and 5.1 [409] audio recordings into B-format, but as an intermediary signal for the purposes of reproduction using Directional Audio Coding (DirAC) [149]. The signals were processed using traditional upmixing techniques, explored in more detail in Chapter 6, to increase coverage around the listening positioning, with the resulting channels then rendered over a virtual loudspeaker array. These signals were then encoded into B-format to allow for reproduction using directional audio coding (DirAC) over the original loudspeaker configurations. However, it is not known whether testing was carried out with respect to rendering the signals converted to B-format over loudspeaker configurations with higher channel counts than the original stereo or 5.1 content. Other methods proposed include the generation of 360 audio based on information extracted from 360 video [410] and a multi-model approach to mono to binaural upmixing, which injects visual features maps into the audio feature vector to enable joint audio-visual analysis. Features are

extracted and encoded using a ResNet-18 [411], which are then transformed into complex spectrogram masks through up-convolutions, which can then be applied as binaural filters to original mono signals.

4.7 Summary

This chapter has presented an investigation into how individuals within the sound design industry have responded to spatial audio production for IME content. A thematic analysis of the data was presented which identified underlying patterns and themes. The results were then discussed within the context of common topics that emerged across the themes.

Immersive experiences aim to provide a user with a more intimate experience than traditional media, often placing them either within narrative or allowing them a more true to life perspective. Alongside the use of technologies utilised to create these experiences, it was felt that the difference in end user experience is what defined this type of content. Specifically, it enabled the user to feel present in the XR environment through the presentation of sensory stimuli comparable to that which would occur in a physical environment, with interactive content providing the user with a further sense of agency and involvement within the narrative. Though there is sometimes a difference in semantics, clear associations can be drawn between what professional practitioners feel is important in generating immersion, and the different dimensions of immersion as explored in more academic literature.

Many of the challenges faced by immersive content producers are technological in nature: results from the data analysis suggested that participants felt that the available audio tools were unable to replicate complex psychoacoustic phenomena such as distance, and those designed to assist in the spatialisation of audio associated with objects in a visual scene can be unreliable. However, with IMEs being new to many users there may be a novelty effect masking some of the current inadequacies of the technology as highlighted by participants. The question raised is how long such a potential novelty effect might be sustained and

will immersive production tools and practices advance ahead of users' awareness of and desire for increased quality. There are also challenges associated with working with non-experts, both in the context of clients commissioning IMEs and other practitioners that are new to the area. While this kind of challenge may fade as the medium becomes more established, education initiatives, like those available through the BBC Academy [412], may not only help to alleviate this, but may also assist the speed at which the wider industry adopts this new form of content.

Spatial audio production for IME content might still be considered to be in its infancy, having only in the last decade started to come into its potential with the rise of affordable consumer level XR technology. This study has highlighted challenges for some of those working in the field and their view on what defines immersive content and demonstrates the value in collaboration with professional practitioners in identifying directions for future research and tools/technology development that satisfy the current needs.

Key challenges noted in this study were the time consuming nature of audio panning, the lack of available spatial sound effects libraries, and the challenge integrating legacy stereo content into spatial audio projects, and finally, the lack of standardisation with respect to source distance emulation and the difficulties in simulating distance using traditional audio production methods.

The rest of this thesis will document the research undertaken to address some of the challenges highlighted within this chapter with Chapter 5 exploring a prototype system created to streamline the process of positioning audio objects within a scene.

Chapter 5

Deriving Audio Metadata from a Visual Scene

5.1 Introduction

The work presented in the previous chapter highlighted some of the challenges associated with spatial audio production for IMEs including the time consuming nature of audio object spatialisation, the difficulties associated with replicating auditory source distance, and the lack of spatial sound libraries coupled with the challenges related to integrating legacy stereo content into spatial audio projects. The results from this study were invaluable in highlighting further areas of potential research and in helping to steer subsequent research. The study was also, to the author's knowledge, the first published investigation into the practice of spatial audio production for IMEs that focused on the perspectives and insights of those practitioners working in the area.

Of the technical challenges highlighted in the previous chapter, it was decided to focus on developing interventions that targeted the integration of legacy stereo content into spatial audio productions, and the tracking of sound-generating objects within a scene to assist with the spatialisation of audio objects . Of the potential research areas outlined in the previous chapter, this challenge of sound spatialisation was chosen as the first to be investigated as it was a common

theme brought up by all those who were interviewed and it was an area that was felt had the potential to be one of the most impactful. This chapter therefore presents the development and evaluation of an early stage prototype system that was used to conduct a feasibility study into whether computer vision algorithms, such as those used for object detection, could be utilised to streamline aspects of the immersive sound design process by facilitating the detection, spatial tracking, and content matching of appropriate audio assets to sound generating objects within a visual scene. For the purposes of this thesis this system will be referred to as the *content matching and tracking system.*

As such, the scope of the study was limited to the use of existing and open source computer vision tools. This study was undertaken using simple 2D scenes from which the system derived stereo panning data was derived and suggested candidate sound effects files from the BBC's sound effects archive [25] were identified.

5.2 Visually Driven Sound Design

Computer vision is an established area of machine learning, which focuses on making sense of the information contained within digital images and videos. These techniques are used within a variety of applications including autonomous vehicles [413], surveillance [414], and estimation of HRTFs [415]. Cross-modal or multi-modal, are also relevant given our desire to derive audio-centric data from visual information. A machine learning model can be considered multi-modal if it is designed to process information from multiple modalities [416, 417].

Within the field of sound design, there are some examples of how visual features can be matched to audio files in a database or used to synthesise sounds from this visual information [417, 418]. Owens *et al.* [418] trained a recurrent neural network (RNN) to map visual features to audio features which were then transformed into a waveform by either matching them to already existing audio files in a database or by *parametrically inverting the features.* The sounds synthesised were of people hitting and scratching different surfaces and objects

with a drum stick. While this is still somewhat distant from complex soundscape creation, this outlines a general approach that could be used in order to produce other plausible sound objects. As acknowledged by the authors, the algorithm performs a very rudimentary version of automatic Foley [418], a sound design process where character driven sound effects are created live and added to films in post-production to enhance realism (e.g. footsteps, rustling of clothes), which may potentially be adapted to replicate a wider array of sound effects. Performance of the model was measured via a psychophysical study using a two-alternative forced choice test, where participants were required to distinguish between real and machine-generated sounds. Results were mixed, with parametric generation performing well for materials which were considered more noisy (e.g.leaves and dirt), but performed poorly for harder surfaces such as metal and wood. It was also found that matching the mapped audio features to existing audio files was ineffective for textured sounds such as splashing water.

Object detection (localisation of objects within a given image) and object classification (estimating which of a given class the object is most likely to belong to) are two common computer vision tasks, with algorithms often having to deal with evaluating multiple objects within a given image. It is systems such as these that will be leveraged to investigate whether computer vision can be used to derive useful metadata about potential sound producing objects within a visual scene.

5.3 System Design

The results from Chapter 4 highlighted that sound spatialisation was considered a challenge in immersive productions due to the labour intensive and procedural nature of the task. One of the goals for this study is to address that challenge by developing a method of automatic panning that utilises the information within a visual scene to derive appropriate panning data for audio objects. The resulting panning data can then be taken as a starting point or spatial template for the scene and be fine-tuned by those working on the project. This section describes the

design and implementation of the proposed content matching system, including how the audio metadata generator and candidate sound effects file recommender are integrated with an existing computer vision system.

5.3.1 Google's Object Detection API

As discussed in Section 5.1, the scope of this study was limited to the use of existing open source computer vision tools. Google's Object Detection API [419], was selected because it was free to use, came with a set of comprehensive tutorials and was open source. These factors contributed to it being an ideal base system for the development of the audio content matching and tracking system which would be built on top of existing computer vision functionality. The API is written in Tensorflow [420] and is designed to detect, locate, and classify content from individual 2D images. When implemented as part of a loop, however, it can be used to iterate over consecutive frames of a video [421].

There are a large collection of models within their *Model Zoo* [422], a repository of detection models pre-trained on various datasets, with options suitable for a variety of memory, speed, and accuracy requirements. For details on the performance of each model see [419]. The models are available with frozen weights trained on the COCO dataset, which can be used for off-the-shelf detection, as well as the facility to retrain the models for specific tasks utilising methods such as transfer learning [423]. When used in this way, however, it does not posses any tracking functionality or method of maintaining object identities or inter-frame relationships. It treats every frame as an isolated standalone image.

The model used for this study was the Single Shot Detection (SSD) meta-architecture, with the inception V2 feature extractor, chosen because it presented a balance between speed, accuracy, and memory usage. This was used with the available frozen weights and inference was run on a machine with an Intel Core i5-600 CPU @ 3.20GHz with 8GB RAM.

The API provides a variety of data related to each frame including number of detections, classes detected, detection scores (confidence), and object bounding box coordinates. The content matching and tracking system first collates the

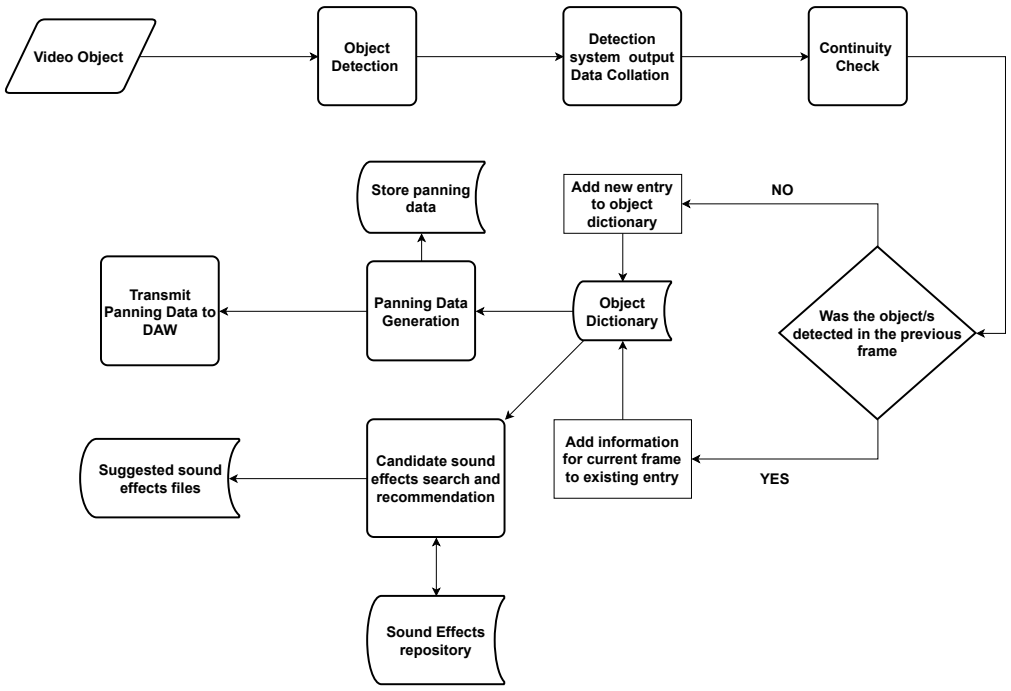


Figure 5.1: Flow chart illustrating order of operations and flow of data within the proposed methodology

data for each frame so it can be used to create the object dictionary. The object dictionary contains a unique ID number for each detected object, class number of the object detected, and the coordinates relating to the bounding box position of each object. Following the collection of this data it is then used in several processes outlined in the following sections. Fig 5.1 shows a block diagram of the complete system.

5.3.2 Tracking

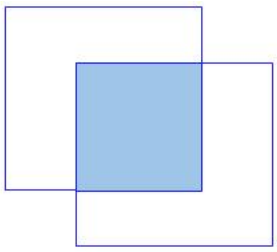
For the system to operate over successive images, as would be required when applied to video data, it must be able to group the data for each detected object that persists over multiple frames and create new object IDs if a detected object is considered as being new to a scene. Video data will usually exhibit temporal continuity and information from earlier frames, such as object location and object class; this can often be used as context in order to improve predictions for the current frame [424]. However, the original API was designed for detection on a

single image and is unable to retain any of the context from information that may have been extracted from previous images or frames. This lack of contextual or temporal awareness as Liu and Zhu [424] refer to it, can have a negative impact on performance when compared to models that do contain temporal awareness. Whilst there is ongoing research to increase the accuracy of video object systems (see [424] and [425]) this was deemed outside of the scope of this research given the focus on deriving audio data from existing computer vision systems.

The content matching and tracking system uses a simple continuity check, based on the Jaccard Index [426] of the bounding boxes generated by the API across two frames, to attempt to minimise between frame misclassifications and to accurately group object data across successive frames. The Jaccard Index, also referred to as the Intersection over Union (IoU), is a statistic used to measure the similarity between two sets of data and is defined as the intersection of the datasets (Figure 5.2a) divided by the the union of the two datasets (Figure 5.2b) as given in Eq 5.1.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{5.1}$$

The Jaccard Index is also a commonly used metric for training and evaluating object detection algorithms as it can be used to compare the similarity between two arbitrary shapes and is scale invariant [427]. During training, the metric is calculated from the ground truth bounding boxes and the predicted bounding boxes. Accuracy is deemed sufficient if the resulting index value exceeds a user specified value e.g > 0.5, with values ranging between 0 and 1. This is also an appropriate metric to use as a simple interframe continuity check since it is expected that an object's position within the current frame will be similar to that in the previous frame. If the resulting $J(A, B)$ is greater than the threshold value, the object is defined as being the same as that identified in the previous frame, otherwise it is treated as a new object and a new entry is added to the object dictionary. The implementation used within this work for deriving the Jaccord Index from two bounding boxes can be found in [428].



(a) Area of Intersection

(b) Area of Union

Figure 5.2: IoU can be calculated by dividing the area of intersection (the area covered by the overlap of the two boxes) by the area of union (total area covered by the two boxes). Within this work it is used as a continuity check on objects within the visual scene taking advantage of the similar locations an object will occupy within the current and previous frame.

5.3.3 Sound Effects Suggestions

Once the object dictionary has been compiled, it is used to generate a list of suggested sound effects from the chosen repository of audio files, which in this case is the BBC sound effects archive (BBCsfx) [25]. A list of candidate sound effects files are compiled by comparing each unique object class detected to the metadata tags from BBCsfx. BBCsfx is an open source repository that, at the time of the study, was made up of 16,011 labelled audio files and has since been increased to contain over 33,000 files. The archive is available to download as WAV files and is subject to terms of use under the RemArc Licence [429], which permits use for personal, educational, and research purposes. It was chosen because it provides a large database of labelled audio files containing a variety of different acoustic scenes and events, with tagging and metadata stored in an associated .CSV file. Table 5.1 shows examples of the tagging and metadata format common

to each audio file in the database. Tagging consists of the description of each sound effect (as taken from the original CD the sound effect was sourced from) and the category (e.g. Engines: Petrol, Engines: Diesel) to which it belongs. The metadata associated with each audio file is the length of the audio file in seconds, the name of the original CD containing the effect, and the track number. There are some inconsistencies within the tagging conventions, such as some audio files lacking an associated category and/or CD origin name. Any inconsistencies within a database's tagging convention may impact its effectiveness when used as data for training and evaluating machine listening systems [430].

Description	Duration (s)	Category	CD Number	CD Name	Track #
Two-stroke petrol engine driving small elevator, start, run, stop.	194	Engines: Petrol	EC117D	Diesel and Petrol Engines	4
Single-cylinder Petter engine, start, run, stop. (1 1/2 h.p.)	194	Engines:Diesel	EC117D	Diesel and Petrol Engines	1
Single hen	63		EC31A	Chickens	1
Motorcycle Scrambling: General atmosphere, pre-1965 machines, 250-500cc	194	Motorcycle Scrambling and General Atmosphere	EC5M4		1

Table 5.1: Examples of the metadata format associated with the BBC's sound effect archive. Available metadata fields consist of a description, duration in seconds, category, CD number, CD Name, and track number. As shown, there is inconsistency within the archive as not all audio files will contain information within the category, CD Number, and CD name fields.

5.3.4 Object Tracking

Object location data can also be used to derive the trajectory of objects over the course of a video. This data can then be utilised to position and pan audio

content. Object trajectory is derived by calculating the centre point of an object's bounding box as shown in Fig. 5.3. The data can then be transmitted to a Digital Audio Workstation (DAW) via OSC [431] to populate automation data for the desired parameter. In the case of stereo panning, the horizontal portion of the trajectory data needs to be normalised to between 0 (hard left) and 1 (hard right). Due to the temporal resolution available within the automation lanes of the DAW used in this study, Cockos Reaper [432], resolution of location data was reduced by a factor of two, resulting in 15 discrete points per second for a 30fps video.

Figure 5.3: Single frame taken from a test video with the preceding trajectory of the detected object overlaid.

5.4 Test Material Specification

Two test videos were created to allow for direct and controlled evaluation of simple scenes containing single and multiple objects. An open source image containing non-human animals was also sourced from the internet to assess the ability of the system to recommend appropriate candidate audio files for non-human objects. Both videos were recorded on an iPhone SE at 1080p 30 frames per second at a

distance of 5m and have the following conditions:

- Video 1 − Single person walking from left to right of scene.
- Video 2 − Two people walking ∼1.5m apart from left to right of scene

Example frames from the two videos are shown in Fig. 5.4 and Fig. 5.5.

Figure 5.4: A single video frame extracted from example Video 1, and used as input for the object detection system to generate candidate audio file recommendations. The location of the detected obkect is indicated by the green bounding box and is assigned the class label of 'person'.

Figure 5.5: Image from a single video frame of Video 2 used to derive panning information for two moving objects with a 2D visual scene. The example video is of two people crossing the field of view from left to right approximately 1.5m apart.

5.5 Results

5.5.1 Run time for data extraction

It took approximately 75 seconds to run detection and information extraction on a 7.97s video @ 30 frames per second (fps). This roughly equates 0.32s per frame but can be reduced to 65s (0.27s per frame) if the output of the detection algorithm is not visualised. It should be noted that inference was run on the CPU as a GPU was, at the time, not available. It is reasonable to assume that runtime would have been several times faster had a GPU been utilised.

5.5.2 Spatial Positioning and trajectory tracking

5.5.2.1 Single Object

Fig. 5.3 shows a single frame taken from Video 1 where the trajectory of the detected object has been plotted. The trajectory appears to accurately track the object travelling across the field of view whilst also taking into account the slight vertical movement of the centre point of the object's bounding box. This

vertical movement occurs due to the vertical movement of the human body whilst walking [433]. Finally, it tracks the variation in the object's speed which in this case is indicated by the non-uniform distribution in the spatial proximity of the data points.

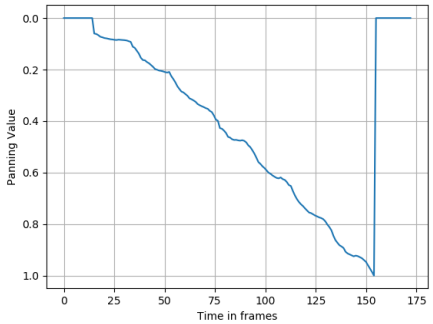
Figure 5.6a shows horizontal panning data plotted over time in frames derived from the positional data of the object as derived from Video 1 as shown by way of example in Figure 5.3. The changes in the gradient of the data represents the variation in the speed of the object as it crosses the field of view. It should be noted that it is the distance moved by the centre of an object's associated bounding box between each frame that is being tracked, rather than the object itself. For objects whose movement causes bounding boxes of varying sizes, such as a human walking with their arms swinging, this may produce variable results. Once the object exits the field of view, the panning value defaults to 0, which may present problems for objects whose audio needs to remain active, even if no longer visible. This, however, is an issue relating to the current 2D only implementation and it would be fairly trivial to introduce an option to maintain an object's last known position once it has exited from the visual field of view. This is less of an issue with 360°audio/visual content as the field of view is dictated by the direction a user is facing, therefore allowing objects outside the field of view to be still be tracked as the video content extends beyond the limits of this region.

Figure 5.6b shows the horizontal trajectory data translated into panning values within a Reaper stereo track automation lane. Upon visual inspection, the reduction in data resolution explained in Section 5.3.4 does not seem to have had an adverse effect on trajectory trends. The linear interpolation generated by Reaper has little impact on the overall trend due to the size of the timesteps but may have a perceptual impact for larger timesteps. The timestep is defined as the length of time between each discrete data point of panning data and is dependent on both the fps of the video and the granularity available with the DAW processing the panning data. At the time of the study the minimum timestep possible for Reaper's panning data was 0.667s, which is equivalent to 15 fps. A reduction in fps results in an increased timestep duration that

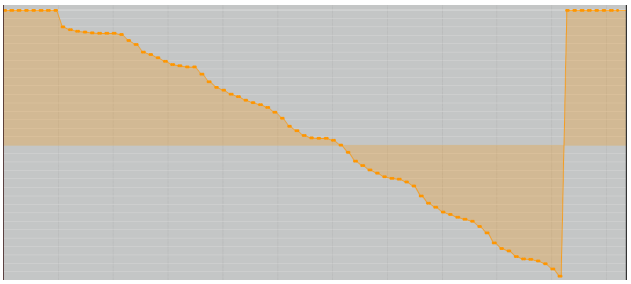
may introduce greater spatial mismatches between the visual object and the associated auditory material. The reported angular offset between visual and auditory stimuli that would result in a perceptually noticeable misalignment varies greatly within the literature. The ventriloquist effect, where a sound source is spatially localised towards a visual stimulus, has been shown in several studies to vary between 20° - 38° azimuth angle [341, 434, 435] in real environments. For screen-based or virtual environments studies have reported offset ranges from 3° to 45° depending on experimental conditions. These are, however, the maximum reported values rather than the JNDs which is the smallest offset at which the location of visual and auditory stimuli can be differentiated. Results have also shown that for speech signals even small audio-visual offsets can subconsciously influence the spatial integration of sources [436].

5.5.2.2 Multiple Objects

Using the API to facilitate the tracking of multiple objects within a scene introduces some additional complexities. If the API is presented with a scene containing multiple objects, it will store and output the data for the detected objects in descending order according to the associated confidence scores. This results in the object data being outputted in a different order for each frame depending on how the confidence scores change between frames. The object data is processed and stored by the object dictionary according to the results of the continuity check, detailed in Section 5.3.2, which uses information from the current and previous frames. The use of inter-frame data introduces a reliance that an object represented by data at $current\_frame(obj\_index)$ is the same object represented at $previous\_frame(obj\_index)$. When this is not the case, it causes the continuity check to compare positional data from different objects, often causing erroneous results. In cases where the two sets of data being compared are from objects with a large enough distance between them to cause the continuity check to fail, this will cause the object at $current\_frame(obj\_index)$ to be defined as a new object and a new entry generated in the object dictionary. This can cause trajectory data from what should be a single object to spread



(a) Original data output from system. Note the y axis has been flipped to match Reaper's and the data has been normalised to between 0 and 1 to match the values used by Reaper.



(b) Stereo panning data was derived by using every second data point to account for the resolution available in Reaper's automation lanes.

Figure 5.6: Horizontal panning data plotted over time as derived from example Video 1.

across multiple entries within the object dictionary. In one instance, the changes in confidence scores over the course of a video resulted in a total of 32 objects being added to the object dictionary for a scene consisting of two unique objects. Due to the object detector in this study being based on a single image detector, overriding the ordering method to create a more consist output order on a frame-by-frame basis is non-trivial. This introduces challenges for situations where object detection alongside the ability to distinguish between pre-existing and new objects over time is required. While outside the scope of this work, multiple object trackers such as those described in [437] may provide a solution if an appropriate open source system is identified.

5.5.3 Sound Effects Recommendations

The system takes approximately 4s to compile a list of candidate sound effect recommendations for Video 2, returning a total of 36 recommendations (a selection of which are shown in Table 5.2), of which 6 were considered usable for the given scene. Those deemed unsuitable were for reasons such as a different environment/activity to the one in the example video e.g. a person exiting a car and a person in an ice skating rink. The current search method takes the class label as a string of characters and compares this to the tags in the metadata. If an exact match is found, it will determine the associated audio file as being a candidate sound effect.

A limitation of this method is the reliance on exact matching between the tags in the repository's metadata and the class labels of the detection system. Due to this, the current search method is unable to recommend audio files which may be suitable but whose tags use different, but related, terms, such as 'man', 'woman', 'child', or 'human' if detecting the class 'person'. At the time of the study, the tagging within the BBC archive was inconsistent (admittedly due to the repository consisting of many decades worth of archived audio files) meaning many potentially suitable sound effects go undetected using the current string comparison method. To avoid the need for an exact match between metadata tags and class labels, a word embedding model, such as Word2Vec [438] or GloVe [439], could be used to evaluate the similarity between the class labels and metadata tags and return the *top-n* most similar tags and the associated audio files.

There are also limitations that stem from the type of detection system used. Google's API is intended for object detection and is limited to the detection of specific objects; it does not have the ability to predict actions or activities taking place within a scene, such as the *walking* present in both test videos. As such, the system did not retrieve the 1,484 sound effects containing the term 'footsteps' which may have been suitable as candidate sound effects. The interested reader is referred to [440], for a review of recent computer vision based methods for human action recognition, which may go towards resolving

Candidate Audio File Recommendations
Walking, 1 person in mud
Footsteps, one person walking in mud
Cars: 1.6 GL (Manual) 1982 model Ford Cortina. Interior, door opens, person exits, door closes
Ice Skating, one person circling close, others in the distance on indoor rink
Footsteps, one person walking in water

Table 5.2: Selection of candidate audio file recommendations generated from Fig. 5.4. Each file was defined by the system as being a potential candidate if the metadata field 'description' contained an exact match for the detected objects class name, in this case 'person'.

this challenge. Additionally, the system proposed in this chapter also lacks the functionality of scene recognition systems to predict more generic scene elements such as location (e.g. living room, beach, city centre) which may help to inform recommendations for audio files relating to environmental/atmospheric sounds. Scene recognition is considered a more challenging problem when compared to other tasks such as object recognition [441], as it often involves the segmentation of the given image into a variety of spatial layouts and not only requires the detection of objects within the scene but also the semantic relations between the detected objects. The inclusion of models, such as those reviewed by Xie et al. [441], would go someway towards addressing this issue.

Figure 5.7 shows that erroneous results can also be produced according to the accuracy of the object detection system in relation to the given scene. In this instance, an animal which would be recognised as a type of antelope to the human eye has been classified as a cow. In turn this has produced recommendations unsuitable for the given scene. The system also failed to recommend several audio files from the repository whose tags contained the word 'zoo', which may have been appropriate for ambience. Again, this is due to lack of ability to infer wider context from the scene.

Figure 5.7: Output of Google Object Detection API, showing correct classification of 'giraffe' (centre) and 'zebra' (right), whilst incorrectly assigning the class label of 'cow' to an antelope (left).

5.6 A Review of Methods to Inform Future Work

The continuation of the work presented in this chapter would have required a more extensive treatment and technical investigation of computer vision technologies. As such, it was deemed outside the scope and aims of this thesis. However, in order to provide both an insight into what may be required to develop the work further and an assessment as to whether the current methods would deliver the creative affordances the work in this chapter explored, a review of the current state of the art computer vision technologies with respect to multiple object tracking and object detection and classification is now presented.

5.6.1 Object Detection and Classification

Object detection and classification describes a task that combines both image classification and image localisation, but for scenarios where the classification and

localisation of multiple objects may be required [442]. According to Wu, Sahoo, and Hoi [443], deep learning approaches can be broadly categorised as either being one-stage detectors [444–450], or two-stage detectors [451–456]. Two-stage detectors, such as R-CNN [451], Fast R-CNN [453], and Mask Scoring R-CNN [456], separate the detection task into two stages. The first stage, proposal generation, attempts to identify regions within the image (or frame if part of a video sequence) which may potentially contain objects of interest. The second stage, the classification stage, then uses a model that attempts to map the proposals to a categorical class label and may additionally try to refine the proposed regions [457]. One-stage detectors, such as the Yolo family of algorithms [445, 448, 449], by contrast, consider all positions on the image as potential objects and attempt to classify each region as either background or a target object [443]. Early deep learning approaches considered object detection as as multi-region classification problem, with Sermanet et al. proposing Overfeat [444], a CNN which modified a CNN classifier model to use the final fully connected layers to output a grid of predictions for each region of the input, indicating the presence of an object. You Only Look Once (YOLO) was then proposed in [445], which approached object detection as a regression problem by spatially segmenting the image into a fixed 7 x 7 grid. Each grid segment was then considered as a proposal region to detect one of more objects. Each grid prediction consisted of, a class label, the bounding box coordinates and size, and whether the location contained an object or just background. Due to the unified architecture, as there was no separate proposal stage, the base model could process images in real-time at 45 fps. Although, it should be noted that the ability to process individual frames at real time speed does not equate the algorithm to being that of a video object detector, as will be discussed later. Some limitations of the original YOLO model include, each grid segment being limited to the prediction of boxes and a single class, which results in the model struggling with small objects, such as flocks of birds, or large crowds if images are captured from a distance. Several iterations of YOLO have been subsequently proposed which improve upon the original design, including YOLOv2 [448], which utilised a convolutional model

Dataset	# images	# train/val/test splits	# classes
Pascal VOC2007	10,022	2,501/2,510/5,011	20
Pascal VOC2012	22,531	5,717/5,823/10,991	20
MSCOCO	163,957	118,287/5,000/40,670	80
Open Images	1.9M	x	600
ImageNet	14M	x	1,000
Wider face	32,203 (400k faces)	12,881/3,220/16,101	x
FDDB	2845 (5171 faces)	x	x
CityPersons	5,000 (35k people)	x	x
CIFAR-100	60,000	50,000/x/10,000	100

Table 5.3: Publicly available annotated MOT datasets.

that was pre-trained on higher resolution images from ImageNet, which enabled it to captured finer detail. Additionally, it adopted the use of anchors proposed in [446], which segmentation of feature maps at varying resolutions.

CNNs are often used as the backbone for neural network based object detectors [458], and act as a feature extractor, which generates the feature maps that will then be used for classification by the fully connected layers. Given the computational resources required to train a SOTA feature extractor from scratch it is common place to utilise a model that has been trained on large scale image classification datasets [451] and, if required, it may be fine-tuned using a smaller dataset more closely related to the target task [459]. For example, if a production studio specialised in a specific format e.g. wild life documentaries, a pre-trained classifier could be taken and fine-tuned on existing content to improve its performance on the target content type.

There a number of common datasets available used for benchmarking difference object detection tasks, for face detection, pedestrian detection, and generic object detection. Table 5.3, provides an overview of the summary statistics for some of these common benchmark datasets and although non-exhaustive, the datasets detailed evidence that a large quantity of training data exists with which to train

object detector algorithms on a variety of different object types. This is beneficial as often within an audio production context there can be multiple different sound producing objects within the screen that could belong to a large number of different object classes. Having a large quantity of publicly available data may also assist in the fine-tuning of models in situations where a content creator may not have the required amount of target content to adequately fine-tune a model. This could often be the case if they working on a project that is outside of their normal format. Another important factor is how the object detection algorithms perform under situations where the visual scene may potentially be densely populated and consist of objects belonging a variety of object classes.

There a number of metrics commonly used to evaluate object detection systems. Two commonly used metrics are precision and recall [196]. Recall measures what proportion of objects were correctly detected and precision measures the proportion of object detections which were correct out of the total number of detections made. In other words, recall can be thought of as a measure of how many true positives were identified out of the total number of possible true positives, whereas precision measures the quality of the predictions, calculating what proportion of positive predictions were true positives. These two metrics can then be used to calculate the Average Precision (AP) for each class as follows:

$$AP = \sum (R_n - R_{n-1}) P_n \tag{5.2}$$

where P_n and R_n are precision and recall at the nth threshold.

To produce a metric to describe the performance on the complete dataset, the mean is then taken to result in the mean Average Precision (mAP). As part of a review into SOTA approaches to object detection, Wu et al. [443] compiled a comprehensive list of model performance spanning the last decade. Table 5.4 presents a selection of the best performing models from the last 6 years. Whilst these may be accurate at the tome of writing, the speed of progress in the areas will likely mean new SOTA scores are established by the time this thesis is published.

As shown by the results, two stage models tend to yield better performance

Model	detector type	Backbone	Proposed Year	mAP (%)	Dataset
DCN+R+CNN	two-stage	ResNet-101 + ResNet-152	2018	84.0	VOC2007
DeepRegionLet	two-stage	ResNet-101	2018	83.3	VOC2007
RFBNet512	one-stage	VGG16	2018	82.2	VOC2007
CentreNet	one-stage	ResNet101	2019	78.7	VOC2007
DeepRegionLet	two-stage	ResNet-101	2018	81.3	VOC2012
DCN+R+CNN	two-stage	ResNet-101 + ResNet-152	2018	81.2	VOC2012
TridentNet	two-stage	ResNet101-Deformable	2019	69.7	MSCOCO
CenterNet511	one-stage	Hourglass-104	2019	64.5	MSCOCO

Table 5.4: Publicly available annotated MOT datasets.

results when compared to one-stage models, however the trade off for this is that usually two-stage models are more computationally expensive because of the separate stage for proposal generation. Scores are generally lower for the MSCOCO dataset, which is indicative of it being a larger and more comprehensive dataset with respect to both the number of images and the number of unique classes. This also provides evidence that given the performance values and the wide variety of datasets with which an object detection system may be trained on, choosing any current SOTA model could be expected to yield good results for object detection within an audio production context, providing that the algorithm is capable of detecting the required classes. But given the varietry of objects available within the datasets, it is not imagined this will present an issue. This would especially be the case when combined with model fine-tuning in order to improve performance on specific target content formats or object classes.

However, it should be noteds that video object detection is a more complex problem when compared to image object detection, as it introduces a temporal dimension to the information. As noted in Section 5.3.1, and highlighted by [457], if an image object detection algorithm is applied directly to video data without any modification, each frame is treated as standalone unrelated images. Video object detection therefore therefore requires that object identities can be maintained across frames, known as object tracking.

5.6.2 Multiple Object Tracking

Multiple object tracking (MOT) typically has two additional requirements when compared to single object tracking, firstly, the ability to determine the number of objects present within the frame, and secondly, maintaining the identities of the objects between continuous frames, a problem known as data association [437, 460, 461]. It was the latter of these two tasks that the proposed system failed to achieve. Luo et al. [437], highlights that there are a number of challenges within MOT which add complexities with respect to those two tasks, including, short and long term occlusions, initialisation and termination of tracks, objects with similar appearances, and interactions among multiple objects. Depending on the how the MOT task is approached and the type of MOT methodology used will dictate what solutions are employed to address those tasks and challenges.

Luo et al. [437], suggests that MOT algorithms can be categorised according to three criteria, 1) *initialisation method*, 2) *processing mode*, and 3) *type of output*. The categorisation by type of output refers to whether the tracking process is considered stochastic or deterministic, which is largely determined by the optimisation methods adopted. As this work is concerned with the novel application of existing technologies to the problem of immersive audio production, and not the development of new tracking methods, this criteria is considered outside the scope of this review. The first criteria, the initialisation method, refers to how the objects are initialised, with MOT approaches broadly be categorised as either detection-based [462], or detection-free [463]. Detection-based tracking, sometimes referred to as model-based tracking, utilises a pre-trained object-detector applied to each frame to obtain a set of object hypotheses, which are then linked to associated object trajectories [462, 464]. Detection-free tracking, sometimes referred to as model-free tracking, in contrast, requires the objects of interest to be manually annotated in the first frame but no requires no further knowledge about the objects [463]. The objects annotated within the initial frame are then tracked through subsequent frames. However, the due to the requirement of manual initialisation, the number of detected objects is fixed and

as such, new objects that may enter the scene will not be identified [437].

The tracking methodology proposed in Section 5.3.2 would be categorised as detector-based, as it used a pre-trained object detector. There are some disadvantages associated with using a detector-based tracking approach, namely that the performance of the MOT algorithm is then largely defined by the performance of the chosen object detector [437, 465]. This is with respect to both the localisation accuracy and the specific object classes the object detector is trained to detect. However, although detector-free tracking may enable perfect initialisation, given that the objects of interest are initialised manually, which additionally enables the tracking of any object, tracking is usually limited to the fixed number of objects annotated in the initial frame. Furthermore, the manual annotation required may increase the associated set up time. As such, given one of the main affordances explored in this chapter was the streamlining of sound design workflows, it is suggested that in most cases a detector-based tracking system would be the preferred approach as it will reduce the required time to initialise the object tracker. Additionally, detector-based tracking also has the advantage of being able to automatically terminate the tracking of objects that disappear and initialise the tracking of newly detected objects [464], which will lend itself to the types of scenes encountered by sound designers where multiple objects may appear, disappear and reappear at any given time.

Which respect to categorising MOT by processing mode, trackers are broadly classified as either being online trackers [461, 466–469] or offline trackers [460, 470, 471]. Online tracking methods receive image data and associated detection results for each frame sequentially [437], and determine updates to the object trajectories, initialise new object tracks, and decide when to terminate object tracks, based on information contained with the current and previous frames [467]. It may also be referred to as sequential tracking [437]. The proposed tracking method for the system presented in this chapter would be considered an online tracker, as each frame is handled sequentially and the tracking algorithm only has access to the current and previous frames.

Offline tracking methods, sometimes referred to as batch methods, use both

past and future frames and thus have access to past and future detections [466]. Depending on memory restraints, offline trackers will either process the whole sequence of frames as a single batch or process the sequence as multiple mini batches. The benefit of offline tracking is that, theoretically, it is possible to obtain an optimal solution given there is access to the whole sequence across which objects are required to be tracked [437]. However, the offline nature means it is unsuitable for any system that requires real-time operation or other scenarios where future frames are not available. Therefore, given that in the context of sound design for IMEs the visual content will be pre-recorded or pre-rendered, both offline and online methods of tracking would be suitable for use within an audio production setting. For a review of different methodologies for addressing the challenge of data association in a MOT context, the reader is directed to [437]

It is important to now review the performance and capabilities of the current SOTA in MOT, as this will help to inform the necessary steps required in developing a SOTA system for content matching and tracking within an audio production context. Additionally, a review of the current SOTA will also provide an assessment of how such a system would be expected to perform when applied under the conditions typically encounter within an audio production setting.

The performance of MOT algorithms is generally divided into those used to measure performance with respect to object detection, and those used to measure performance with respect to object tracking [437]. Recall and precision are common metrics for evaluating object detection [196]. Recall measures what proportion of objects were correctly detected and precision measures the proportion of object detections which were correct out of the total number of detections made. In other words, recall can be thought of as a measure of how many true positives were identified out of the total number of possible true positives, whereas precision measures the quality of the predictions, calculating what proportion of positive predictions were true positives.

For the evaluation object tracking, the CLEAR metrics are often used [472], which include, Multiple Object Tracking Accuracy (MOTA) and Multiple Object

Tracking Precision (MOTP). These are both derived by measuring the IoU of the predicted bounding boxes and the associated ground truth bounding boxes. MOTA takes into the account the number of missed detections, false positives, and ID switches in the predicted output trajectory for a given ground-truth trajectory. ID switches, as defined in [473], is a measure of the number of times a tracked trajectory changes its matched ground-truth ID. This can occur either when an object becomes associated with the trajectory of another object, or when an object trajectory is fragmented through missed detections. MOTP calculates a precision score by using the spatiotemporal overlap between the reference trajectories and the predicted trajectories [472]. A study by Leal-Taixé et al. [474], showed that the MOTA is the measure that best aligns with human visual assessment, followed by Mostly Tracked, which is the ratio of ground-truth trajectories that are covered by the predicted trajectories for at least 80% of their respective life span [473]. Leal-Taixé et al. [474], also note that ID switches do not have much impact on the human quality assessment, reflecting that human observers place more importance on objects, in this case people, being detected rather than tracked correctly. In our context, however, ID switching is a highly important metric as it evaluates whether object associations are being correctly determined across frames, which in turn will determine the accuracy of the derived object trajectories and the subsequent audio panning data derived from the object trajectories.

Table 5.5, details a selection of MOT algorithms that have been evaluated on the MOT challenge datasets [475–477], which were chosen as they are one of the most widely used set of MOT datasets and baselines. As can be seen from the MOTA, MOTP, and ID switches scores, there has been a great deal of progress in the performance of MOT systems over the last decade, even when taking into account the increasing complexity of subsequent datasets. For example, MOT20 contains three sequences where crowd density can reach values of 246 pedestrians per frame [477], yet MOTA scores for both the TransTrack [461] and MOT correlation learning [478] are 64.5% and 65.2% respectively, whereas the SOTA for earlier models, such as SORT [467], proposed in 2016 was 33.4%, with

scenes much less densely populated. This increased complexity and object density also explains why the absolute number of ID switches is higher for more recently released datasets, but the percentage based scores are also generally higher. This does, however, highlight that any system deriving audio object positional data from MOT tracking systems, such as those detailed in Table 5.5, would require a robust way to deal with missed detections and fragmented trajectory paths. However, due to the available use of offline processing, fragmented trajectory paths can be accounted for using a solution that utilises the information and results from all frames within the tracking sequence, which may prove more robust than online methods that can only utilise current and previous frames.

Table 5.6 details a non-exhaustive but varied selection of publicly available MOT datasets. One thing that seems clear is that, at present, from both the target objects of the MOT algorithms in Table 5.5, and the datasets detailed in Table 5.6, is that many of the MOT datasets and benchmarks are concerned with the tracking of a single, specified, class of object. As a result, these existing methods can perform poorly when presented with unseen objects [465, 483]. Whilst there are recently presented datasets, such as the Track Any Object (TAO) dataset presented in [484], there appears, to the authors knowledge, limited studies currently using it as benchmark. Additionally, as noted by Fan et al. [485], although TAO contains a diverse range of classes, not all instances of each target class are annotated in a video sequence, potentially making it unreliable if a prediction is required for each object for each frame. Also available are datasets and benchmarks for adjacent tasks, such as ImageNet VID [486], for multiple object detection within videos, these are not annotated with object trajectory information and so would require new annotation and benchmarking. So, even given some of the SOTA performance scores associated with the algorithms in Table 5.5, their application within an audio production setting would be limited due to their lack of ability to accurately track a range of different object classes.

There have been recent studies presented that investigate generic MOT (GMOT), a MOT paradigm that requires no prior knowledge of the objects to be tracked [465, 483]. However, as highlighted by Bai et al. [465], despite its broad

Method/Model	Year	Online/Offline	Target object (single, multi-class, generic)	Architecture	MOTA (%)	MOTP (%)	# ID switches	dataset
SiameseCNN [470]	2016	offline	single	CNN	29.2	71.2	639	MOT15
SORT [467]	2016	online	single	Faster R-CNN	33.4	72.1	1001	MOT15
AMIR [479]	2017	online	single	RNN	37.6	71.7	1,026	MOT15
MOT correlation learning [478]	2021	online	single	CNN	62.3	65.7	x	MOT15
Dual Matching Attention Networks [480]	2018	online	single	RNN/LSTM	46.1	73.8	532	MOT16
LMP [481]	2016	online	single	CNN	46.3	75.7	663	MOT16
DeepFlowNet [471]	2017	online	single	CNN	29.19	x	142	MOT16
AMIR [479]	2017	online	single	RNN	47.2	75.8	774	MOT16
MOT correlation learning [478]	2021	online	single	CNN	76.6	x	74.3	MOT16
Deep Affinity Network [461]	2021	online	single	CNN	52.4224	79.9071	1648.08	MOT17
MOT correlation learning [478]	2021	online	single	CNN	76.5	73.6	x	MOT17
TransTrack [482]	2021	online	single	transformer	74.5	80.6	3663	MOT17
MOT correlation learning [478]	2021	online	single	CNN	65.2	69.1	x	MOT20
TransTrack [482]	2021	online	single	transformer	64.5	80.0	3564	MOT20
DeepFlowNet [471]	2017	online	single	CNN	73.75	x	89	KITTI
Z-GMOT [483]	2023	both	generic	CNN	62.76	x	x	GMOT-40
GMOT one-shot [465]	2021	online	generic	CNN	19.93	24.16	x	GMOT-40

Table 5.5: Details of current SOTA MOT algorithms.

Dataset	# video sequences	# frames	GT	# annotated classes	Target classes	indoor/outdoor
MOT15 [475]	22	11,283	Yes (train set only)	1	pedestrians	outdoor
MOT16 [476]	14	11,235	Yes	4	people	both
MOT20 [477]	8	13,410	Yes	13 (people, vehicles, and occluders)	people	both
KITTI [487]	50	x	Yes	8	car, pedestrian	outdoor
PETS 2009 [488]	3	x	Yes	1	people	outdoor
ETH Pedestrian	8	4,000	Yes	1	people	outdoor
Urban Tracker [489]	5	x	Yes	3	pedestrians, cycles, motorised vehicles	outdoor
GMOT-40 [465]	40	9,643	Yes	All objects annotated	Generic object class	both
VAR19 [490]	30 (test)	2,000 (train)	Yes	8	car, bus, train, truck, minivan, bike/scooter, bicycle, person	outdoor
Tracking Any Object (TAO) [484]	2,907	x	Yes	833	numerous	both

Table 5.6: Publicly available annotated MOT datasets.

applications, GMOT is still under researched when compared to MOT and there is, at present, a lack of GMOT benchmarks and baselines. The baselines presented in [465] were built upon publicly available trackers, with the best performing model still vastly under performing that of traditional MOT algorithms with scores of MOTA = 19.92% (±1.84%) and MOTP = 24.15% (±0.27%). These scores are exceptionally low when compared to some of the results in Table 5.5, where even the lowest scores from earlier methodologies are approximately MOTA = 29% and MOTP = 65%, although it is acknowledged a true comparison is difficult given that scores were obtained on different datasets. It should be noted, however, that when ground truth detections were provided, the results in [465], increased significantly, with the best performing model then achieving MOTA = 80.60%. The increase in performance when ground truth detections are provided suggest that the challenge lies with the generic object detection portion of the problem, especially relating to objects that are not seen during training. Tran et al. [483], offered a solution to the problem of unseen object classes by introducing text prompts through the use of Grounded language pre-training, which utilises existing mappings for image-text pairs [491], and resulted in an improved MOTA score on the GMOT baseline, when using the OC-SORT detector [492], of 62.76 percentage points. The utilises of text prompts to condition a GMOT algorithm may prove beneficial in an audio production setting as it would provide the flexibility to use the same algorithm for a range of different objects and scenarios, without the need for retraining.

There also exists a growing interest in multi-class multi-object tracking (MCMOT), where multi-class object classification is undertaken as part of the tracking process. However, as noted by Jo et al. [493], many tracking datasets are annotated for human only tracking, or in some cases human and vehicle [487], although many will contain other unlabelled objects within the scenes. A multi-class multi-object tracker presented in [494], had to provide separate evaluations for the tracking and detection components, utilising ImageNet VID to assess its detection performance and MOT15 to assess its tracking performance. For detection on ImageNet VID the model achieved a mean average precision

of 74.5% over 30 object classes and MOTA = 62.4% on the MOT16 benchmark. Although this provides indicative performance for both components, it is not ideal as there is a lack of evidence provided for how the system would perform when having to track multiple classes, especially when faced with additional challenges like ID switching, missed detections, and miss-classification of objects. Within the context of an audio production setting, a MOT that is able only able to track a single class of object will have limited application given the variety of sound producing objects that are present in most productions. Given that, further progress may be required to integrate the SOTA object detections algorithms outlined in Section 5.6.1, with the tracking methodologies of the models outlined in Table 5.5, before the kind of system explored in this chapter can be fully realised. GMOT has the potential to afford a greater flexibility than traditional object-specific trackers as it would allow the user to specify on a case-by-case basis which types of objects require tracking and additionally provide varying levels of specificity depending on what the scenario requires. Not only would this provide more flexibility to content creators, but it may also alleviate the issue of surplus information generation, as it would limit the amount of objects that are detected and subsequently tracked.

In conclusion, although MOT and object detection and classification have all seen great improvements in the SOTA in recent years, creating a unified system that encompasses and performs highly in all three tasks is still an open problem. Especially within the context of MCMOT or GMOT. Largely, it appears, that one of the main barriers to progress stems from a lack of suitable annotated datasets where a single dataset could be used as a unified benchmark. The greatest challenge for further developing a system such as the one presented in this chapter is within the capability to track multiple objects of multiple class types across complex visual scenes.

5.7 Summary

This chapter detailed the investigation and development of an early stage methodology for deriving audio metadata for objects within a simple visual scene. The Google Object Detection API was described as the object detection system used for this pilot study and how it can be adapted to provide basic object detection for video content and the data used to compile an object dictionary from which audio metadata can be derived. Following this, the Jaccard Index, calculated from the object's bounding box data from the current and previous frame, was suggested as a simple method for checking inter-frame object continuity. Additionally, it was suggested as a method for deriving trajectory data from the location data of an object over successive frames. This resulting trajectory data can then be transformed into panning values and transmitted into a DAW, such as Reaper, via OSC. A simple method for suggesting sound effects from the BBCsfx library using string comparison between class labels and audio file metadata was described along with the limitations of this method and how more complex but robust solutions exists through the use of word embedding models.

Results were then presented for each component of the content matching and tracking system. The recommendation of relevant and appropriate candidate sound effects files was limited due to the use of string comparison to identify metadata tags that matches the class labels of predicted objects in the scene. Potential candidate audio files can also be missed due to inconsistent or incomplete tags within the chosen sound effects repository and/or the class labels used by chosen object detector. The results provide evidence that the output from computer vision algorithms can be used to search arbitrary sound effects repositories to return a selection of potential candidate audio files. Recommendations were then given for alternative and more robust search methods, such as the use of word embedding models. Additionally, combining multiple computer vision algorithms, such as object detection, scene detection, and action detection, was suggested as a way to potentially provide a more complete representation of the visual scene which can then be used to suggest both sound effects and ambience files.

In this study, the best performance with respect to spatial position and trajectory tracking was for scenes containing a single object. Whilst correct panning information could be derived for discrete frames containing multiple objects, sensitivity to the differences in the outputted order of object data resulted in the simple continuity check being insufficient to maintain continuity over successive frames for scenes containing multiple objects. The results suggest that there is potential to use object detection algorithms to facilitate audio object positioning and dynamic panning. Computer vision algorithms that are able to maintain object identities over successive frames, however, would provide a more robust system from which to derive object trajectory data that can then be transformed into audio panning data.

Overall, the results of this study indicate that utilising computer vision algorithms to search large-scale, labelled, audio repositories and derive both static and dynamic audio panning data is a valid approach. Potential areas for further investigation would be the use of computer vision algorithms with the ability to provide a more complete representation of the scene, which persists across frames, with review by Jiao et al. [457] presented a comparison of 31 video detection and tracking algorithms which may address the issue of inter-frame continuity. Given the progress and availability of object tracking algorithms designed for 360 video [495, 496] this would also facilitate taking the current approach and extending it for 3D space. However, the availability of open source code to facilitate building upon such video specific systems still appears sparse when compared to traditional single image detection algorithms.

Chapter 6

Predicting time-frequency spatial parameters for use in stereo upmixing using a Residual U-Net

6.1 Introduction

The previous chapter proposed and tested a proof of concept methodology for automatic panning and candidate sound effects suggestion developed using visual object detection. This was in response to some of the findings outlined in Chapter 4. Although the results demonstrate the potential for computer vision to track the trajectory of objects and generate appropriate panning data, further improvement would require a more extensive investigation and utilisation of computer vision techniques. This was, however, deemed to be outside the scope of this thesis. The work in this chapter addresses additional challenges outlined in Chapter 4 relating to both the integration of stereo content within spatial productions and the perceived lack of spatial sound effects libraries. A Multi-Channel U-Net with Residual connections (MuCh-Res-U-Net) is presented, which was trained on a novel dataset of stereo and parametric time-frequency spatial audio data

to predict 360° time-frequency spatial parameters from a stereo input signal. The predicted parametric features can then be applied to a number of different spatial audio encoding methodologies to then upmix stereo content to a 3D representation for both reproduction and storage.

This chapter also presents a dataset of IRs in stereo and Ambisonic format collected to facilitate the training of the propose model. It contains IR for all positions on a 50-point Lebedev quadrature in 9 stereo configurations, 32 capsules from the Eigenmike, and up to 4th order Ambisonics derived from the Eigenmike capsules. Details are given on its collection, availability and use within this thesis.

Finally, this chapter proposes two example stereo upmixing pipelines. The first demonstrates how the predicted spatial parameters can be used as part of parametric audio coding and upmixing methods, such as those proposed in [149] and [497], to upmix stereo signals to arbitrary known multi-loudspeaker configurations. However, unlike traditional stereo upmixing approaches, where the repositioning of direct signal components is limited to the frontal section of a channel-based loudspeaker array, and often only on the horizontal plane [405, 406, 497–501], the directional parameters predicted by the proposed MuCh-Res-U-net will cover 360°, enabling repositioning to any point on a sphere. The second example uses the time-frequency directional features to extract and remap signal components to target spherical harmonic components to facilitate the generation of a full spherical representation of the upmixed sound field.

6.2 Relevant Background

Given the amount of two-channel stereo content that exists when compared to multi-channel content, it may sometimes be desirable to convert or upmix two-channel (low-order) stereo content into a format with a higher order spatial representation [406, 497, 498]. Many of the upmix algorithms in the literature provide channel-based upmixing as they aim to generate additional signals to directly drive additional loudspeakers in a known configuration [405, 406, 498–506], such as 5.1, or by using methods such as (VBAP) [135] to upmix to arbitrary

2D or 3D configurations [499]. Within this context, upmix algorithms can be more simply defined as generating a higher number of output channels from a smaller number of input channels. It is often the case with IMEs that the number of desired audio channels outnumber that of the programme material. For instance, the programme material may consist of a stereo recording where the target system is a 5.1 configuration that requires five full-range signals and one band-limited low frequency signal. Additionally, it is worth noting that even binaurally rendered Ambisonic audio is commonly first rendered to a virtual multi-channel loudspeaker array before being rendered to binaural [69, 75, 161].

Upmixing can be classified into one of two types. The first is upmixing as decoding, where an algorithm upmixes or decodes multi-channel content that has been previously encoded [507]. For instance, Dolby Pro-Logic encoding/decoding can encode 4-channel, 5-channel, and 7-channel surround sound into a two-channel matrix encoded signal that can itself be decoded to retrieve an approximation of the original multi-channel signals [508]. These algorithms are effective as the encoded input signal often contains signal cues such as relative channel phase, which can be used to aid the upmix process. The second, blind upmixing, is where additional channels are generated based solely on analysis of the input signal. As the vast majority of stereo content has not been downmixed from existing multi-channel content, it is the latter type of upmixing algorithm that is of interest in the context of this thesis.

Early upmix algorithms, such as those proposed by Gerzon [509], Dressler [510, 511], Irwan [512, 513], and Usher [514], upmixed (decoded) and/or dowmixed (encoded) using matrix and/or linear filtering operations in the time domain. Passive matrix decoders such as those proposed by Gerzon [509], are signal-independent and require optimisation for different loudspeaker configurations. The signals for the surround loudspeakers are often derived by taking the left, S_L, and right, S_r channels and calculating the difference $S_L - S_R$. However, the difference signals can often still contain direct signal components, especially for those sources panned hard left or right, which may cause distortions to the stereo image. Active matrix decoders, such as those proposed in [510–513],

introduce a steering algorithm to maintain more stable sound source positions compared to passive decoders. The wide-band analysis used, however, is not able to sufficiently differentiate and separate temporally overlapping dominant sources, which, again, can introduce spatial distortions. More recent approaches tend to favour processing in the time-frequency domain [497, 498, 500, 506, 515–520], where frequency components are extracted and remapped from the original signal to the target channel configuration [500, 506] and it is these methods which are of particular interest in this thesis.

Avendano and Jot proposed two classes within which upmix algorithms could be classified [498]:

- Multi-channel converters, which derive additional channels (e.g. for a centre loudspeaker) with the aim of increasing the listening area while preserving the stereo image.
- Ambience generation, which aim to extract and/or synthesise the ambient component of a recording to be reproduced by the surround channels.

As noted by Kraft [497], most upmix algorithms use a combination of both approaches: the input signal is first analysed and decomposed into its direct and diffuse components, with diffuse component usually being decorrelated and sent to the surround (rear and overhead) loudspeakers. The direct component is processed using the multi-channel converter approach to create the required number of output channels to drive all frontal loudspeakers.

6.2.1 Stereo Signal Model

Many stereo upmixing methods decompose a stereo signal into direct signal components and diffuse signal components, the details of which are explored in Section 6.2.2. However, firstly, a signal model is defined similar to that found in the literature [497, 500]:

$$x_i(t) = \sum_{j=1}^{N} \bar{s}_j(t) * \vec{d}_{ij}(t) + \sum_{j=1}^{N} \bar{s}_j(t) * d_{ij}^{\psi}(t) + n_i(t) \tag{6.1}$$

where x_i are the channel signals that result from the summation of the weighted summation of sources $\bar{s}_j$ convolved with the direct signal component $\vec{d}_{ij}$ and the weighted summation of sources convolved with the diffuse signal component d^{ψ}_{ij} and where n_i are background noise signals contributing to the ambience. $\vec{d}_{ij}(t)$ and $d^{\psi}_{ij}(t)$ are also components of a room/system impulse response such that:

$$h_{ij}(t) = \vec{d}_{ij}(t) + d^{\psi}_{ij}(t) \tag{6.2}$$

which simplifies equation 6.1 to be identical to Equation 2.35 such that:

$$x_i(t) = \sum_{j=1}^{N} \bar{s}_j(t) * \vec{d}_{ij}(t) + \sum_{j=1}^{N} \bar{s}_j(t) * d^{\psi}_{ij}(t) + n_i(t) \tag{6.3}$$

$$x_i(t) = \sum_{j=1}^{N} \bar{s}_j(t) * h_{ij}(t) + n_i(t) \tag{6.4}$$

are equivalent.

6.2.2 Direct-Diffuse Decomposition

A common approach for many recent upmix systems is to transform the signal into a time-frequency representation through techniques such as the STFT. The time-frequency signal is then decomposed into its estimated direct and diffuse components [407, 498, 500, 505, 517], sometimes also referred to as the primary and ambient components respectively, enabling more effective separation of temporally overlapping sources. Direct components are defined as those signal components which are highly correlated with existing channels and diffuse components are those signal components which have low correlation with the existing channels [521]. For a detailed review of existing direct and ambient decomposition methods see [411]. It is also generally assumed that only a single dominant source is active within each time-frequency tile and that the direct signal power is greater than the ambient signal power [149, 411, 497]. Within the context of the signal model defined in equation 6.1 this can be defined as in [497]:

$$|S_u(m,\omega)| \gg \sum_{\forall \neq u} |\bar{S}_i(m,\omega)| \tag{6.5}$$

where S_u is a single dominant source for the time-frequency tile centred at time instant m and in frequency band ω and where $\bar{S}_i$ is all other sources present.

Several approaches have been proposed for direct-diffuse decomposition of stereo signals. Goodwin and Jot [506] proposed the use of Principle Component Analysis (PCA) where the primary (direct) component for each channel is estimated as the projection of the channel signal onto the principal vector derived from the largest eigenvalue, while the ambient components are assumed to be the residuals showing low correlation. Vickers [522] proposed a similar geometric decomposition utilised for center channel extraction. Goodwin and Jot [506] note, however, that if the primary component does not have substantially more energy than the ambient component, an amount of the ambient component can remain present in the principal PCA component. This can cause erroneous directional analysis of the direct components and suboptimal rendering of the diffuse components. While decomposition through traditional PCA approaches utilise intensity differences between the channels, it does not take advantage of any time differences that may be present between the signals. He [411] proposed a PCA based approach that analyses the time difference between the two channels to aid in the decomposition. Ibrahim and Allam [523] also propose the use of a weighting factor to estimate the presence of the dominant primary source to improve the accuracy of ambient source separation.

Avendano [498, 500, 505] proposed a spectral method which calculates the short-time coherence using cross-correlation and auto-correlation of the two stereo channels to derive estimates of time-frequency panning and ambience indexes. These indexes are then used to derive a time-frequency mask to extract the direct and diffuse components. Faller [517] also proposed a similar time-frequency approach but utilised a least-squares estimate to extract the direct and diffuse components by minimising the error between the extracted signal and a stereo signal model. A method based on subband mid-side decomposition was used by Kraft [404] to estimate azimuth directions which were then used to separate direct and diffuse signal components.

6.2.3 Directional Estimation

For stereo signals, directional estimates for direct components are often based on estimated panning coefficients $\hat{\alpha}$, which are derived from channel and inter-channel comparisons such as cross-correlations and auto-correlations [497]. The estimated panning coefficients are then used to calculate an estimated panning index $\hat{\Psi}$ which indicates the position of a signal component between the limits of the stereo field; for instance, $\hat{\Psi} = 0$ would signify a centre panned source, $\hat{\Psi}$ = -1 would signify a source panned hard left, and $\hat{\Psi} = 1$ would signify a source panned hard right. An approximated source angle is then estimated using the obtained panning index [497, 498, 517–519]. Kraft [497] conducted a comparison of the panning estimation approaches in [498, 517–519], which found that once the different approaches were unified under a common notation scheme the panning coefficient estimates could be simplified to:

$$\hat{\alpha} = \frac{\hat{\alpha_R}}{\hat{\alpha_L}} = \sqrt{\frac{r^0_{RR}}{r^0_{LL}}} \tag{6.6}$$

and the positional index is then simplified to:

$$\hat{\Psi} = \frac{\hat{\alpha} - 1}{\hat{\alpha} + 1} = \frac{\sqrt{r^0_{RR}} - \sqrt{r^0_{LL}}}{\sqrt{r^0_{RR}} + \sqrt{r^0_{LL}}} \tag{6.7}$$

where r^0_{XX} is the power of the respective channel averaged over time-frames. In [404] Kraft further simplifies this as:

$$\hat{\Psi} = \frac{\hat{\alpha} - 1}{\hat{\alpha} + 1} = \frac{|X_R| - |X_L|}{|X_L| + |X_R|} \tag{6.8}$$

where X_L and X_R are the STFT magnitude spectrum of the left and right channels respectively. The reader is directed to Appendix A in [497] for the full comparisons and derivations.

As the panning index is estimated based on the panning law of sines [127], it represents an approximate lateral position of a source between two loudspeakers based on the inter-channel amplitude differences. From this index an estimated source angle can be approximated using the angular position of the loudspeakers. The relationship between $\hat{\Psi}$ and estimated source angle $\hat{\theta}$ has been shown to be

generally linear for typical two-channel stereo configurations of -45° $\leq \theta_0 \leq -30°$ and be calculated as in [497]:

$$\hat{\theta} = |\theta_0|\hat{\Psi} \tag{6.9}$$

with source angle being approximated using:

$$\hat{\theta} = \arcsin(-\sin(\theta_0)\hat{\Psi}) \tag{6.10}$$

The panning coefficients and positional indexes can then be used to remap the components to the new target array using the chosen panning method.

6.2.4 Existing Tools

At the time of writing, there are a limited selection of plug-ins offering stereo to Ambisonic upmixing available from companies such as, Blue Ripple [524], Nugen [525], and Penteo [526] with Cardew [527] having released a freely available plugin designed for horizontal only upmixing of existing stereophonic music recordings. However, only Cardew provides in depth details on the algorithms used, whilst the other plug-ins operate as "black boxes", which is understandable given they are commercial products. Blue ripple does, however, state that their upmixing algorithm is designed for use with material that has been mixed using conventional panning methods and is more akin to presenting the stereo material as a stage that can be moved around the spatial scene, with an option spread the left and right channels to varying degrees over the whole 3D space. It appears, at present, there are no algorithms that enable the synthesis of B-format signals from a stereo recording, that would be comparable to those generated had a multi-channel or spherical microphone array been used to capture the original source material.

6.2.5 Limitations of current approaches

There are, however, some limitations to the current approaches for stereo upmixing, particularly around the directional estimation of components. Firstly, many of the algorithms in the literature are developed and tested using synthetic

stereo material consisting of individually recorded sources positioned within the mix using amplitude panning [411, 497, 498], or where live recordings have been used these are normally using coincident microphones pairs, which again encode the scene using predominately inter-channel amplitude differences and though not explicitly stated, the sources of interest would most likely have been placed in front of the microphone array. In this context we define the front as being the direction the capsules are facing, and within the context of a bidirectional polar patter it is the direction of positive polarity. In some cases, information about how the material was captured is not provided [521]. As detailed in Section 2.6.6, stereo signals traditionally only account for a source's lateral position, providing insufficient information for traditional methods to discern its elevation or whether it is positioned in front or behind the array. It is the practice of stereo signals being replayed over frontally placed loudspeakers that introduces a conceptual *front* and *back*. Conceptually, this seems reasonable, as although a stereo signal can be viewed as a lateral representation of a sound field, it is reasonable to assume that the microphone array would be pointed towards the sources of interest, and when reproduced, the listener would be orientated towards the reproduction system, thus creating a frontal representation of a given sound field.

Upmixers aim to enhance this representation by generating ambience around the listener that seeks to simulate the reflections and reverberation of the recorded or synthesised environment [503]. They in effect create a frontally focused sound field with additional surrounding ambience, which is generally adequate for traditional screen based media where the action will be coming from the front and therefore the attention of the audience will be directed towards the front. This approach, however, introduces challenges for stereo signals recorded in real environments as sources may be located at varying positions on both the median and horizontal planes. Consider the example presented in Figure 6.1, which shows a spaced stereo pair with 4 loudspeakers positioned at azimuth, $\theta = [45°, 135°, 225°, 315°]$. A sound is played from each speaker sequentially, starting with the speaker at 45° and continuing in an anti-clockwise direction.

Applying traditional methods of panning estimation would yield near identical values for the sources at 45° and 135° as well as identical values for those positioned at 225° and 315°. The identical value pairs are a result of traditional stereo localisation estimation methods being limited to lateral position, usually based on either TDOA between the two microphone signals or the inter-channel amplitude difference. This is a similar principle to that explored in Section 2.4.3, where it is possible for each set of inter-signal differences, when based purely on time and level difference, to exist for multiple locations.

Subsequently, were these signals to be upmixed using systems such as those proposed in [404, 405, 498, 518] and reproduced over a 5.1 configuration the perception of source movement around the array would not be congruent with that observed during the recording. Instead, the direct components of the two source positions at 45° and 135° would be reproduced at the front left of the array and the two sources at 225° and 315° reproduced at the front right, while each time the surround speakers would predominately contain the decorrelated diffuse component.

The aim of this work is to develop a deep learning approach where, given appropriate input features containing time, amplitude, and phase information, a NN can be trained to approximate a mapping function that predicts spatial features for a 360° space from the information contained within and derived from a stereo signal. These spatial features can then be used to facilitate upmixing methods that move away from frontally biased systems to ones that aim to reproduce a sound field that approximates the spatial characteristics that would have been present at the time of recording.

6.2.6 Machine Learning Approaches

A number of machine learning approaches to upmxing have been presented in recent years, although they have predominately focused on channel-based methods. Ibrahim and Allam [521] approach the task of direct-diffuse composition as a classification problem, training a feed forward NN to classify each complex valued time-frequency tile as either direct and diffuse. When used as part of an upmixing

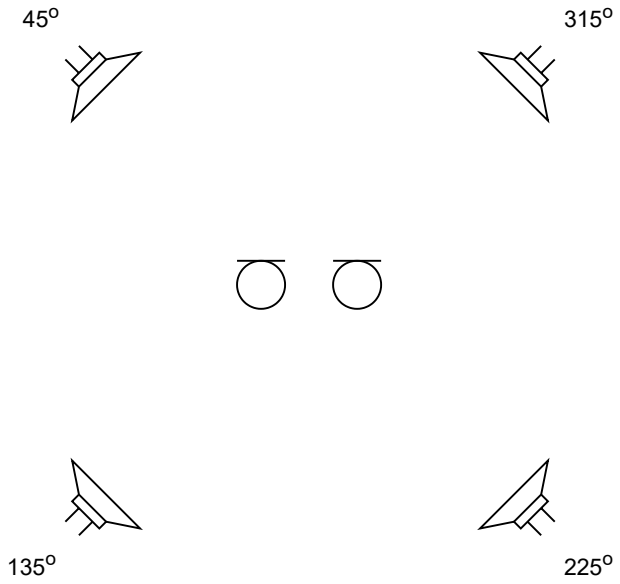


Figure 6.1: Spaced pair capturing sources from locations 45° 135° 315° and 225°. This illustrates the frontally biased nature of traditional stereo upmixing systems as the direct components for sources at 45° and 135° would both be reproduced out of the front left speaker and direct components for sources at 315° and 22°5 both replayed out of position 315°

system to upmix from stereo to a quad array, 10 out of the 11 listeners preferred the NN method above traditional methods such as those proposed in [506] and [498], as well as achieving the highest signal to distortion ratio which was tested on each of the extracted direct and ambient components. Park et. al. [405] proposed a deep neural network (DNN) to upmix from stereo to 5.1 within the MPEG-H 3D framework [120]. A DNN was trained using log-spectral magnitudes of quadrature mirror filter subbands to predict the center and surround channels from the input stereo signals. The input signals are then mapped in the subband space to the center and surround channels where they are transformed back into audio signals via quadrature mirror filter synthesis. The approach is based on the assumption that the center channel is some combination of the left and right channels and the surround channels are derived as some amount of the difference between two channels. The method proposed in [406] uses two DNNs where one is trained to perform direct-diffuse decomposition and the other then

renders the diffuse component. Both networks are trained to jointly minimise the Mean Squared Eerror (MSE) between the magnitude spectra of the original and the upmixed/decoded five channel signal as well as minimising the loss for the ICLD. The network predicts spectral weights which are then multiplied with each frequency bin in the stereo signal and acts as a mask to separate the direct and diffuse components. In all cases, the current methods are concerned with deriving signals to directly drive additional loudspeakers for use within channel-based upmixing.

6.3 Dataset

6.3.1 Existing Datasets

An investigation into ML driven upmixing of stereo signals requires a dataset that contains the relevant input-output pairs with which to train and evaluate the model. In the case of this thesis, a dataset containing equivalent stereo and Ambisonic signals was desired. Such a dataset would facilitate the training of a model to approximate the mapping function from a given stereo scene to the 360° time-frequency spatial parameters for that scene, with the target time-frequency spatial parameters being derived from the Ambisonic signals.

Given the increase in both the interest in ML applied to audio signal processing and the use spatial audio within IMEs, there are a number of existing open source spatial audio datasets that are available. For instance, a number of spatial audio datasets have been used and released as part of the Detection and Classification of Acoustic Scenes and Events (DCASE) challenges [430]. The first DCASE challenge used a dataset consisting of 24 recordings of individual sounds, alongside 14 ≈1-minute scripted sequences [430]. The recordings were captured in an Ambisonic format, although only the stereo mix downs were publicly released. More recent DCASE challenges, specifically the sound event localisation and event detection (SELD) task, have utilised datasets consisting of synthesised scenes in both FOA and four-channel tetrahedral array format [14, 528, 529], whilst the 2022 challenge introduced a dataset containing recordings of real scenes [530]. A comparison

	DCASE2019	DCASE2020	DCASE2021
# rooms	5	13	13
# spatial RIRs/positions	504 discrete positions	≈200 spatial trajectories (continuously captured SRIRs)	≈200 spatial trajectories (continuously captured SRIRs)
Source-to-receiver distances	1m-2m	1m-5m	1m-5m
Spatial ambient noise	30dB SNR	6-30dB SNR	6-30dB SNR
Moving sources	No	Yes	Yes
Non-target interfering events	No	Yes	Yes
# polyphony/overlapping events	≤2	≤2	≤3 (+ ≤1 interf. event)
% same-class overlapping events	low	low	high
# target classes	11	14	12
# event samples	220	≈700	≈500 (target events) ≈400 (interferer events)

Table 6.1: Comparison of DCASE SELD datasets. Taken from [531].

of the DCASE datasets is shown in Table 6.1. Each of the datasets has been rigorously produced encompassing different environments, a variety of different sound events, varying levels of background noise and polyphony, and a high spatial sampling density. The DCASE datasets therefore provide ideal data for many tasks associated with machine listening and spatial audio processing, two such examples being SELD and multi-channel source separation.

Green and Murphy presented Eigenscape [532], a dataset of soundscape recordings consisting of 64 x 10 minute recordings evenly distributed over 8 classes of soundscape and available up to 4th order Ambisonics. This results in 80 minutes of audio data for each class and a total dataset length of just under 11 hours. Eigenscape has been predominately used to investigate auditory scene classification and has shown to be a suitable dataset for training both classical ML algorithms, such as Gaussian Mixture Models [115], and DNNs, such as CNNs [533], using both FOA and HOA signals. FOA data has also been collected from YouTube to train a self-supervised audiovisual model for aligning spatial video and audio clips extracted from different viewing angles [410]. As part of the Learning 3D Audio Sources 2021 (L3DAS21) Challenge [534], two datasets were released for both 3D speech enhancement and 3D SELD. The datasets consisted of FOA recordings that were synthesised by convolving monophonic sources with IRs taken at 252 positions (168 on a fixed point grid and 84 from positions within

a 3D uniform random distribution) within a large office room [534].

There are also a number of datasets consisting of spatial IRs, which one can then use to synthesize scenes through the convolution of the IRs with monophonic and anechoic source material. The IRs used as part of the dataset synthesis in [14] were also released seperately as part of the TAU Spatial Room Impulse Response Database (TAU-SRIR DB) [535] and consist of IRs captured in 9 different rooms with SRIRs being extracted from noise recordings of sources moving slowly across specified trajectories at intervals of $\approx 1^{\circ}$ from the microphone. This does mean that the exact SRIR directions differ slightly with each room, however, this increases the ease with which moving sources can be emulated. Lübeck, Arend, and Pörschmann presented a high-resolution SRIR dataset in [536], which was captured using VariSphear [537], an automated single-microphone measurement system. The Varisphear systems comes with a spherical microphone array extension, which is a rigid spherical baffle that houses the measurement mic and rotates about its axis to sample the sphere using a given grid configuration. This facilitates the capture of high resolution spherical microphone array IRs, whilst using only a single microphone. IRs were captured for 2702 sampling positions using a 44 point Lebedev grid spherical microphone array configuration [538]. BRIRs were also captured using a Neumann KU100 dummy head for 360 directions along a horizontal circle in 1° steps. Furthermore, given that the IRs captured are equivalent to those captured using spherical microphone arrays, Ambisonic signals may also be derived through additional processing.

It is suggested by Cobos et al. [539], that one of the limiting factors in developing deep learning derived spatial audio methods is the lack of sufficiently large multichannel audio datasets that would adequately facilitate the training of DNNs. It is clear from this non-exhaustive review that there exists, at present, a selection of spatial audio datasets available that are both of high quality and open source that might contribute towards this challenge. The ones detailed in this section contain both real and synthesized scenes, as well as spatial IRs, from which one could synthesize new scenes. All the datasets discussed in this section, however, lack the property of containing equivalent stereo and Ambisonic data.

The only dataset that the author was able to find, possessing both spatial and stereo data, has been presented by Gao and Grauman in [540], and contained both binaural and stereo recordings captured simultaneously by mounting a GoPro HERO6 Black [541], which records stereo, on top of a 3Dio binaural microphone [542]. At the time of writing, and to the author's knowledge, there are currently no open source datasets that feature recorded or synthesised sound scenes in both Ambisonic and stereo formats, where the stereo signals have not been derived from the Ambisonic signals. Whilst coincident stereo signals can be derived from Ambisonic recordings, and there is ample evidence of this derivation being used in studies requiring both Ambisonic and stereo signals [534, 543, 544], the resulting signals would differ from those captured with real stereo microphone techniques (detailed in Section 2.6.3) in that the two Ambisonically derived signals would be in theory, perfectly coincident, something which is physically impossible given real microphone arrays as it would require them to occupy exactly the same location in physical space. It was therefore decided to synthesise a novel dataset of equivalent stereo and Ambisonic sound scenes. Additionally, given the complexity of the intended task, it would be beneficial for initial algorithm development to be conducted with simple spatial scenes whose attributes, such as number of objects, polyphony, and level of background noise could be controlled and quantified, similar to the datasets presented in [14, 528–530]. This level of control over the training data would then enable further testing and development on more complex scenes as appropriate. From this portion of the work two distinct datasets are produced; the stereo and Ambisonic IRs and the sound scenes which are synthesised using the IRs and an ambient recording. The target features and input features are then derived from the resulting sound scenes.

6.3.2 Dataset Formats

The IR dataset consists of two-channel stereo IRs for 9 stereo configurations, the 32 channels captured from an Eigenmike, and spherical harmonic components up to 4th order derived from the Eigenmike signals using the Eigenunits plugin [165]. Details of the microphone configurations are shown in Table 6.2. The stereo IRs

encode spatial information differently depending on the stereo configuration used. All stereo pairs will contain, to some degree, time of arrival differences due to the spacing between capsules. This holds true even for coincident techniques such as XY and Blumlein since the capsules are physically unable to occupy exactly the same point in space. There will also be ICLD due to the physical space between capsules, the angle of incidence between the source and the two capsules when using directional pickup patterns, and the angle and distance between the two capsules. The Blumlein pair will also encode the directional of arrival with phase information/differences due to the bi-directional pickup pattern of each microphone.

The naming convention for the IR is:

- {Mic_config}_IR_{loudspeakerNumber}_azi_{position}_el_{position}.wav

with the scene naming convention being:

- fold{num}_{Mic_config}_mix_{mix_num}.wav

6.3.3 Sound Events

Sound event samples used were from the NIGENS General Sounds Events Database [545], as also used in [14]. The Database contains 714 sound samples distributed across 14 classes of, *alarm, barking dog, burning fire, crash, crying baby, female and male scream, female and male speech, footsteps, knocking on door, piano, ringing phone, and running engine.* It contains an additional 303 samples within the *general* class which are any sounds that do not fit into the previously mentioned 14 classes, giving a total of 1017 audio files. The samples vary in length between 1s and 5 minutes with most being sampled at 44.1 kHz and 32-bit precision. The dataset was originally curated for use within computational auditory scene analysis tasks, such as sound event detection and classification. As such, it 8 pre-defined splits which divides up the dataset into 8 folds of equal size.

The dataset consists of sounds compiled from other open source libraries. *Speech* samples are taken from the GRID [546] copora, *scream* samples from

freesound.org [547], *general* samples were taken from both freesound.org and StockMusic, with all other classes being attained from StockMusic. For further details on the recordings and the database the reader is referred to [545].

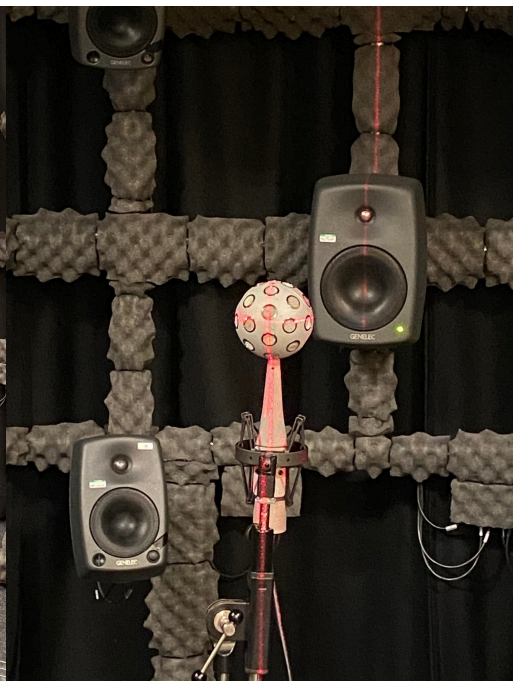
6.3.4 Impulse Response Specification and Acquisition

The dataset was captured using the mh Acoustics Eigenmike em32 [165], a spherical microphone array with 32 captures arranged on a rigid sphere and capable of up to fourth-order Ambisonic capture, and the selection of microphones listed in Table 6.2 were used to capture a variety of stereo configurations including spaced, coincident, and near-coincident. All IRs were measured using 10 second exponential sine sweep following the method proposed in [106]. Genelec 8040A's captured all positions at 0° 90° and -90° elevation whilst all other positions were captured using 8030Bs. All sweeps were played back at a peak A-weighted amplitude of approximately 80 dB. The Eigenmike IRs were recording using the proprietary Firewire interface while the stereo configurations were recorded using a Presonus DigiMax DP88 microphone pre-amplifier. All microphone arrays were aligned in the centre of the loudspeaker using laser level meters, with spaced and coincident pairs then having their distance and orientation with respect to each other set using a stereo microphone bar. All IRs were captured and extracted at 24-bit resolution and 48 kHz sampling rate. For IR capture the heavy theatre curtains were drawn around the rig to limit interference from reflections. Due to the location of the measurement rig there was a risk of interference from noise sources in and outside of the building, therefore sweep recordings were repeated if necessary. IRs were generated through the 'deconvolution' of the measured sweeps with the inverse of the original sweep, as proposed in [106]. Figure 6.2 shows an example set up for IR capture with a spaced pair, a Blumlein pair, and an Eigenmike.

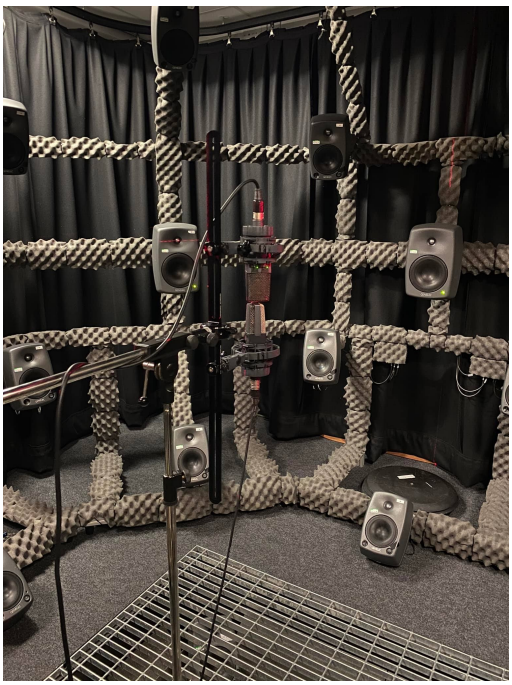
6.3.5 Spherical Harmonic IR Encoding

The raw IRs from the Eigenmike were also converted into spherical harmonics up to fourth-order using the EigenUnits® software tool described in [548]. Although

(a) Spaced pair



(b) Eigenmic



(c) Blumlein pair

Figure 6.2: Microphone configurations set up for IR capture and positoned using laser level meters.

Set Name	Array Configuration	Microphone/s used	Microphone Directivity Pattern	Spacing	Orientation angle
AB_Omni_30	AB Pair	AKG C414 XLS	Omnidirectional	30 cm	Parallel
AB_Omni_40	AB Pair	AKG C414 XLS	Omnidirectional	40 cm	Parallel
AB_Cardioid_30	AB Pair	AKG C414 XLS	Cardioid	30 cm	Parallel
AB_Cardioid_40	AB Pair	AKG C414 XLS	Cardioid	40 cm	Parallel
Blumlein	Blumlein	AKG C414 XLS	Bidirectional	Coincident	90°
DIN	DIN	Rode NT5	Cardioid	20 cm near coincident	90°
NOS	NOS	Rode NT5	Cardioid	30 cm near coincident	110° near coincident
ORTF	ORTF	Rode NT5	Cardioid	17cm near coincident	90°
Eigen_SPH	Rigid Spherical Baffle	Eigenmike	up to 4th Order Spherical Harmonics	8.4cm diameter spherical array	
Eigen_raw	Rigid Spherical Baffle	Eigenmike	omnidirectional	8.4cm diameter spherical array	
Coincident	XY	Rode NT4	Cardioid	Coincident	90°

Table 6.2: Details of IR sets captured including configuration, spacing, capsule angle, and microphone used.

the spatial aliasing frequency for the Eigenmike is stated as approximate 9 kHz in the official documentation, analysis by McKenzie [549] found spatial aliasing occurs above approximate 5.1 kHz. Due to the physical limitations of the array configurations 2nd-4th order components are also by default highpass filtered with cut-off frequencies set to 400 Hz, 1 kHz, and 1.8 kHz respectively. The spherical harmonic IRs follow the ACN and SN3D conventions detailed in Section 2.6.5; referred to collectively as the AmbiX format.

6.3.6 Dataset Availability

The dataset of IRs is available under a Creative Commons license downloadable as .wav files. The dataset contains the IRs for 9 stereo configurations, raw capsule records from the Eigenmike and the encoded spherical harmonic conversions [550].

6.3.7 Sound Scene Synthesis

Sound scene synthesis follows a similar procedure proposed in [14], but is adapted to generate scenes in both Ambisonic and stereo format and is explained here for completeness. The procedure is the same for both Ambisonic and stereo scene synthesis. The sampling rate of the synthesised scenes is user defined and in this

instance was set to 44.1 kHz. As discussed in Section 6.3.3, the NIGENS database provides 8 pre-determined splits, splits 1-6 were for the creation of the training set, split 7 for the validation set, and split 8 for the test set. This ensures none of the sound samples used to validate or test the model are used during training, which reduces the possibility of data leakage across the training and evaluation. The training strategy is discussed in detail in Section 6.6. The onset of each sound event within the scenes were randomly distributed but adhered to the specified level of polyphony which could range between one to five. All sound events locations are static and were spatialised by convolution with the respective IRs for their randomly assigned DoA from the available 50 positions with IRs being resampled if required. Overlapping sound events have a user defined minimum angular distance between them. Finally, a random portion of the two-minute ambient recording was selected and added to the synthesised sound scene at a specified signal-to-noise ratio, in this case 30dB. All scenes generated were 7 seconds in length and a total of 6000 unique scenes were generated. This resulted in a dataset comprising of 11 hours and 36 minutes of sound material. Figures 6.3 and 6.4 show the log-magnitude spectrograms of a scene synthesised for the training set in both stereo and Ambisonic format respectively.

6.3.8 Target Feature Extraction using Directional Audio Coding Analysis

The desired target features are a diffuseness index and a direction of arrival for each time-frequency component in spherical coordinates, azimuth θ and elevation ϕ, in radians. These are common features often used in traditional upmixing systems to extract, reposition, and render the direct and diffuse components of a given signal. As the aim of the network is to predict these features within a 360° space, target features were extracted from the synthesised Ambisonic scenes using Directional Audio Coding (DirAC) analysis [149]. Developed from an existing method for impulse response reproduction [173], DirAC was developed as a flexible, perceptually motivated method of parametric spatial sound representation, reproduction, and transmission [551] that could be used to re-

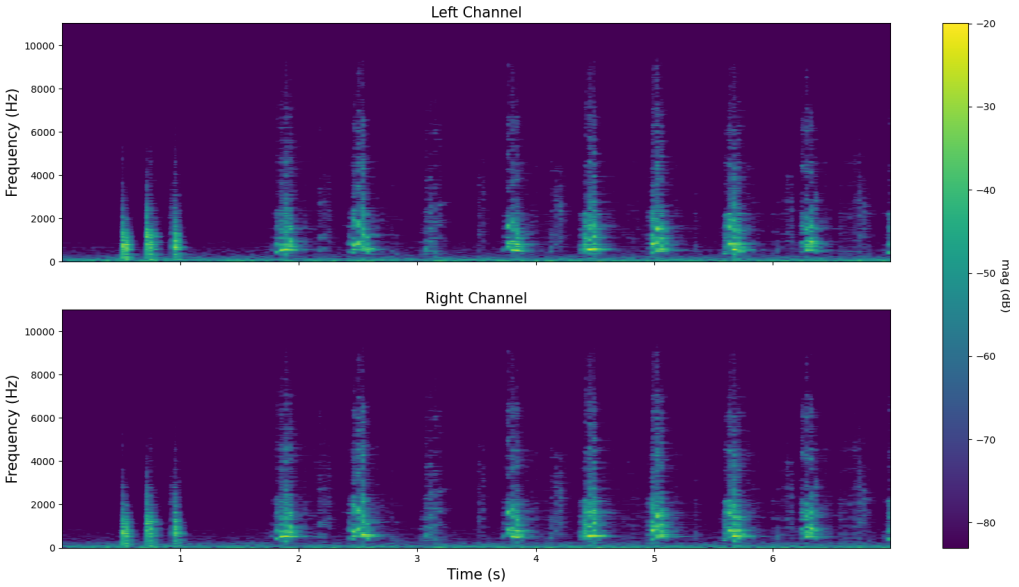


Figure 6.3: Log magnitude spectra extracted from a stereo scene synthesised using the methodology outlined in Section 6.3.7

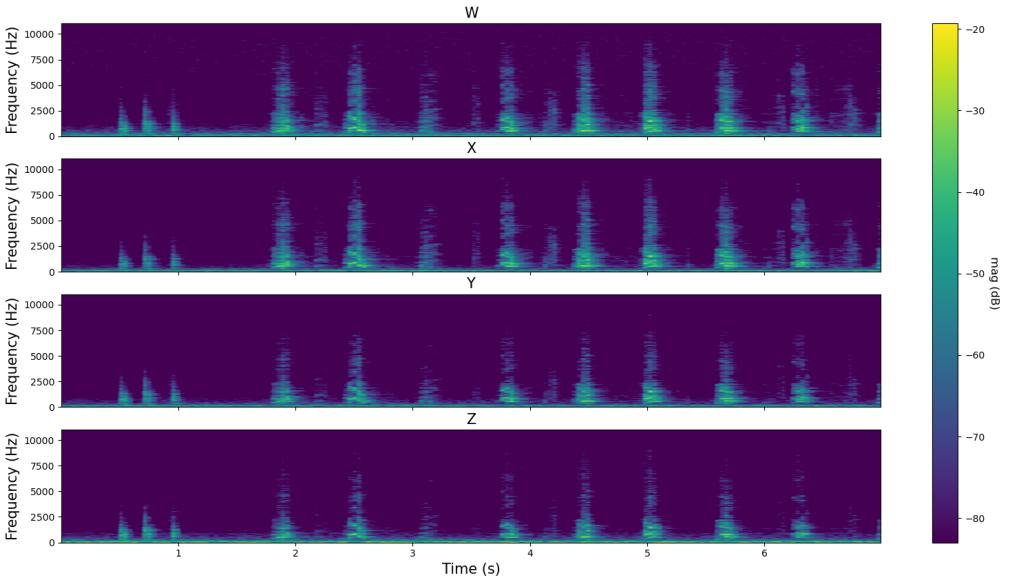


Figure 6.4: Log magnitude spectra extracted from a B-format scene synthesised using the methodology outlined in Section 6.3.7

produce 2D and 3D sound fields using arbitrary audio reproduction methods [150]. Applications of DirAC include improved spatial reproduction of B-format signals using time-frequency parametric direct and diffuse rendering [149, 150], enabling reduced data rate transmission for telecommunication systems by only transmitting the omnidirectional pressure component (W channel) along with the necessary spatial metadata for reconstruction [151], and the development of parametric spatial audio effects based on the manipulation of parameters derived from DirAC analysis [551, 552]. Spatial features derived from DirAC analysis have also been used to successfully train sound scene classifiers [11, 532].

The design of DirAC is based on a number of assumptions about the interaction between a sound field and human spatial sound perception, for a full review see [149]. The most relevant to this work is the assumption that at any one time instant humans can only decode single cues within each critical frequency band from the summed signals received from the ear canals, an assumption supported by evidence presented in [553]. Simply put, this means that for each time-frequency tile there will be a single perceptually dominant cue for each parameter. These parameters are the DOA of the incident sound energy and the diffuseness/inter-aural coherence index and will determine how a listener perceives the spatial impression of the given sound field. As such, DirAC analysis derives a single parameter value for each time-frequency tile for each parameter.

For parameter derivation, the B-format signals were first transformed into the time-frequency domain using the STFT method as detailed in Section 2.5.5. The STFT was calculated using a non-symmetric Hann window with a length of 1024 and a hopsize of 512. This corresponds to a window length of duration 23 ms at a sampling frequency of 44.1 kHz with 11.5 ms between successive frames. The non-symmetric nature of the window ensured COLA compliance as also detailed in section 2.5.5. It is important to ensure that window length is sufficient in length for correct low-frequency analysis but also that it is able adequately capture sound events that are very short or transient in nature.

Directional analysis utilising B-format signals is performed as per [149] and [551], using an energetic analysis of the sound field based on the STFT domain

representations of the sound pressure $P(m, \omega_k)$ and particle velocity $\vec{U}(m.\omega_k)$ at the recording position, where m, ω_k are time and frequency indices respectively. The W channel signal is regarded as proportional to the sound pressure, while the three orthogonal pressure gradient signals X, Y, and Z capture signal properties considered to be proportional to sound velocity. This gives the relationship [551]:

$$P(m, \omega_k) = W(m, \omega_k) \tag{6.11}$$

$$\vec{U}(m, \omega_k) = -\frac{1}{\sqrt{2}Z_0}\vec{X}'(m, \omega_k) \tag{6.12}$$

where $\vec{X}'(m, \omega_k) = [X(m, \omega_k), Y(m, \omega_k), Z(m, \omega_k)]^T$ is the vector of B-format pressure gradient signals and Z_0 is the characteristic impedance of air. The 3-dimensional instantaneous intensity vector is an estimate of the direction of the net flow of energy and is calculated for each frame m and frequency bin ω_k as:

$$\vec{I}(m, \omega_k) = \Re\{\mathrm{E}\{P^*(m, \omega_k)\vec{U}(m, \omega_k)\}\} \tag{6.13}$$

where * represents the complex conjugate of a complex number and $\mathrm{E}\{\cdot\}$ is a short time averaging operation which for an un-averaged intensity vector $\vec{I}_{raw}$ can be expressed as:

$$\mathrm{E}\{\vec{I}(m, \omega_k)\} = \varepsilon\vec{I}_{raw}(m, \omega_k) + (1 - \varepsilon)\mathrm{E}\{\vec{I}_{raw}(m - 1, \omega_k)\} \tag{6.14}$$

where $\varepsilon \in [0, 1]$ is the time-constant in seconds of the exponentially decaying estimation window:

$$T = \frac{1}{\varepsilon fs} \tag{6.15}$$

where fs is the STFT sampling frequency.

As the intensity vector is said to point in the direction of the net flow of energy, the direction of incidence is defined to be the opposite direction of the intensity vector and points towards the source [149]. This can simply be defined as:

$$\vec{D}(m, \omega_k) = -\frac{\vec{I}(m, \omega_k)}{||\vec{I}(m, \omega_k)||} \tag{6.16}$$

The resulting matrix $\vec{D}$ contains time-averaged directional of arrival (DOA) estimates for each time-frequency tile. The desired azimuth and elevation angles

in radians can be derived from this as follows [115]:

$$\theta = arctan\left(\frac{I_3}{I_1}\right) \tag{6.17}$$

$$\phi = arccos\left(\frac{I_2}{||\vec{I}||}\right) \tag{6.18}$$

where I_1, I_2, and I_3 are the first-order channel matrices contained within $\vec{I}$.

The diffuseness index is estimated in the STFT domain as [517]:

$$\psi(m,\omega_k) = 1 - \frac{\sqrt{2}||\Re\{\mathrm{E}\{P^*(m,\omega_k)\vec{U}(m,\omega_k)\}\}||}{|\mathrm{E}\{P^*(m,\omega_k)\}|^2 + ||\mathrm{E}\{\vec{U}(m,\omega_k)\}||^2} \tag{6.19}$$

where a value of $\psi = 0$ indicates the net flow of energy from a given time-frequency tile corresponds to the total energy within that time-frequency tile. A value of $\psi = 1$ indicates there is no net transfer of acoustic energy within that time-frequency tile and thus indicates a completely diffuse sound field.

Lastly, the short-time averaged energy vector can be derived as in [554]:

$$\vec{E}(m,\omega_k) = |\mathrm{E}\{P^*(m,\omega_k)\}|^2 + ||\mathrm{E}\{\vec{U}(m,\omega_k)\}||^2 \tag{6.20}$$

Initially, features were derived from sound scenes sampled at 44.1 kHz, which, with the previously detailed STFT parameters, yielded a frequency resolution of approximately 43 Hz and a temporal resolution of 23 ms and resulted in a matrix of size 604 x 513 x 4 for a single training example. Due to memory constraints it was decided to derive features from sound scenes resampled to 22.05 Hz. Keeping the same window length to maintain the absolute number of frequency bins resulted in a matrix of size 303 x 513 x 4. The window length is now of duration 46 ms with the start of each successive window being separated by 23ms. The frequency resolution is now approximately 21.5 Hz with a lowest detectable frequency of approximately 100 Hz.

6.4 Input Features

The selection of appropriate input features is an important consideration when designing any neural network (NN) system. NNs are often referred to as universal

function approximators, this does mean, however, that assuming an appropriate network architecture, there must also exist some mathematical function capable of deriving the desired target from the information available within the input features. If the input features are ill-conditioned for the chosen problem and do not contain the requisite information, then the network will be unable to approximate the desired mapping function. As the aim of the network is to predict multiple target features it is important that the selected input features contain appropriate information that can be applied to the prediction of each target feature.

6.4.1 Pre-processing

Prior to feature extraction the raw audio signals are normalised to zero-mean and unit-variance such that:

$$X_{\text{stereoNorm}} = \frac{X_{\text{stereo}} - mean(X_{\text{stereo}})}{std(X_{stereo})} \tag{6.21}$$

The channel signals must be normalised as a stereo pair to maintain their inter-channel relationship. This normalisation method centers the data around zero which has been shown to be beneficial for training NNs [555]. For networks that utilise multiple types of input features, which may be quite different with respect to numeric scale, normalisation ensures that all features are within a similar numeric range and that the values are not too large when compared to the networks initial weight values. Input features with large and/or heterogeneous values can potentially cause large gradient updates which can at best slow down convergence or at worst prevent convergence and cause the network to become unstable [196].

6.4.2 Short-time log-magnitude spectrum

Each of the stereo channels are transformed into a time-frequency representation using the STFT with identical parameters to those used during the target feature extraction explained in Section 6.3.8. A logarithmic function is then applied

to retrieve the log-magnitude spectrum and this is typically done as another form of normalisation. Once the log-magnitude spectrum has been obtained, zero-mean unit-variance normalisation can again be applied. This process can be represented mathematically by the following:

$$S_{LdB}(m, \omega_k) = 20 \log_{10} \left(\frac{S_L(m, \omega_k)}{ref} \right) \tag{6.22}$$

$$S_{RdB}(m, \omega_k) = 20 \log_{10} \left(\frac{s_R(m, \omega_k)}{ref} \right) \tag{6.23}$$

where $S_L(m, \omega_k)$ and $S_R(m, \omega_k)$ are time-frequency domain representations of the left and right stereo channels, while the respective time-frequency log-magnitude representations are $S_{LdB}(m, \omega_k)$ and $S_{RdB}(m, \omega_k)$. For the TorchAudio implementation used within this work, $ref = 1.0$ [556].

6.4.3 Generalised Cross-Correlation Phase Transform (GCC-PHAT)

Although this work does not focus on predictions relating to discrete sound events, and therefore it is not classed as a sound event localisation and detection (SELD) problem, the aim of the network predicting the dominant direction of arrival for each time-frequency tile can be seen as a related task. For this reason, the Generalised Cross-correlation with phase transform (GCC-PHAT) [557] was chosen as one of the input features as it is widely used for estimating time difference of arrival (TDOA) and is commonly used for SELD based machine listening tasks [558]. However, tasks that utilise stereo signals often focus only on the detection of frontal objects [558, 559], while tasks interested in locating objects within a 3D space often utilise multi-channel microphone arrays and derive the GCC-PHAT for each pair of microphones within the given array [14–16]. The intuition for this experimental work is that given an appropriate dataset combined with an appropriate model, the network may be able to recognise and utilise complex patterns within the stereo data to map to parameters in a 360° space.

The GCC-PHAT is calculated by first transforming the channels into the frequency domain and combining them through a generalised cross-correlation as

defined in [560]:

$$\Psi G[\omega_k] = X_1^*[\omega_k] X_2[\omega_k] \tag{6.24}$$

where X_n is the frequency domain representation of the given channel. The phase transform (PHAT) is then applied such that the magnitudes are normalised and any effects due to amplitude are eliminated:

$$\Psi P[\omega_k] = \frac{\Psi G[\omega_k]}{|\Psi G[\omega_k]|} \tag{6.25}$$

The iFFT is then applied which results in a histogram-like representation and is obtained by:

$$\psi P[\omega_k] = \mathcal{F}^{-1}\left\{ \frac{\Psi G[\omega_k]}{|\Psi G[\omega_k]|} \right\} \tag{6.26}$$

where $\mathcal{F}^{-1}$ is the iFFT and which results in the feature that will be used as input into the proposed network. The delay between the signals can be estimated by reading the histogram such that:

$$\tau = \arg\max \psi P[n] \tag{6.27}$$

It should be noted that when being used within machine learning applications it is common for the GCC-PHAT to be captured for each timeframe resulting in a 2D feature map.

6.5 Architecture

The Multi-channel Residual-U-Net (MuCh-Res-U-Net) proposed in this thesis combines the multi-channel U-Net approach detailed in [561] with a similar Residual-U-Net backbone to that used in [562] and [563]. Originally developed for image segmentation tasks [220] the U-Net architecture has been found to be effective when applied to a number of audio related tasks including source separation [561, 564–568], musical score following [569], voice conversion and cloning [570, 571], denoising [572, 573], and audio synthesis [410, 574]. An additional reason for choice of a U-Net style architecture is that it lends itself to tasks where the input and output data are of similar dimensions due to the symmetry of the encoder and decoder paths. For the purposes of this study an

original U-Net architecture is chosen as the baseline with which the performance of the proposed MuCh-Res-U-Net will be compared against.

6.5.1 U-Net Baseline

The original U-Net architecture proposed by Ronneberger, Fischer, and Brox [220] , is used in this work as a baseline with which to compare the proposed MuCh-Res-U-Net. The original U-Net, as shown in Figure 6.5, consists of an encoding path and a decoding path with skip connections that are passed from the encoding layer to the corresponding decoding layer. The encoding path is similar to traditional convolutional neural networks (CNN) where the resolution of the feature maps decrease through consecutive layers while the number of feature maps/number of filters increases. The encoding path in the original U-Net network consists of convolutional blocks that contain two successive 3 x 3 convolutions followed by a ReLU activation function and a max pooling layer. The original U-net utilises 4 such blocks. The decoding path then upsamples the resulting feature maps using 2 x 2 *up-convolutions*, typically referred to as transposed convolutions within common deep learning frameworks such as Pytorch [575]. Each transposed convolution is followed by two 3 x 3 convolutional layers and a ReLU activation. The final stage includes an additional 1 x 1 convolutional layer to map to the desired number of output channels. The intuition behind the use of skip connections between the encoder-decoder pathways is that it allows the decoder blocks to recover and utilise spatial information that may be lost during the downsampling process. The baseline will also use the original number of features map for each layer which correspond to $[64, 128, 256, 512]$

6.5.2 Residual Connections

The Residual U-Net adopts the residual connections introduced within the ResNet architecture[576]. Residual connections have been shown to improve the training of deeper NNs by mitigating the *vanishing gradient problem* which is caused by small derivatives being multiplied together during back propagation and can result in increasingly small gradients for earlier network layers. Li [577] also

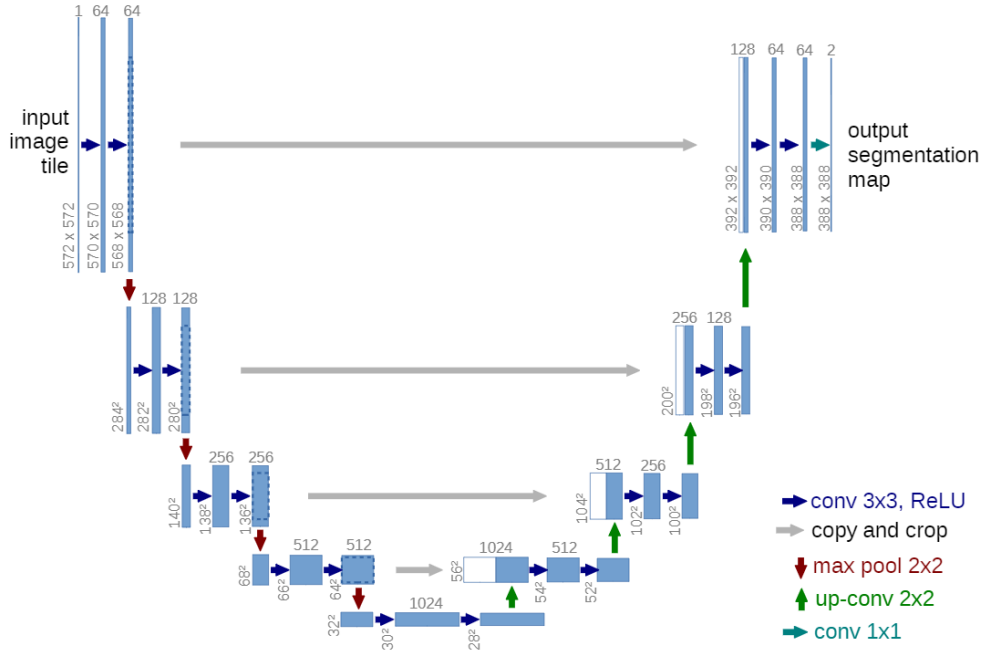


Figure 6.5: Original U-net architecture taken from [220]. Blue boxes correspond to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations.

demonstrates that the loss landscape, and by extension the ease and stability with which a network can be trained, changes with the inclusion of skip connections as shown in Figure 6.6. Residual connections reduce this issue by taking the feature map from one layer and element-wise adding it to a deeper layer in the network. Not only does this serve to preserve the information from earlier feature maps, it also changes the function that the layer has to approximate. Rather than being required to approximate the mapping function $\mathcal{H}(x)$, it instead has to approximate the residual function $\mathcal{H}(x) - x$. The complete block can therefore be formulated as in [578]:

$$y_l = \mathcal{F}(x, \{W_i\}) + \mathcal{I}(x) \tag{6.28}$$

$$x_{l+1} = f(y_l) \tag{6.29}$$

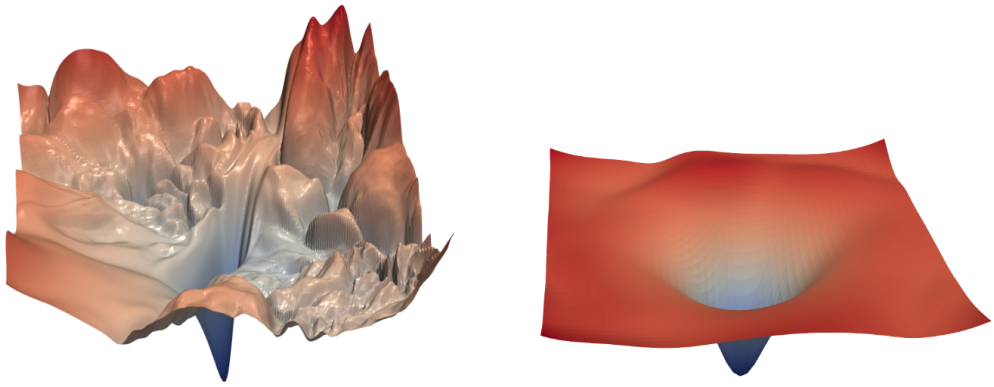


Figure 6.6: Loss surfaces for a ResNet-56 without skip connections (left) and with skip connections (right). Visualisation taken from [577]

where x_l and x_{l+1} are the input and output of the residual unit respectively, $\mathcal{F}(\cdot)$ is the residual function, $f(\cdot)$ is the activation function, and where $\mathcal{I}(\cdot)$ is the identity mapping function where generally $\mathcal{I}(\cdot) = x_l$. This assumes that the optimal desired function is closer to an identity mapping than to a zero mapping [576]. Figure 6.7 shows the structural difference between a regular convolutional block and one with a residual connection. Additionally, the derivative of a sum operation, such as the identity mapping, is 1.0, and this allows the gradient to flow back through that operation unaffected which again helps to mitigate the vanishing gradient problem inherent in deep networks.

6.5.3 Multi-channel Residual-U-Net (MuCh-Res-U-Net)

The proposed MuCh-Res-U-net is an encoder-decoder DNN that utilises the advantages of both the U-net and residual NN architectures. The skip connections between the encoder and decoder pathways allows for information to propagate from the encoding layers to the decoding layers. This serves to preserve and propagate localised features that may otherwise be lost due to the dimensionality reduction of the deeper encoding layers. The residual connections within the encoder and decoder blocks facilitate two main advantages; firstly, they reframe the modelling problem to one of modelling the residual between the input and targets, as opposed to the complete transform from input to target [563]. Secondly,

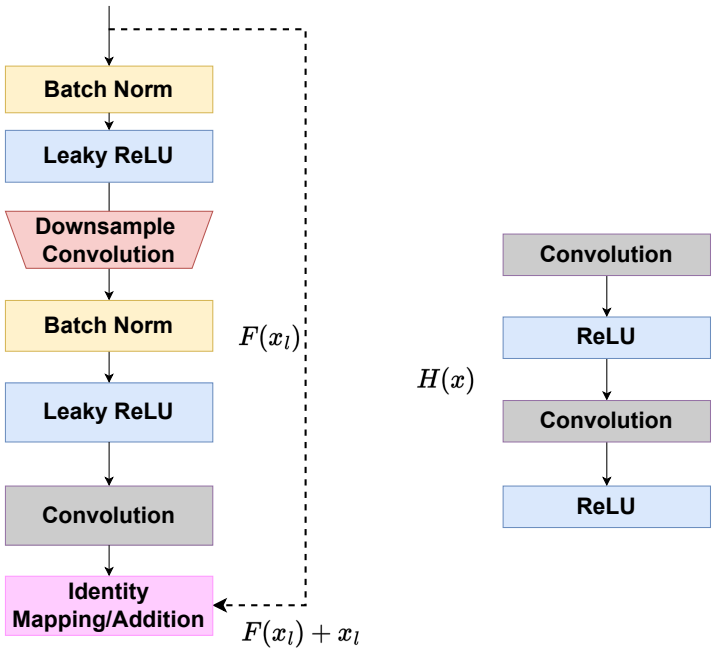


Figure 6.7: Regular convolutional block used in U-Net (right) and Residual unit used in Res-U-Net (left)

they allow gradients to be backpropagated unimpeded to earlier initial layers due to the nature of derivatives of summation operations. This mitigates the vanishing gradient problem previously mentioned which causes gradients to approach zero for earlier layers due to sequential multiplications of small numbers.

This thesis utilises a 9-level Res-U-Net architecture, as shown in figure 6.8, with a multi-channel output to predict time-frequency parametric spatial features equivilant to those resulting from the DirAC analysis of B-format signals. As shown in Figure 6.8, the network comprises three main stages: encoding, bottleneck, and decoding. The encoding pathway encodes the input features into a high number of low dimensional representations. The bottleneck serves to connect the encoding and decoding pathways and has an internal structure identical to the encoder blocks, the decoding pathway then decodes and extracts the target features with the final convolutional layer extracting the required number of output feature maps. All stages utilise residual connections, convolutional blocks, and identity mapping. Each convolution block contains a batch normalisation

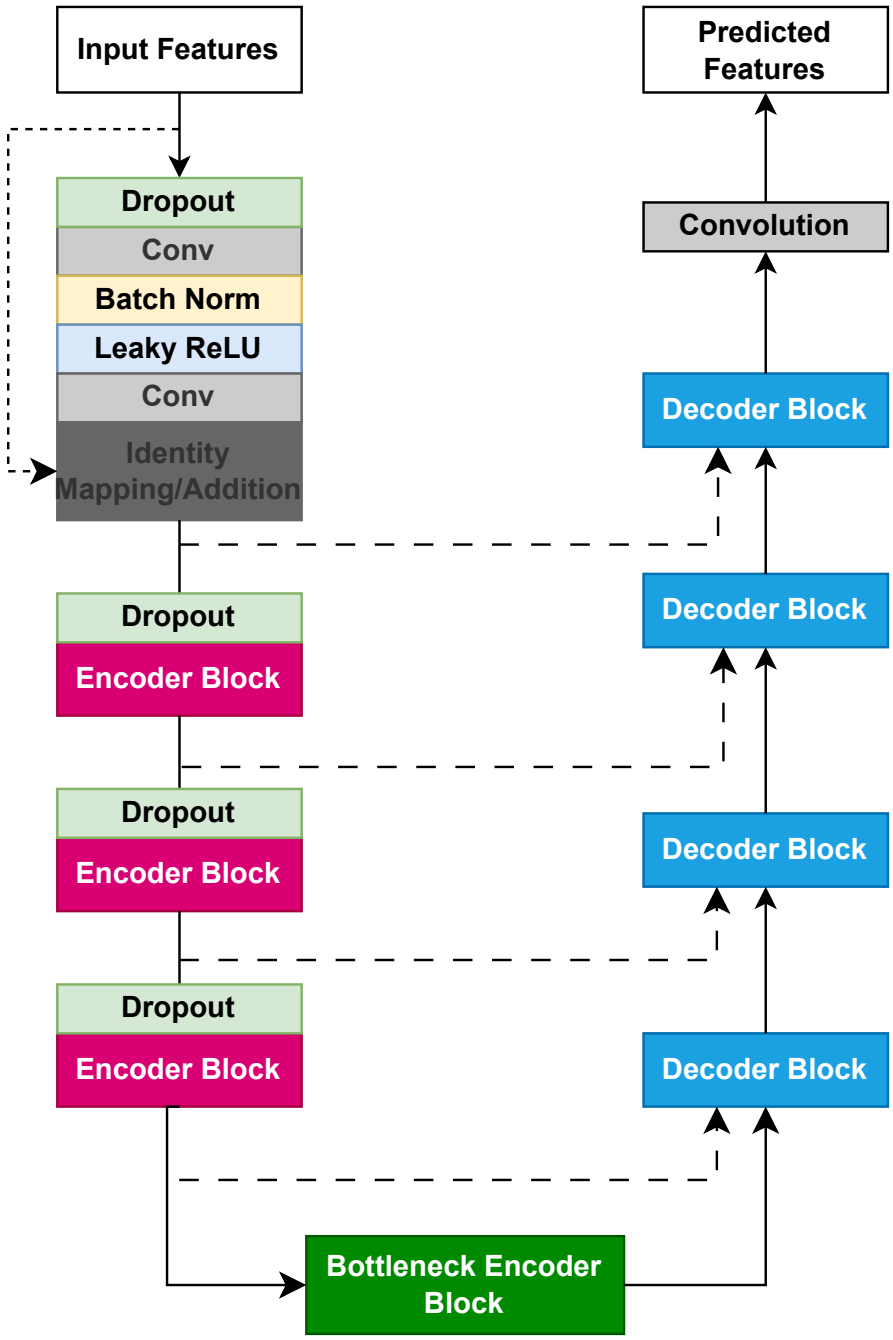
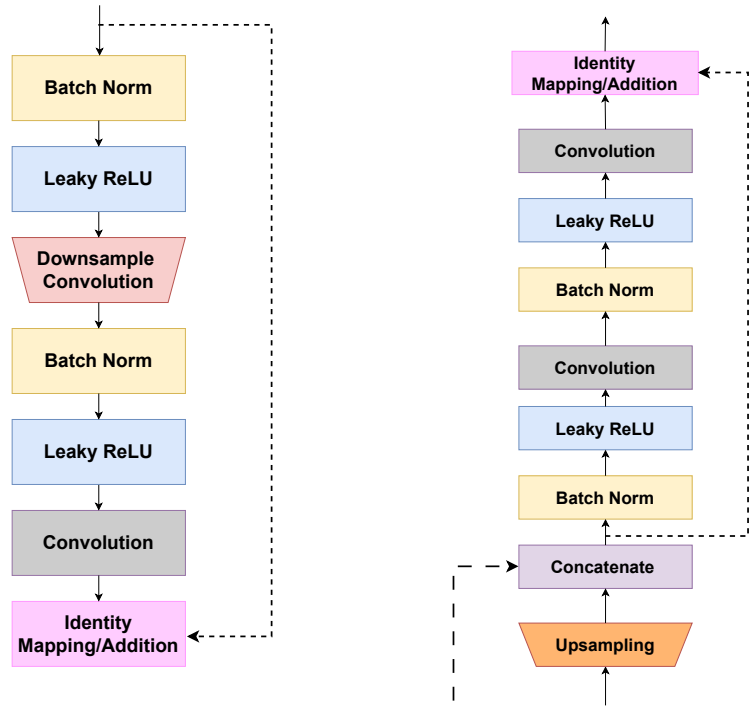


Figure 6.8: Proposed MuCh-Res-U-Net architecture

layer, a Leaky ReLU activation layer and a convolutional layer. As the model is convolutional it can process input sequences of arbitrary length [563], only being limited by the amount of available compute resources. Apart from the first block, two convolutional blocks are stacked sequentially as can be seen in Figure 6.9.



(a) Encoder Block (b) Decoder Block

Figure 6.9: Encoder and decoder blocks for MuCh-Res-U-Net.

At the feature extraction stage, prior to the time-frequency transform, there is a noise injection layer which randomly adds Gaussian noise to the time-domain signals based on a given probability. Noise injection has been shown as an effective regularisation method as it serves as a type of data augmentation to prevent the network overfitting through continuous sampling of the noise inherent in smaller datasets [579, 580]. The amount of noise added is scaled according to each example in order to achieve a SNR of 20 dB, a value reached through iterative testing.

The encoding path consists of 4 encoding blocks, each of which has a dropout layer prior to the input into the block. Dropout is another simple regularization

method designed to prevent overfitting where neurons and their connections are effectively deactivated for that training step [581]. The neurons selected to be dropped are chosen at random according to a user defined probability between 0 and 1; for example, a value of 0.25 would mean each neuron has a 25% chance of being dropped. This serves to reduce the co-adaption of neurons. As a different set of neurons are dropped each time, this can be viewed as a form of architecture augmentation and results in the network being trained from a sample of reduced or *thinnned* networks [581]. Whereas the original U-net used a maxpooling for dimensionality reduction, the proposed network uses a convolutional layer with a stride of 2 at the start of each encoding block, which results in a downsampling of the feature maps by a factor of 2. The identity mapping is achieved through the use of a convolutional layer with a kernel size of 1 and serves to broadcast the input to the correct number of channels at the block output. The input signal is then summed with the output of the final convolution layer in the block.

The decoding path also comprises 4 blocks, the structure of which is shown in Figure 6.9b. Each block first upsamples the feature maps by a factor of 2 followed by a concatenation with the skip connection from the corresponding encoding block. The rest of the decoding block structure is then identical to that of the encoding block. After the last decoding block a convolutional layer with a kernel size of 1 x 1 is then used to approximate the final mapping from the multi-channel feature maps to the target features. The output layer consists of either 2 feature maps if predicting time-frequency azimuth and elevation values or 3 feature maps if also predicting the time-frequency diffuseness index.

Whilst it is common for many networks based on convolutional blocks to have fewer filters in early blocks, typical starting values being 32 and 64 [220, 561, 572], in this instance it was found through initial experimentation that the network under investigation began to converge earlier and more stably with fewer layers, but containing a higher number of filters.

Although both the original U-net [220] and Resnet [576] architectures utilise a ReLU activation function, the MuCh-Res-U-Net instead uses the Leaky ReLU activation. Due to the ReLU function being zero when $x < 0$ problems can

arise similar to that of the vanishing gradient issue. The dying ReLU problem [582] refers to a situation where ReLU neurons become inactive and only output zero for any input, this could be caused, for instance, by a large negative valued bias. This in turn causes the derivative with respect to that neuron to also be constantly zero, effectively stopping the flow of gradients back through that neuron which results in the weights not being updated and causing it to become stuck in a local minima [583]. Even in situations where the ReLU does not remain inactive indefinitely, training can be slowed as during optimisation the gradient is 0 whenever the unit is not active, and thus will not have its weights adjusted by the optimiser [584].

6.6 Training

6.6.1 Dataset

The proposed MuCh-Res-U-Net was trained on the *AB_Omni_40* set, detailed in Table 6.2. The choice to train the network on a single stereo configuration was made in order to limit the complexity of the problem space for this initial investigation. The intuition is that the required mapping function for a single stereo configuration is going to be mathematically simpler to approximate than a mapping function that is sufficient to account for multiple stereo configurations that each possess varying time-frequency characteristics including, directional response, frequency response, and inter-channel differences resulting from the spaced distance and degree of coincidence. The *AB_Omni_40* set was chosen based on the results of an initial set of experiments conducted on 60 training examples to ascertain which stereo configuration had the potential to converge the fastest. It is acknowledged that 60 examples is too small a dataset on which to base any definitive conclusions of training potential, however, as the work was practically limited by available compute power it was decided this would be adequate in deciding on a configuration with which to conduct this initial investigation. Additionally, the omnidirectional signals allow for either channel to be taken as an approximation for the omnidirectional pressure component

which will simplify the upmixing pipelines as detailed in Section 6.7. The 40 cm separation allows the network to take advantage of both TDOA information and some degree of inter-channel level differences, whilst the distance from either microphone to the furthest loudspeaker is not enough to cause a substantial drop in magnitude compared to that of a source coming from the closest loudspeaker. For example, the closest loudspeaker will be approximately 1.3m away from a given microphone and the furthest would then be approximately 1.7m. Using Equation 2.9 this can be calculated to result in an approximate maximum SPL difference of 3dB. 6000 x 7 second samples were synthesised at 22.05 kHz for training, validation, and testing. Reduced bandwidth audio was used to reduce the computational cost. The dataset of synthesised scenes was split into folds according to the NIGENS split the sound events originated from. This resulted in 4500 examples for training, 750 for validation, and 750 for final testing which yields a percentage split of 75% training, and 12.5% for each of the validation and test sets.

6.6.2 Experimental Set-up

The input feature vector was of shape 303 x 513 x 3 corresponding to 303 time frames, 513 frequency bins, and 3 features maps for the GCC-PHAT and the STFT of each stereo channel. As there exists a large number of adjustable hyperparameters it is often not possible to conduct an exhaustive search of the the complete n-dimensional hyperparameter space. Therefore, a hyperparameter search was conducted using a grid search approach facilitated by the Weights and Biases library [585]. The list of hyperparameters explored can be found in Table 6.3. The maximum number of epochs for each tuning run was 100, although some runs were terminated early if they were seen to be overfitting or if the model became unstable and produced NaN values for at least 3+ epochs. The final selection of hyperparameters were based on the validation performance of the model. It is important, however, to note that while weight updates were not directly effected by the validation loss, basing design choices on validation performance will introduce some inherent data leakage in the training process,

Parameter	Range	Step size	Final selection
Learning Rate (η)	0.0001 to 1	$\eta_{new} = 2\eta_{old}$	8 x 10^{-4}
Warmup Steps	$[1, 100, 1000, 10000, 15000]$	N/A	1000
Dropout	[0, 0.5]	N/A	0.0
Noise injection probability	[0, 0.5]	N/A	0.5
Zero-mean Unit-var norm	[True, False]	N/A	True
log transform	[True, False]	N/A	True
No. of Conv. layers	3,4	N/A	4
No. of filters	$[128, 256, 512, 1024]$ $[64, 128, 256, 512]$ $[128, 256, 512]$	N/A	$[128, 256, 512, 1024]$

Table 6.3: Details of hyperparameter sweeps including parameters and defined search range.

hence the need to reserve a final completely unseen test set. The selected model had 154.8 million parameters.

A learning rate schedule was adopted that consisted of a linear warm-up over 1000 steps to a maximum learning rate of $\eta = 8 \times 10^{-4}$. The learning rate remained static for 2 $\times$ warm-up steps before following a scheme of Cosine Annealing with warm restarts [586] with 10 epochs for the initial restart with the number of epochs between subsequent restarts increasing each time by a factor of 2. Defined in [587] as:

$$\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min}\left(1 + \cos\left(\frac{T_{cur}}{Ti}\right)\pi\right) \tag{6.30}$$

where η_t is the learning rate for current step, η_{max} is the maximum learning rate, η_{min} is the minimum learning rate, T_{cur} is the number of epochs since the last restart, and T_i is the number of epochs between two restarts.

Training and validation loss were recorded after each epoch for both total loss and individual feature loss. An adaptive gradient clipping method proposed in [588] was used which sets a clipping threshold based on the history of gradient norms observed the training run. This helps to minimise the risk of exploding gradients caused by the often non-smooth nature of NN loss landscapes [589] and

allows for an appropriate selection of the clipping threshold parameter without having to include it in a hyperparameter search. It was set to clip to the 10th percentile of the derived threshold as this would help to ensure any outliers would not have a disproportionate impact the clipping threshold. This work uses the Mean Squared Error (MSE) between the estimated time-frequency parameter values and the ground truth parameter values as the loss function and can be defined as:

$$loss(\hat{y}_i, y_i) = \sum_{i=0}^{I} \frac{1}{K} \sum_{k=0}^{K} (\hat{y}_{ik} - y_{ik})^2 \tag{6.31}$$

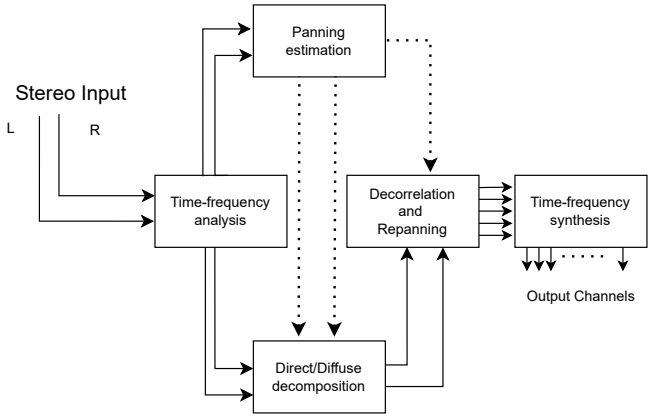
where $\hat{y}_{ik}$ is the prediction for the k_{th} time-frequency tile in the i_{th} target feature map. The losses from each feature are summed to get the final loss.

The model was trained on a single Nvidia RTX 3090 [590] using mini-batch gradient descent with a batch size of 6 and optimised using Adam with decoupled weight decay regularisation [587]. To increase the effective batch size and negate some of the issues associated with small batch sizes, such as larger inter-batch variance, gradient accumulation [591] was utilised to create an effective batch size of 45. The final model configuration was trained for 100 epochs with model checkpointing each time the validation loss reached a new minimum.

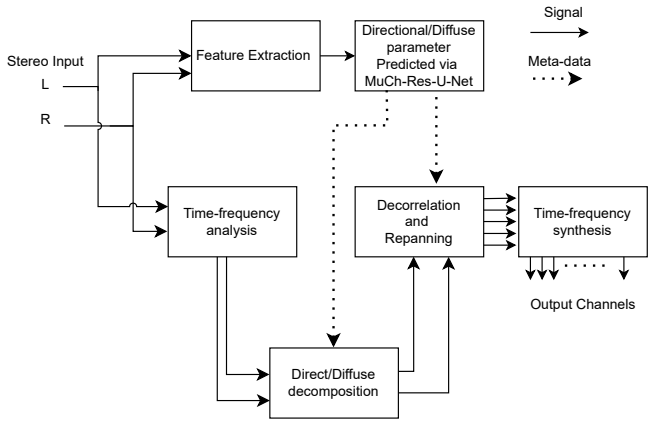
6.7 Example Upmixing pipeline

6.7.1 Upmixing using Directional Audio Coding

Figure 6.10a shows a generic time-frequency parametric stereo upmix processor. The main differentiation between such stereo upmixers and other spatial time-frequency parametric processors, such as DirAC [149], is usually the number of accepted input channels and whether the panning estimation is done on a lateral or 360° basis. This in turn determines the behaviour of the decorrelation and repanning block. As previously mentioned, parametric stereo upmixers such as those proposed in [497–500, 503], are usually limited to lateral positional estimates on the horizontal plane with direct components assumed to be coming from the front. Replacing the panning estimation block with the proposed



(a) Generic time-frequency parametric stereo upmix processor



(b) Stereo upmix processing with panning estimation block replaced by proposed MuCh-Res-U-Net

Figure 6.10: Block diagrams showing a) generic time-frequency parametric stereo upmix processor and b) A stereo upmix processor with the panning estimation block replaced by the proposed MuCh-Res-U-Net that predicts direct/diffuse parameters for 360° space

MuCh-Res-U-Net, as shown in Figure 6.10b, would allow for the directional and diffuseness parameters be predicted for 360° space. Many upmix processors are designed and tested using synthetic stereo material, where placement in the stereo field is determined solely by inter-channel amplitude differences. This may introduce challenges when using traditional directional estimation methods to upmix stereo signals recorded using spaced configurations due to the addition of TDOA between the two stereo signals and the comparatively small ICLD. If the same time-frequency tile from each signal is repositioned to the same

location the phase differences caused by the TDOA may potentially lead to comb filtering artefacts during playback. In these situations it may be beneficial to instead derive a mono downmix from the original stereo signals prior to the time-frequency analysis as shown in Figure 6.11. For this work, the mono downmix is approximated by simply processing only one of the two stereo channels. This assumes that the inter-channel amplitude differences should be negligible for most standard spaced stereo pairs and therefore each signal would have a very similar magnitude spectrum if the TDOA were removed. As discussed in Section 6.3.8, an omnidirectional signal can approximated for the sound field pressure component, which can then be used, given adequate spatial metadata, to reconstruct a spatial sound field using parametric time-frequency spatial audio system such as DirAC [151].

6.7.2 Upmixing to B-format

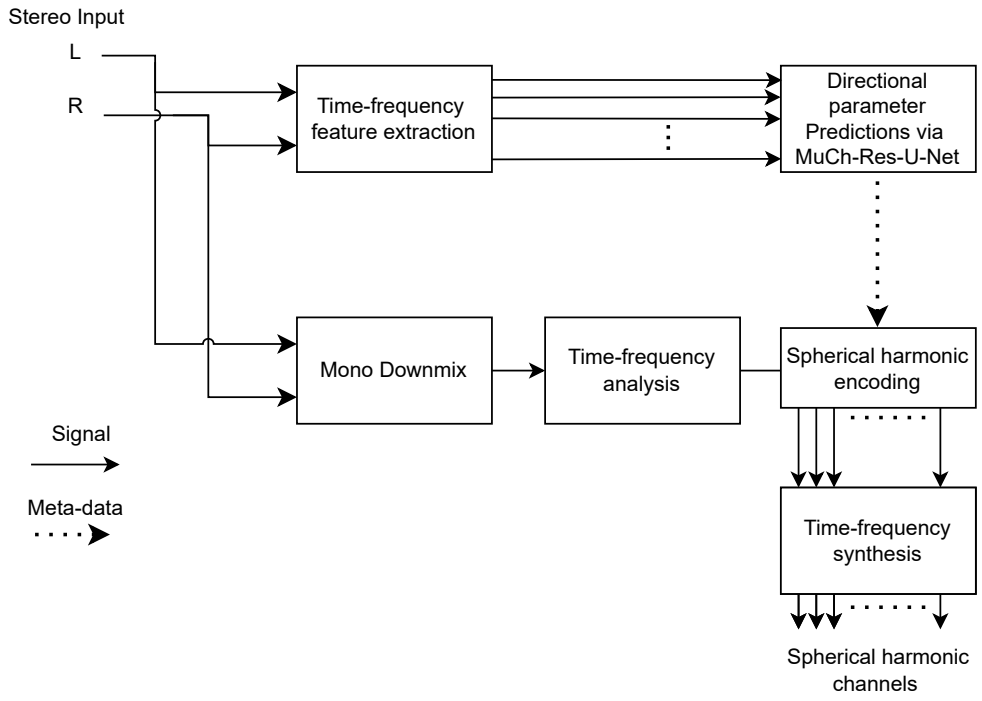


Figure 6.11: Block diagram of proposed stereo to B-format upmixer utilising directional parameters predicted by MuCh-Res-U-Net

The proposed MuCh-Res-U-Net can also be applied as part of a scene-based upmixing pipeline whereby the predicted time-frequency directional parameters can be used to extract and remap frequency components into target spherical harmonic components. Although the pipeline itself is relatively simple, as shown in Figure 6.11, it allows the frequency components to be mapped in 3D space in such a way that the resulting spatial representation could be comparable to that which would have resulted had the scene been captured with a suitable microphone array, given an accurate enough model.

An example is now presented of a pipeline to upmix a stereo scene captured with a spaced stereo pair to first order spherical harmonic components, which will be referred to by their B-format channel labelling. First, a mono signal must be derived to represent the W channel which, as detailed in Section 2.6.5.1, is an omnidirectional pressure signal. Due to the low inter-channel amplitude differences, the mono signal can be approximated using one of the two stereo signals such that:

$$W \simeq S_L \| S_R \tag{6.32}$$

where S_L and S_R are the left and right stereo channels respectively. The W channel is then transformed into the time-frequency domain using the process detailed in Equation 2.48 which for brevity will be represented as:

$$W(m, \omega_k) = \mathcal{STFT}\{W\} \tag{6.33}$$

Directional features are then predicted by the network and these are used to extract and weight the frequency components according to the target spherical harmonic coefficients [160]:

$$\beta^{\sigma}_{mi}(m, \omega_k) = W(m, \omega_k) Y^{\sigma}_{mi}(\hat{\theta}(m, \omega_k), \hat{\phi}(m, \omega_k)) \tag{6.34}$$

Where:

- $\beta^{\sigma}_{mi}(m, \omega_k)$ is the time-frequency representation of the Ambisonic channel representing the spherical harmonic Y^{σ}_{mi},

- $W(m, \omega_k)$ is the time-frequency representation of the W channel from which the frequency components are being extracted and remapped. This approach is similar to that proposed in [151] where DirAC for telecommunications only transmits the metadata and W channel, discarding the other B-format channels after DirAC analysis.

- $\hat{\theta}(m, \omega_k)$ and $\hat{\phi}(m, \omega_k)$ are the predicted time-frequency directional parameters for azimuth and elevation respectively.

Lastly, the resulting time-frequency Ambisonic channels can then be returned into the time-domain expressed as:

$$\beta^{\sigma}_{mi} = \mathcal{STFT}^{-1}\{\beta^{\sigma}_{mi}(m, \omega_k)\} \tag{6.35}$$

where $\mathcal{STFT}^{-1}\{\cdot\}$ is the inverse STFT.

6.8 Results and Discussion

6.8.1 Neural Network

Table 6.4 shows the performance on the test set with respect to the MSE loss while Table 6.5 shows the hyperparameters of each model. Figure 6.12 shows the validation loss for each model over epochs. The baseline model (black line) takes longer before it starts to noticeably optimise and once it does it begins to converge much slower than the MuCh-Res-U-Net models and also settles into a local minima with a higher loss value. The loss for MuCh-Res-U-Net-Best appears to be continuing to reduce, albeit it at a very slow rate, indicating that an increase in training time or further model/training optimisation may continue to yield improvements in the model's accuracy. The slight increases in the loss observed around epochs 30 and 59 coincide with the learning rate warm restart. Whilst MuCh-Res-U-Net-overfit followed the same optimisation rent for the first approximately 30 epochs, it began to overfit due to lack of regularisation, and specifically we can see from Figure 6.13 that it began fitting to the background noise of the training set. Most networks will begin to overfit at some stage

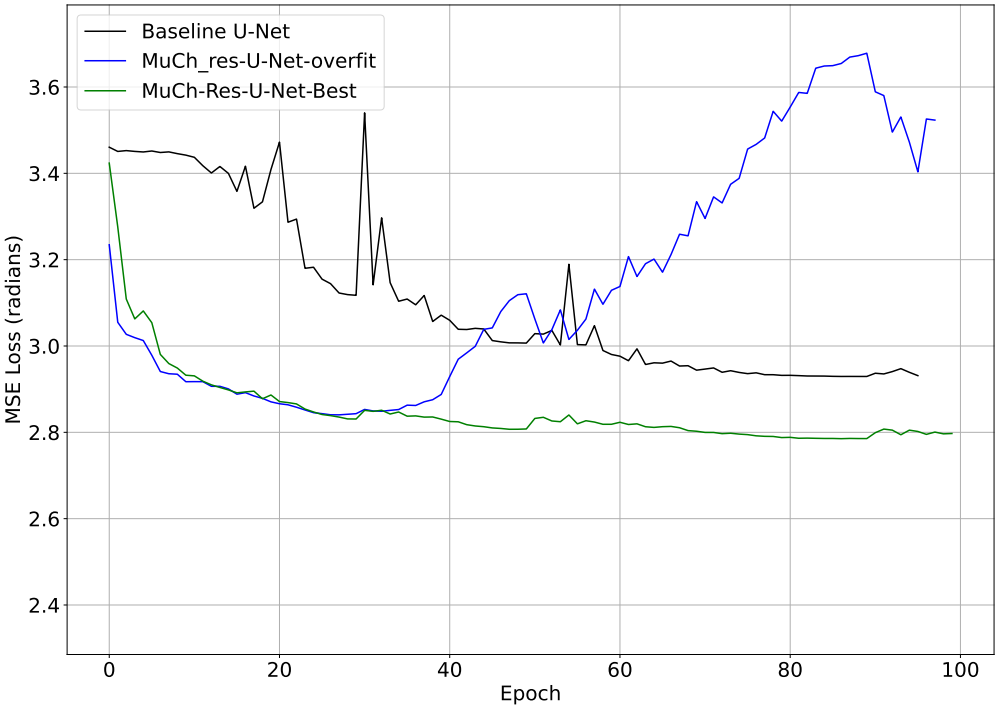


Figure 6.12: Validation loss curves for baseline mode, best performing model, and a model that is representative of overfitting. The measured loss Baseline and MuCh-Res-U-Net-Best continue to decrease slowly over time while the MuCh-Res-U-Net begins to overfit at around epoch 35 as evidenced by the increase in its loss value. The sharp peaks in the loss curves coincide with the learning rates warm restart.

without regularisation, however, one potential cause for the overfitting in this instance is due to how the ambient noise for each training sample was synthesised. As discussed in Section 6.3.7, a 2 minute long recording was made to capture the ambient noise of the measurement space, but given that there are 4500 training samples in total, each 7 seconds in length, the same sections of that background noise will be present in many of the training samples. The repeated sampling of the noise makes it more likely for the network to start to converge on a mapping function for the specific noise distribution present within the training set.

The MSE, whilst a common loss function for regression tasks, is measured in squared units and does not provide an intuitive loss value with respect to "how far away" the predictions are from the ground truth in the units of interest. With respect to the directional parameters, the MSE is providing an error value in

Model	RMSE			
	Total	azimuth (θ°)	Elevation (ϕ°)	Diffusness (ψ)
Baseline U-Net	1.75	95.11	31.51	0.034
MuCh-Res-U-Net-Overfit	1.74	94.19	30.88	0.176
MuCh-Res-U-Net-Best	1.72	92.82	30.37	0.170

Table 6.4: MSE results for the test set. Results are given for both individual parameter loss and total loss. Total loss is calculated as the sum of parameter losses. Loss θ and ϕ was calculated in radians but have been converted into degrees for clarity. Results show across all parameters MuCh-Res-U-Net achieved the lowest loss value.

radians squared. Taking the root of the MSE, which results in the Root Mean Square Error (RMSE), provides a loss value in the units of interest. Upon initial examination, the model appears to be performing poorly as the RMSE indicates an error value of 92.8° for azimuth predictions and 31.5° for elevation predictions. However, it should be noted that these values are derived from 155,439 equally weighted time-frequency tiles that represent the entire time-frequency spectrum, and as such may not be appropriate given the complex inter-tile relationships that exist within sound spectra. Additionally, the importance of accuracy with respect to different parts of the time-frequency spectrum will vary depending on the sources contained within them. It would arguably be more important to correctly predict directional values for time-frequency tiles relating to sound sources than it would be for the parts of the spectrum only containing diffuse background noise. Consequently, the MSE may not be the best placed loss function to optimise for the intended mapping function given that all time-frequency tiles contribute equally to final loss value. From an analysis perspective it also fails to provide a comprehensive insight into the model's performance across the time-frequency spectrum.

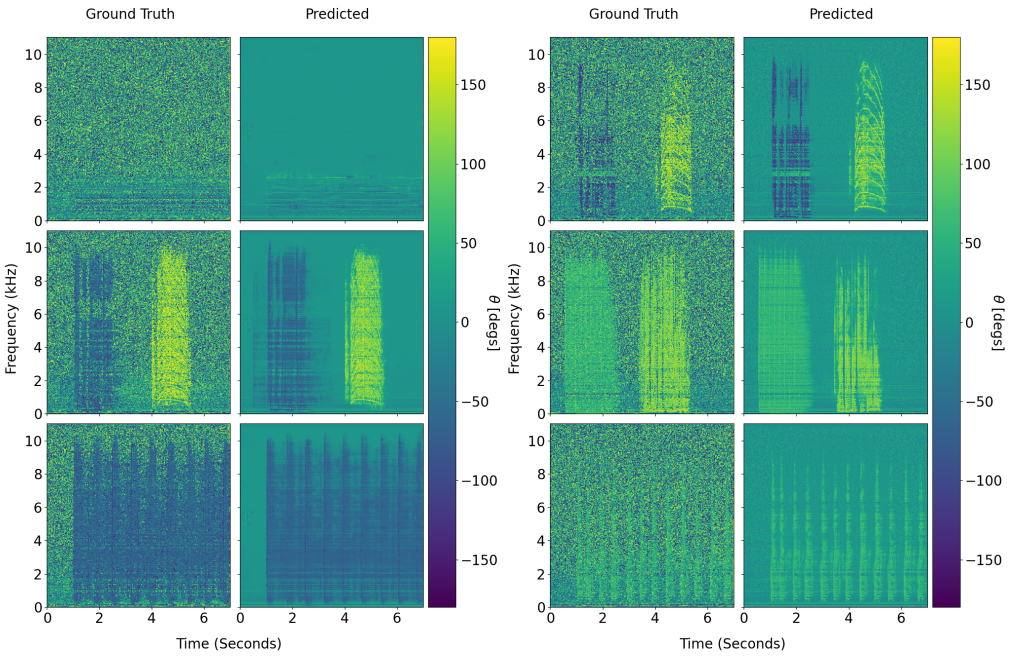
Figure 6.13 shows predicted and ground truth azimuth values for 3 randomly selected examples from the test set for the baseline and MuCh-Res-U-Net-Best models. The ground truth values are those derived directly from B-format signals using DirAC analysis whilst the predicted values are from the models using stereo

Parameter	Baseline	MuCh-Res-U-Net-OverFit	MuCh-Res-U-Net-Best
Learning rate	0.0008	0.0008	0.0008
Warmup Steps	1000	1000	1000
Dropout	0.0	0.0	0.0
Noise injection probability	0.0	0.0	0.5
Zero-mean Unit-var norm	True	True	False
Log transform	True	True	True
No. of Conv. layers	4	4	4
No. of filters	$[64, 128, 256, 512]$	$[128, 256, 512, 1024]$	$[128, 256, 512, 1024]$

Table 6.5: Hyperparameters for the models shown in Figure 6.12

signals as input. Although the MSE values in Table 6.4 would give the impression that the model is performing poorly, as can be seen from visualising the data the model is in fact beginning to generalise relatively well to the parameters as they relate to the part of the time-frequency spectrum occupied by the source within the scene. The noise injection utilised by MuCh-Res-U-Net-Best, which randomly applies Gaussian noise to 50% of input samples, succeeds in regularizing the model, although, the model fails to improve substantially in the remaining $\approx$60 epochs. This could indicate that there is a greater challenge in optimising the network to approximate a general mapping for the parts of the spectrum where the frequency content can be considered random or diffuse noise, which in this case is related to the ambient background noise of the training example. A similar trend can be seen for both elevation and diffuseness predictions as shown in Figures 6.14 and 6.15.

When the model was allowed to overfit an interesting effect is observed on the predictions made on the validation set. Figure 6.16 shows prediction and ground truth data for MuCh-Res-U-Net-Overfit after training for 100 epochs. As the model begins to fit to the noise in the training data it starts predicting similar noise like spectra in the ambient portion of the validation set examples, as expected. This causes an increase in validation loss as the noise component predicted is not aligned with the noise component present in the validation set. However, given the nature of diffuse background noise it is unimportant in this

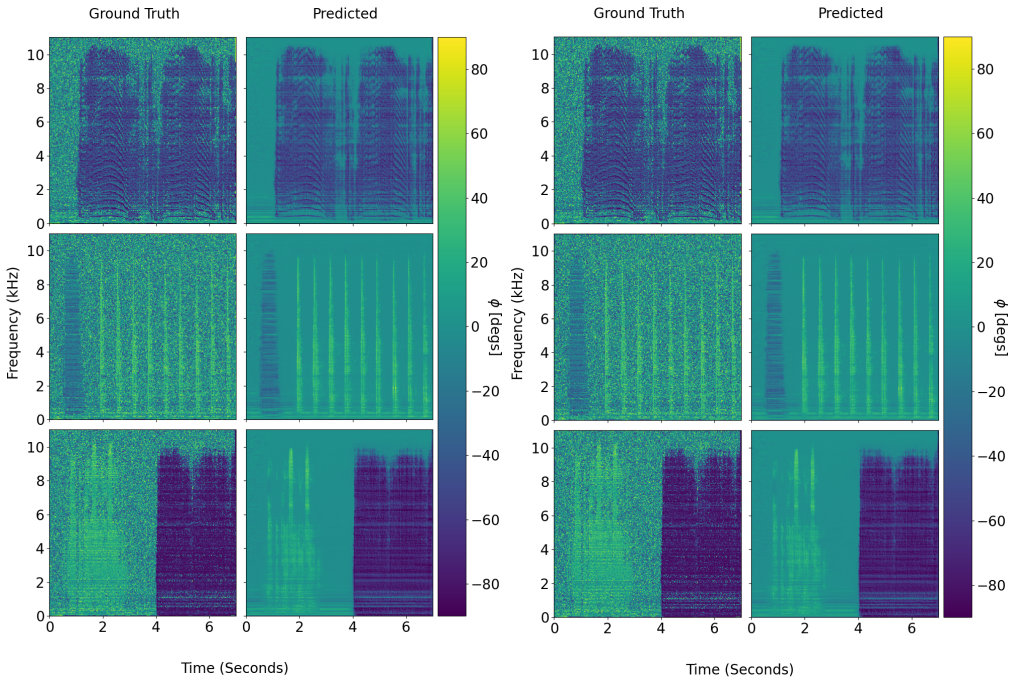


(a) Baseline

(b) MuCh-Res-U-Net-Best

Figure 6.13: Ground-truth and predicted time-frequency azimuth parameter values for the a) baseline U-Net and b) MuCh-Res-U-Net-Best. Each row is a randomly selected example from the test set, with the left hand column containing the ground truth data and the right hand column containing the output from the model.

context as to whether the predicted directional parameters for the frequency components related to the noise are closely aligned with that of the ground truth, as long they are still sufficiently diffuse in nature. Although the predictions from the model are objectively getting further away from the ground truth, the portion relating to the direct sound source are still generalised well, and the diffuse distribution of directional parameter values relating to the ambient component of the training example could, in fact, be desirable. It is also worth noting that the time-frequency tiles occupied by the ambient background noise tend to also have higher diffuse index values as shown in Figure 6.15 and will therefore likely contain much less directional energy. However due to the limitations of the DirAC analysis used to derive the target features, all time-frequency tiles will be allocated a given direction even if there is little directional energy contained within it.



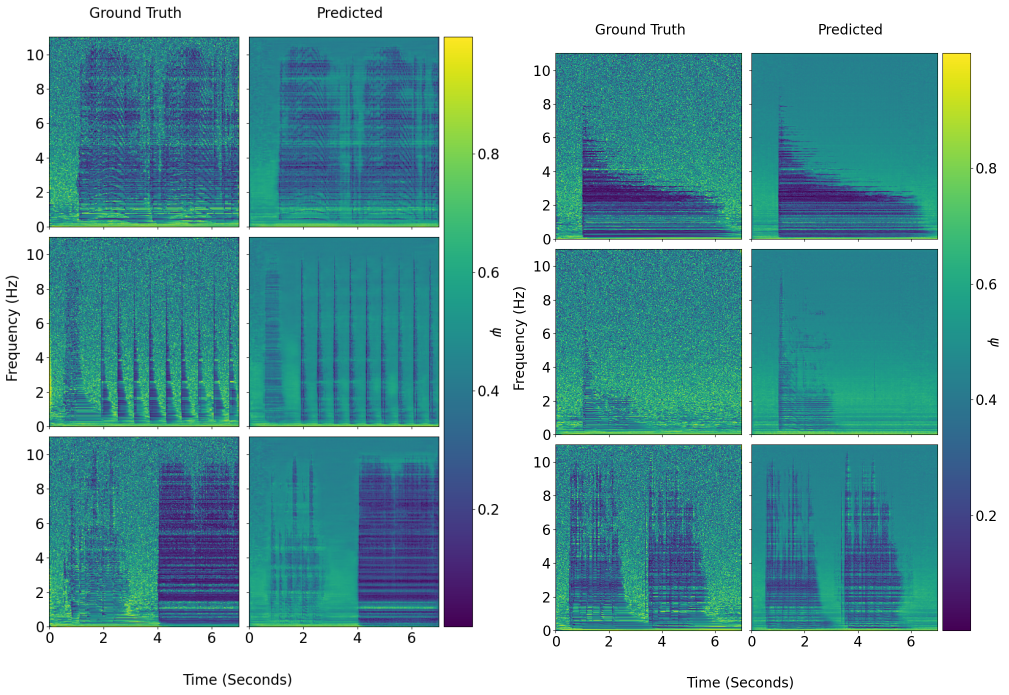
(a) Baseline (b) MuCh-Res-U-Net-Best

Figure 6.14: Ground-truth and predicted time-frequency elevation parameter values for the a) baseline U-Net and b) MuCh-Res-U-Net-Best.

6.8.2 Evaluation of B-format upmix pipeline

Several stereo sound scenes were upmixed into B-format using the pipeline detailed in Section 6.7.2 to provide some preliminary evaluation. Figures 6.17 and 6.18 show the spectrogram of the original B-format channels and the upmixed B-format from stereo signals using the method proposed in Section 6.7.2. Given these two signals were captured with two different microphones, in slightly different locations within the measurement rig, and that the W channel channel derived from the Eigenmike has already gone through a filtering process, there are differences between the the spectra as expected. A consequence of this is that many of the perceptually driven metrics, such as those presented in [592], which base the comparison on how similar the predicted signal is to the ground truth will likely score this system poorly.

To illustrate, Table 6.6 shows scores given to two scenes that have been upmixed from stereo to B-format and an unrelated upmixed and B-format scene



(a) Baseline (b) MuCh-Res-U-Net-Best

Figure 6.15: Ground-truth and predicted time-frequency diffuseness parameter values for the a) baseline U-Net and b) MuCh-Res-U-Net-Best.

for comparison, all of which use original B-format signals as the reference, these include the Aggregate STFT [593] loss, the multi-resolution STFT loss [594], and a perceptually motivated NN model trained on JNDs scores [595].

Model	Loss		
	Aggregate STFT	Multi-resolution STFT	ML JNDs
Fold4 mix 003	2.01	2.12	2.22
Fold6 mix 265	2.01	2.13	2.73
Unrelated scenes	3.94	4.0	3.10

Table 6.6: Results for audio loss metrics comparing upmixed B-format to original B-format.

As expected, none score particularly highly, however the upmixed scenes do score better with their original counterpart than unrelated scenes. What this

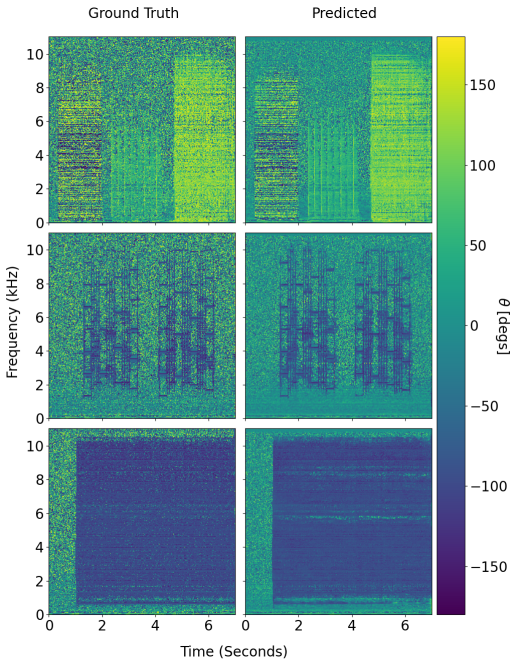


Figure 6.16: Ground-truth and predicted time-frequency azimuth parameter values for MuCh-Res-U-Net-overfit taken from the validation set. Note how when the model overfits it begins to predict similar to noise like spectra in the ambient portion of the training example.

establishes is that similarity metrics, although useful as perceptually motivated loss functions, may not be the most appropriate type of metric for assessing this particular class of upmix algorithm as there will always be inherent differences that stem from the microphones, recording equipment, and any processing required used to capture and encode the respective input and 'target' signals. In this instance, evaluations based on preference scores, such as mean opinion scores, may be more appropriate.

As the perceptual metrics used earlier in this section are primarily intended to be used as perceptual loss function for training neural networks, as opposed to perceptually evaluating spatial scenes, additional evaluation of the pipeline was performed to quantify the spatial accuracy of the upmixed signals, when compared to known ground truth signals. The IRs used to synthesise the training data were also used to spatialise a 3s pink noise burst, followed by 0.5s of silence,

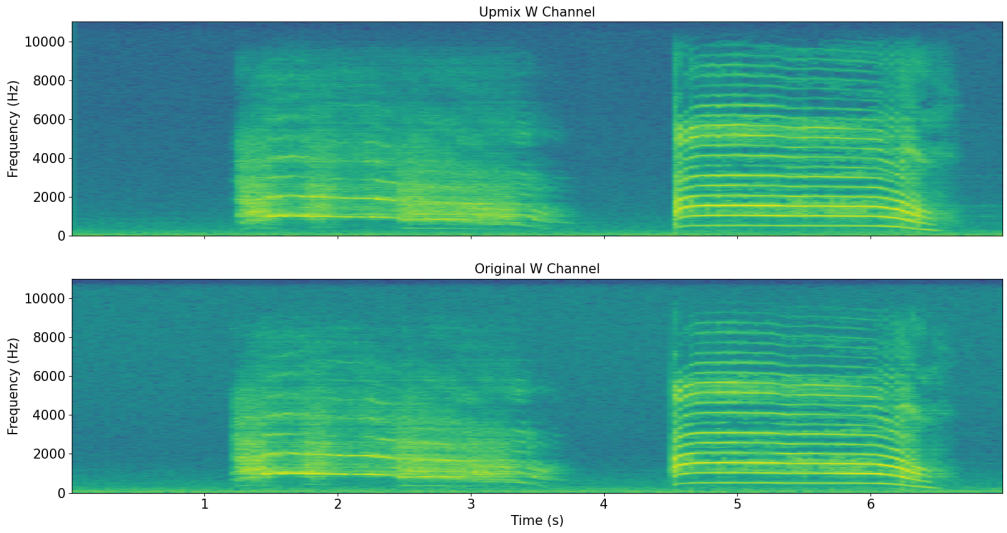


Figure 6.17: Upmixed W channel (Top) original W channel (bottom). Perceivable different in spectra may be a consequence of the microphone, recording equipment, and any subsequent processing that went into capturing and encoding the shown signals.

at all sampled locations on the horizontal and all elevation locations directly frontal to the receiver, which due to the Lebedev sampling scheme were located at azimuth positions 0°, 18°, or 45°. The directional performance of the upmix algorithm is evaluated based on the spherical distance, as defined in [15], between the DOA estimations (DOA-Est) for the upmixed B-format signals and the ground-truth B-format signal and will be referred to as the Total DOA error. It can be calculated as follows:

$$\Delta DOA^{3D} = \arccos(\sin(\hat{\phi})\sin(\phi) + \cos(\hat{\phi})\cos(\phi)\cos(|\theta - \hat{\theta}|)) \tag{6.36}$$

where ΔDOA^{3D} is the Total DOA error as spherical distance in degrees $^\circ$ and $\hat{\theta}, \hat{\phi}$ are the DOA-Est from the upmixed B-format signals and θ, ϕ are the DOA-Est for the ground-truth B-format signals.

When referring to the DOA error for a single direction, either θ or ϕ, the 2D angular distance used, as defined in [596]:

$$\Delta DOA^{2D\theta} = |\theta - \hat{\theta}| \tag{6.37}$$

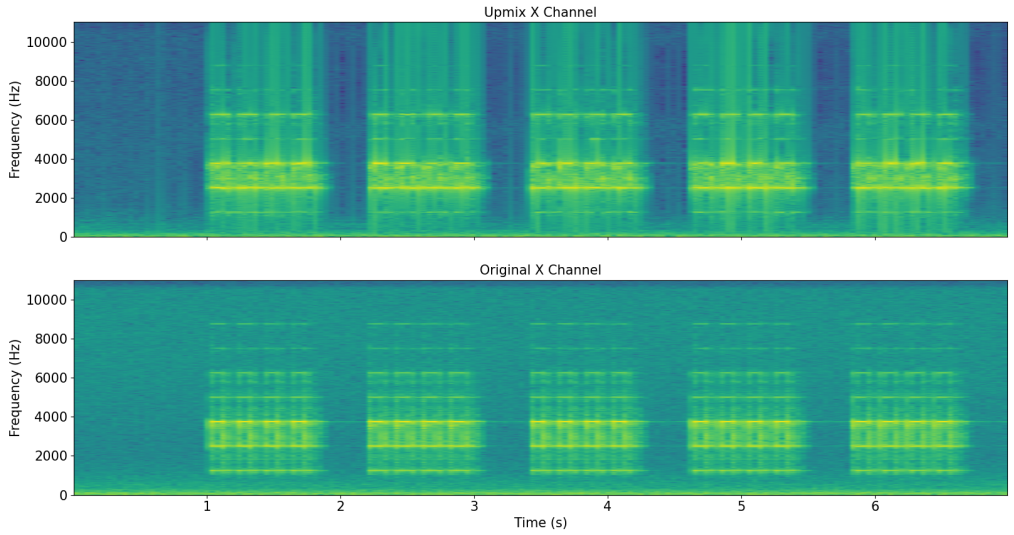


Figure 6.18: Upmixed X channel (Top) original X channel (bottom). Perceivable different in spectra may be a consequence of the microphone, recording equipment, and any subsequent processing that went into capturing and encoding the shown signals similar to that observed in figure 6.17

where $\Delta DOA^{2D\theta}$ is error in the azimuthal direction and where error in the elevation direction, $\Delta DOA^{2D\phi}$, is calculated by :

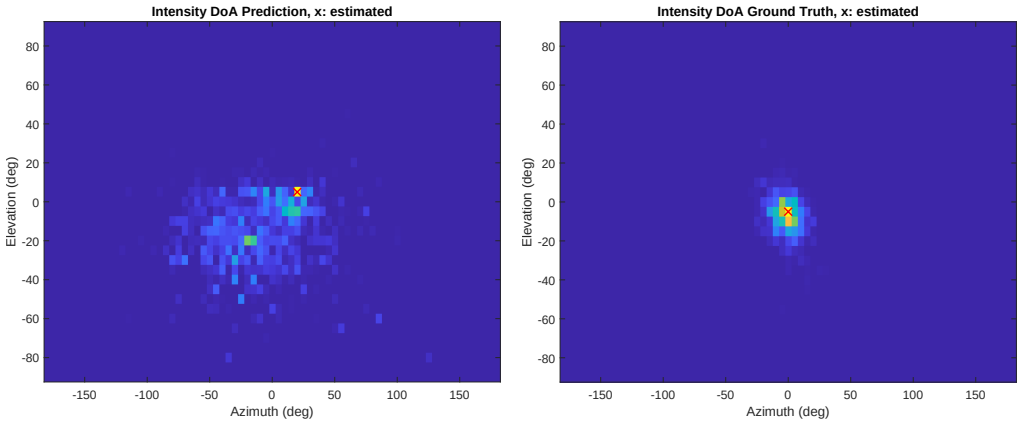
$$\Delta DOA^{2D\phi} = |\phi - \hat{\phi}| \tag{6.38}$$

The DOA-Est are derived from the unsmoothed intensity vector, as defined in Section 6.3.8, using the MATLAB library presented in [597]. The acoustic intensity measurements are sampled across time using a window length of 100 samples with an overlap of 50% and are used to compute histograms of their estimated DOAs, weighted by the magnitude of the vectors. DOA-Est are made on a vector of spherical grid points with a resolution of 5°. The grid locations associated with the greatest number of DOA estimates are assumed to represent the directions of the dominant sound sources and are determined based on Von-Mises peak-finding, presented in [598], which facilitates DOA estimates for a specified number of sources over the length of the given signal. Table 6.7 details the DOA error for all examples given in this section.

Figure 6.19 shows the the DOA-Est histograms for a pink noise burst spatialised to $\theta = \phi = 0°$ for both the upmixed and ground truth B-format signals.

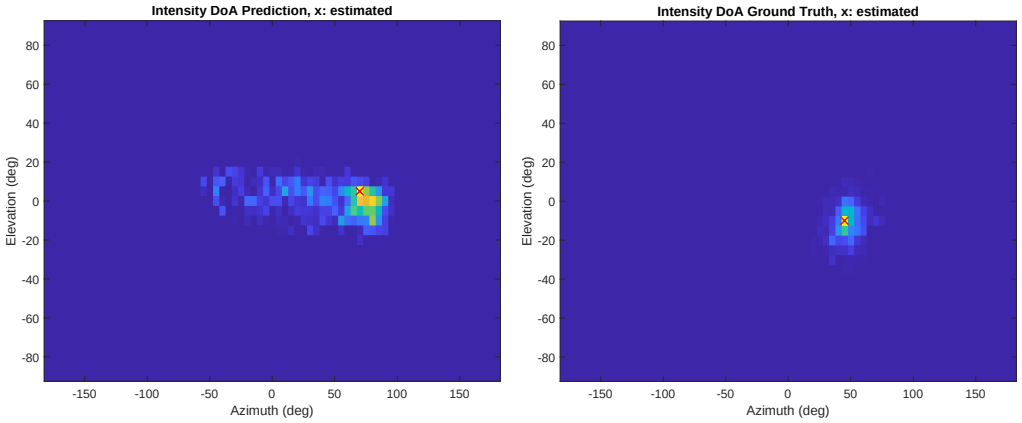
Ground truth location (°) θ, ϕ	Ground-truth DOA-Est (°) θ, ϕ	Predicted DOA-Est (°) θ, ϕ	DOA error θ (°) $\Delta DOA^{2D\theta}$	DOA error ϕ (°) $\Delta DOA^{2D\phi}$	Total DOA error (°) ΔDOA^{3D}
0,0	0,-5	20,5	20	5	22.34
45,0	45,-10	70,5	25	15	29.07
90,0	95,-5	80,0	15	5	15.79
135,0	140,-5	25,-10	115	5	113.55
180,0	-175,0	-25,-5	150	5	149.62
-135,0	-135,0	-90.-5	45	5	45.22
-90,0	-90,0	-90,0	0	0	0.00
-45,0	-40,-10	-110,-5	70	5	69.47
0,90	0,90	-40,70	40	20	20.00
45,65	40,60	95,40	55	20	39.07
0,45	5,35	20,45	15	10	15.19
18,18	20,10	110,15	90	5	87.42
18,-18	15,-25	105,-25	90	0	79.71
0,-45	-5,-45	10,-40	15	5	12.11
45,-65	45,-70	85,-55	40	15	23.07
0,-90	0,-90	-15,-65	15	25	25.00

Table 6.7: DOA errors derived from DOA historgram estimates for upmixed B-format signals when compared to ground truth B-format signals.



(a) DOA estimates for upmixed B-format signal (b) DOA estimates for ground-truth B-format signals

Figure 6.19: Directional grid of DOA estimates resulting from the time-sampled intensity vectors of the (a) upmixed B-format signals and (b) ground truth B-format signals. Stereo input source is a 3s pink burst spatialised to $\theta = \phi = 0°$ using IRs from the AB_omni_40 set.



(a) DOA estimates for upmixed B-format signal (b) DOA estimates for ground-truth B-format signals

Figure 6.20: Directional grid of DOA estimates resulting from the time-sampled intensity vectors of the (a) upmixed B-format signals and (b) ground truth B-format signals. Stereo input source is a 3s pink burst spatialised to $\theta = 45°$, $\phi = 0°$ using IRs from the AB_omni_40 set.

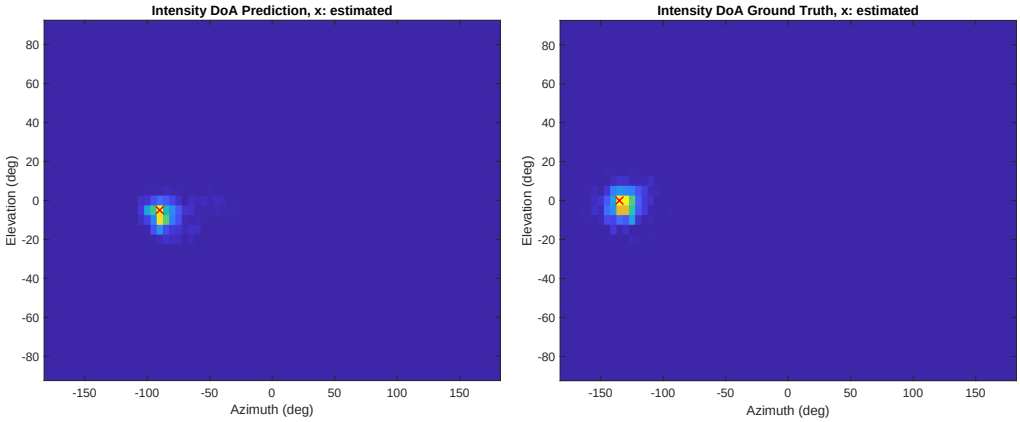
As well as a Total DOA error of 22.34°, there is also evidence of greater fluctuations and variability in the DOAs estimated for the upmixed B-format signal. This infers that there exists some spatial instability between time frames within the predicted directional parameters, where the predicted values cause DOA estimates to fluctuate between time-frames to a greater extent than is present in the ground truth data. Figure 6.20 shows the results for a pink noise burst spatialised at $\theta = 45°$, $\phi = 0°$, which resulted in a DOA error of 29.07°. There is also similar evidence of spatial instability with respect to the spatialisation derived from the predicted parameters. However, in this instance the instability seems to be much more localised to the horizontal plane within ±20° elevation with the DOA-Est having a higher concentration around the dominant peak, which suggests a more stable spatial image.

Figure 6.21 shows the DOA-Est for a pink noise burst at $\theta = \pm 135°$, $\phi = 0°$. Although these positions are symmetric about the median plane, there are clear differences between the results for each. For $\theta = -135°$, shown in Figure 6.21a, the DOA error is 45° and the predicted parameters have been unable to produce

an upmix where the source is positioned to the rear of the receiver but instead positioned it at the extent capable of traditional stereo directional estimates. The source appears, however, to be spatially stable, as evidenced by the high concentration of DOA-Est within a smaller number of grid locations. However, a source position of $\theta = 135°$, shown in Figure 6.21c, not only results in a higher DOA error of 113.5°, but additionally, results in greater fluctuations in DOA-Est that span across the entire frontal region, which may be indicative of spatial artefacts. Also of note is the high secondary peak at $\theta = -85°$, $\phi = -10°$ indicating that the source may be perceived as fluctuating between those two positions, which was confirmed by informal listening. For these positions, symmetric about the median plane, the predicted spatial parameters have failed to result in any DOA-Est to the rear of the receiver, which means the network was unable, in this instance, to predict directional parameters that result in the sources being remapped to the rear by the upmix process.

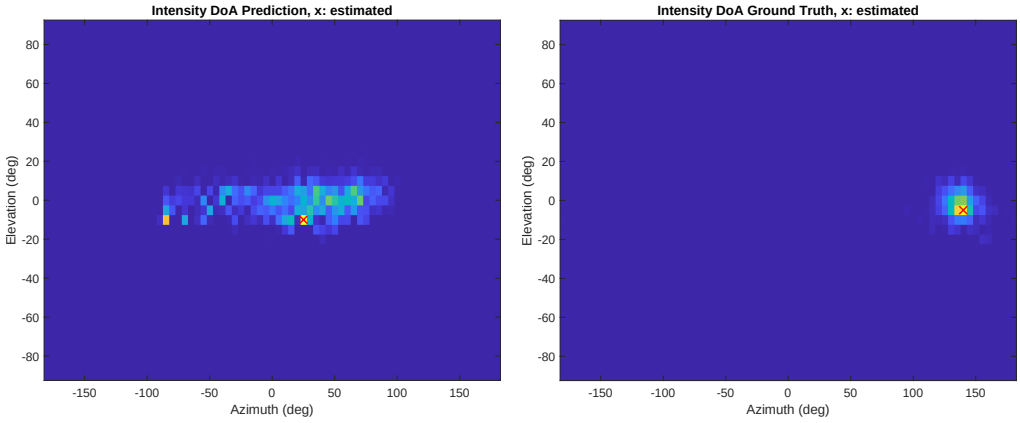
For a pink noise burst spatialised at θ =180°, $\phi = 0°$, Figure 6.22 shows the predicted parameters appear to cause a front/back reversal in the upmix process, which causes the upmixed source to again be placed frontal to the listener with the DOA-Est being concentrated around $\theta = -25°$. It should be noted that this does not mean that the model is incapable of predicting parameters in the range of $90° < \theta < -90°$ Instead, it cannot do so with enough consistency over the time-frequency tiles containing the spectral content of the source as to ensure enough of the spectral energy of the source is remapped to the correct positions in order for the source to be perceived as coming from that position.

For pink noise bursts spatialised on the horizontal, the best performance, as shown in Figure 6.23, appears to be for sources positioned at ±90° with DOA errors of 15.79° and 0.0°, for +90° and −90° respectively. There also appears to be much less variability in the DOA-Est than for the other evaluated positions, which is again evidenced by the concentration of DOA-Est within fewer grid locations. This could be due to the inter-channel differences for the stereo signal, and thus also for the extracted features from the stereo signal, being greatest at ±90° and enabling the network to more easily approximate the mapping function



(a) DOA estimates for upmixed B-format signal

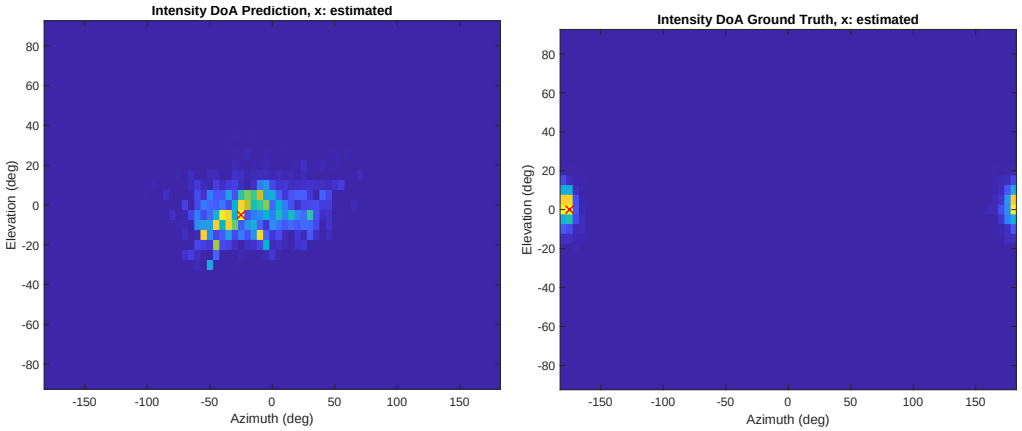
(b) DOA estimates for ground-truth B-format signals



(c) DOA estimates for upmixed B-format signal

(d) DOA estimates for ground-truth B-format signals

Figure 6.21: Directional grid of DOA estimates resulting from the time-sampled intensity vectors of (a) and (c) the upmixed B-format signals and (b) and (d) the ground truth B-format signals. Stereo input source is a 3s pink burst spatialised to (a), (b) $\theta = -135°$, $\phi = 0°$, and (c), (d) $\theta = 135°$, $\phi = 0°$, using IRs from the AB_omni_40 set.

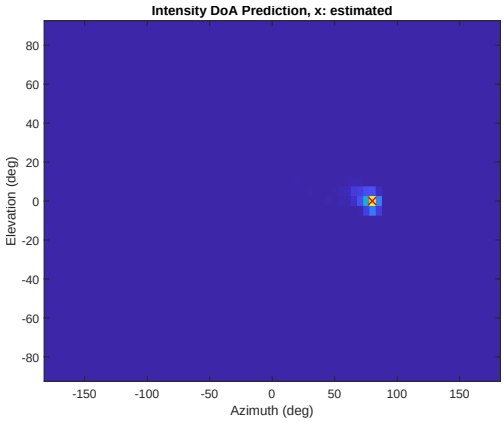


(a) DOA estimates for upmixed B-format signal (b) DOA estimates for ground-truth B-format signals

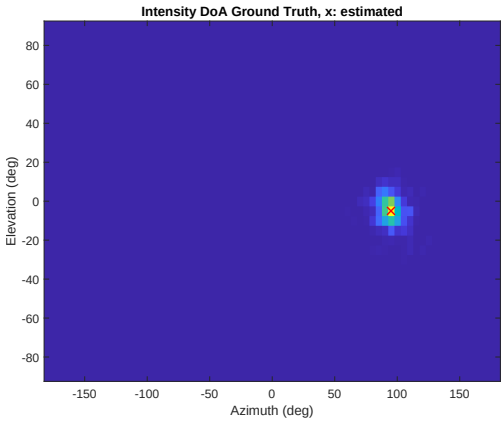
Figure 6.22: Predicted DOA estimates resulting from the time-sampled intensity vectors of the (a) upmixed B-format signals and (b) ground-truth B-format signals. Stereo input source is a 3s pink burst spatialised to $\theta = 180°$, $\phi = 0°$ using IRs from the AB_omni_40 set.

for these positions.

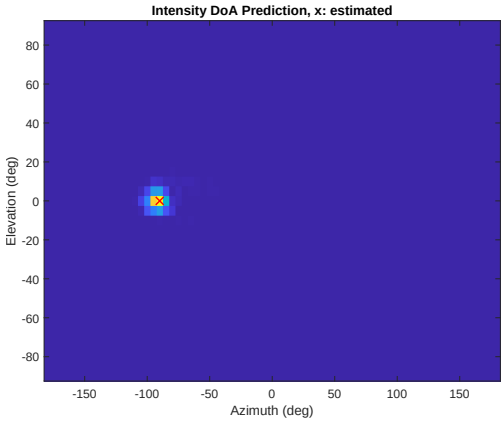
Figures 6.24 show the DOA-Est for sources located at elevation values of $\pm 90°$, which equates to directly above and below the receiver. Whilst the predicted parameters are able to facilitate the upmix algorithm in positioning sources at both positive and negative elevations their position is underestimated in both the above and below cases. The Total DOA error is $20°$ and $25°$ for elevations of $+90°$ and $-90°$ respectively. It is worth noting that for elevation values for $\pm 90°$, azimuth error does not impact the Total DOA error, as those elevation values represent the points on the sampled sphere where all azimuth values converge. A consequence of this is that if the DOA-Est of a source has a elevation value of $\phi = \pm 90°$ it is perceived to be coming from all azimuth directions. This can be seen in Figure 6.24 and is evidenced by the high DOA-Est count across all azimuth angles, indicated by the yellow band at the top and bottom of the grids for positions of $\phi = 90°$ and $\phi = -90°$ respectively. The results shown in Figures 6.25, 6.26, 6.27, confirms the pattern of underestimation with respect to elevation position, which is greatest for positions greater/less than $\pm 65°$. The



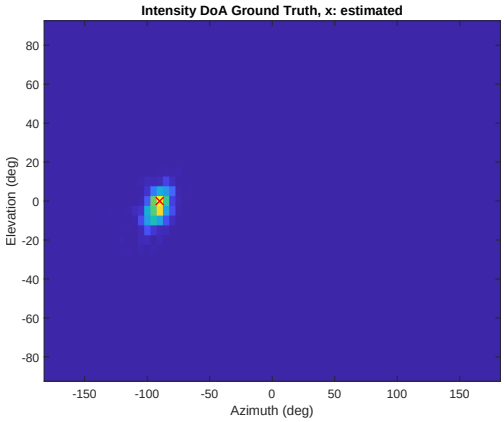
(a) DOA estimates for upmixed B-format signal



(b) DOA estimates for ground-truth B-format signals



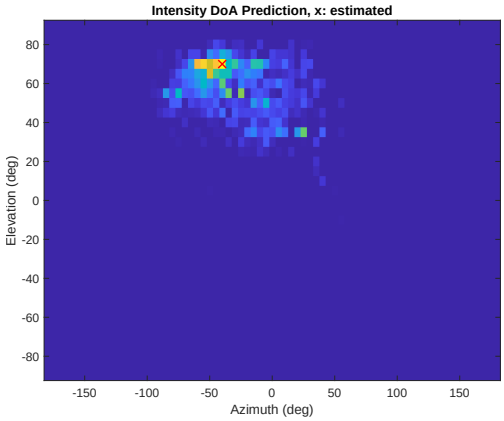
(c) DOA estimates for upmixed B-format signal



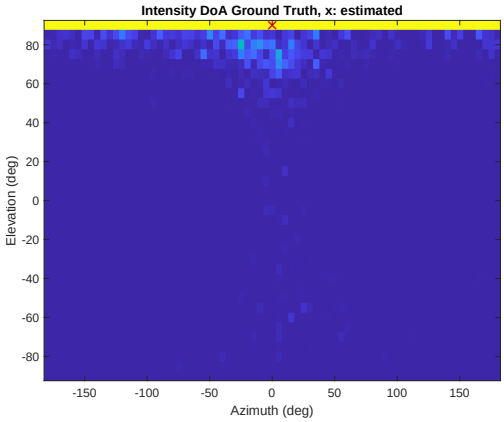
(d) DOA estimates for ground-truth B-format signals

Figure 6.23: Directional grid of DOA estimates resulting from the time-sampled intensity vectors of (a) and (c) the upmixed B-format signals and (b) and (d) the ground truth B-format signals. Stereo input source is a 3s pink burst spatialised to (a), (b) $\theta = 90°$, $\phi = 0°$, and (c), (d) $\theta = -90°$, $\phi = 0°$, using IRs from the AB_omni_40 set.

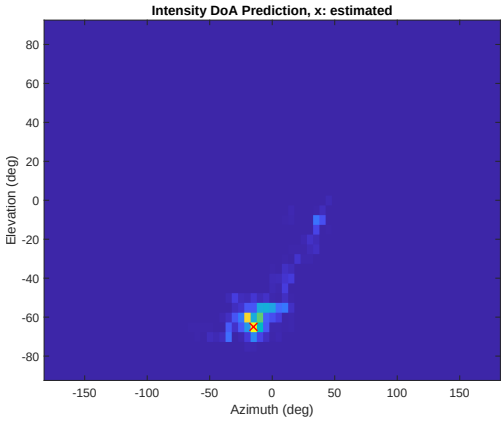
Total DOA error for the sources positioned at $\theta = 0°, \phi = \pm 45$ (Figure 6.25) is 19.08° and 12.11° for $+90°$ and $-90°$ respectively. Whilst source positions of $\theta = 45°, \phi = \pm 65$ resulted in Total DOA errors of 39.07° and 23.07 for $\phi = 65°$ and $\phi = -65°$, respectively, and source positions of $\theta = 18°, \phi = \pm 18$ yielded Total DOA errors of 87.42° and 79.71° for $\phi = 18°$ and $\phi = -18°$, respectively. It should also be highlighted that, as shown in Table 6.7, a large contributor to the error values for the elevated source positions are due to larger errors in the



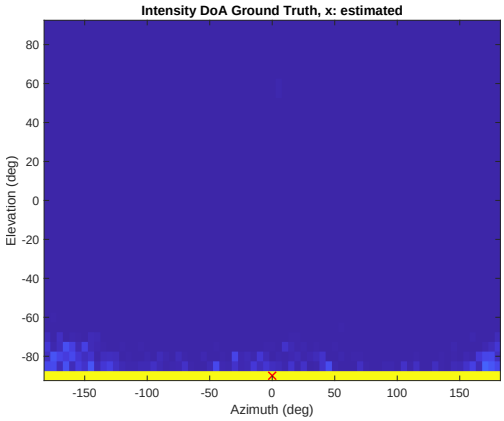
(a) DOA estimates for upmixed B-format signal



(b) DOA estimates for ground-truth B-format signals

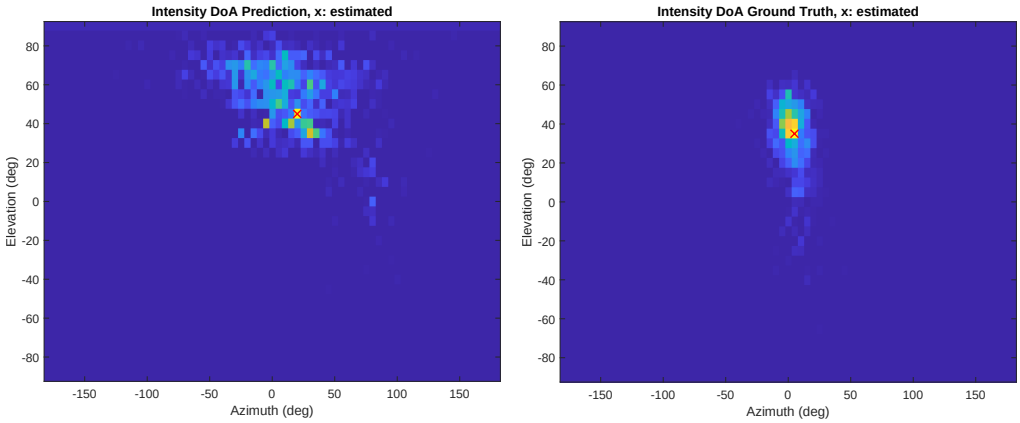


(c) DOA estimates for upmixed B-format signal



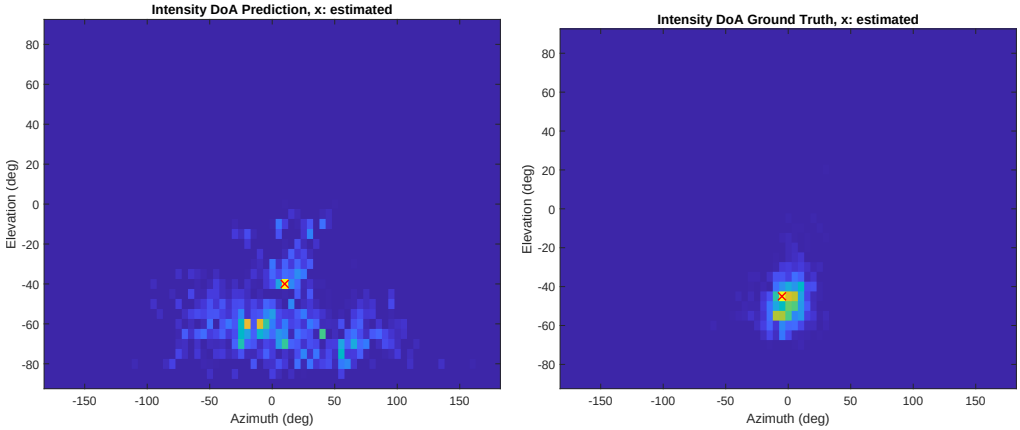
(d) DOA estimates for ground-truth B-format signals

Figure 6.24: Directional grid of DOA estimates resulting from the time-sampled intensity vectors of (a) and (c) the upmixed B-format signals and (b) and (d) the ground truth B-format signals. Stereo input source is a 3s pink burst spatialised to (a), (b) $\theta = 0°$, $\phi = 90°$, and (c), (d) $\theta = 0°$, $\phi = -90°$, using IRs from the AB_omni_40 set.



(a) DOA estimates for upmixed B-format signal

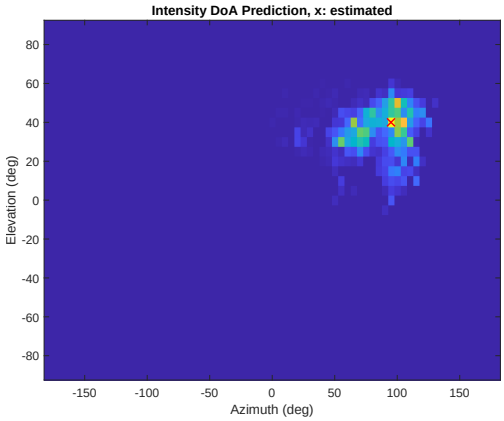
(b) DOA estimates for ground-truth B-format signals



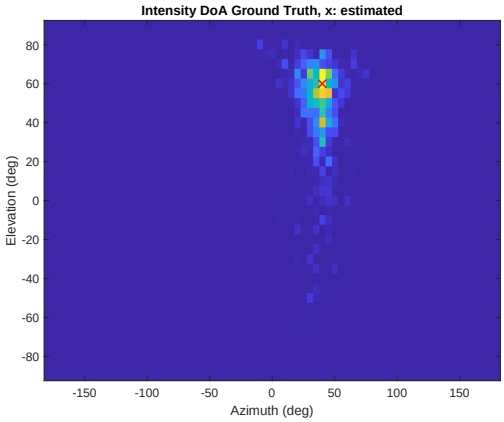
(c) DOA estimates for upmixed B-format signal

(d) DOA estimates for ground-truth B-format signals

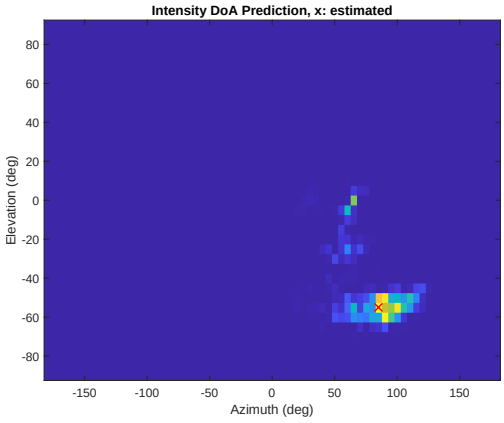
Figure 6.25: Directional grid of DOA estimates resulting from the time-sampled intensity vectors of (a) and (c) the upmixed B-format signals and (b) and (d) the ground truth B-format signals. Stereo input source is a 3s pink burst spatialised to (a), (b) $\theta = 0°$, $\phi = 45°$, and (c), (d) $\theta = 0°$, $\phi = -45°$, using IRs from the AB_omni_40 set.



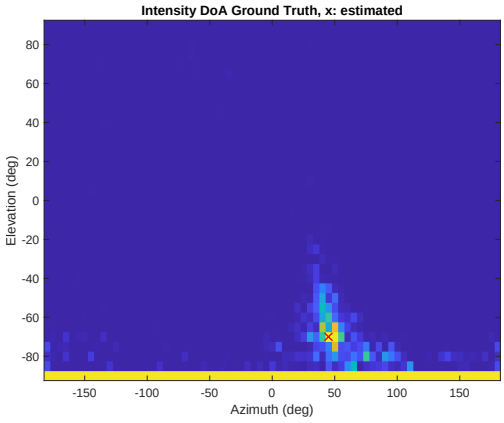
(a) DOA estimates for upmixed B-format signal



(b) DOA estimates for ground-truth B-format signals

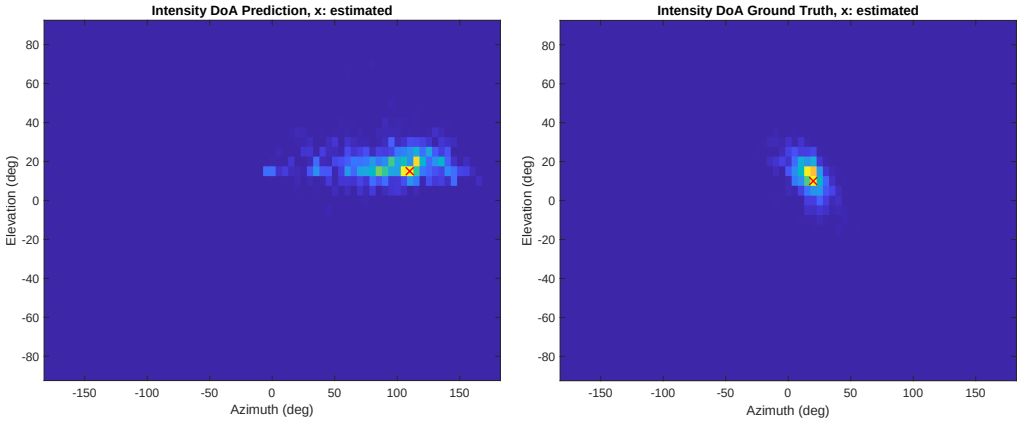


(c) DOA estimates for upmixed B-format signal



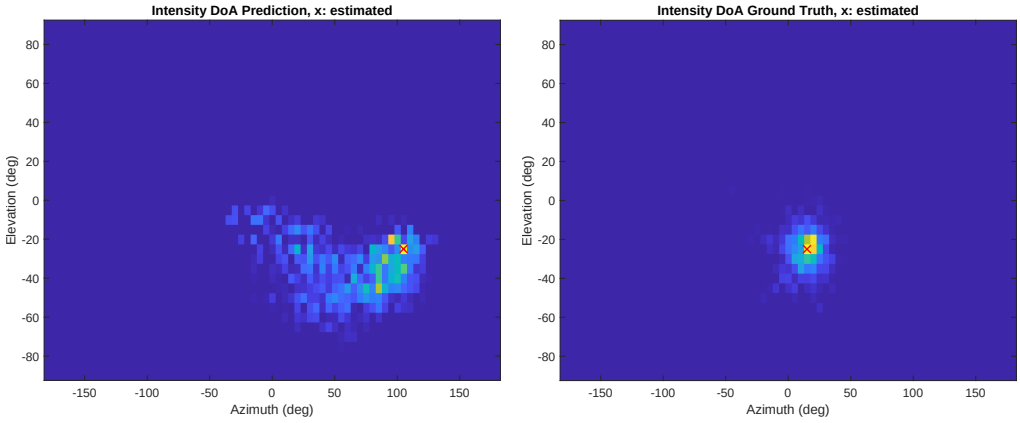
(d) DOA estimates for ground-truth B-format signals

Figure 6.26: Directional grid of DOA estimates resulting from the time-sampled intensity vectors of (a) and (c) the upmixed B-format signals and (b) and (d) the ground truth B-format signals. Stereo input source is a 3s pink burst spatialised to (a), (b) $\theta = 45°$, $\phi = 65°$, and (c), (d) $\theta = 45°$, $\phi = -65°$, using IRs from the AB_omni_40 set.



(a) DOA estimates for upmixed B-format signal

(b) DOA estimates for ground-truth B-format signals



(c) DOA estimates for upmixed B-format signal

(d) DOA estimates for ground-truth B-format signals

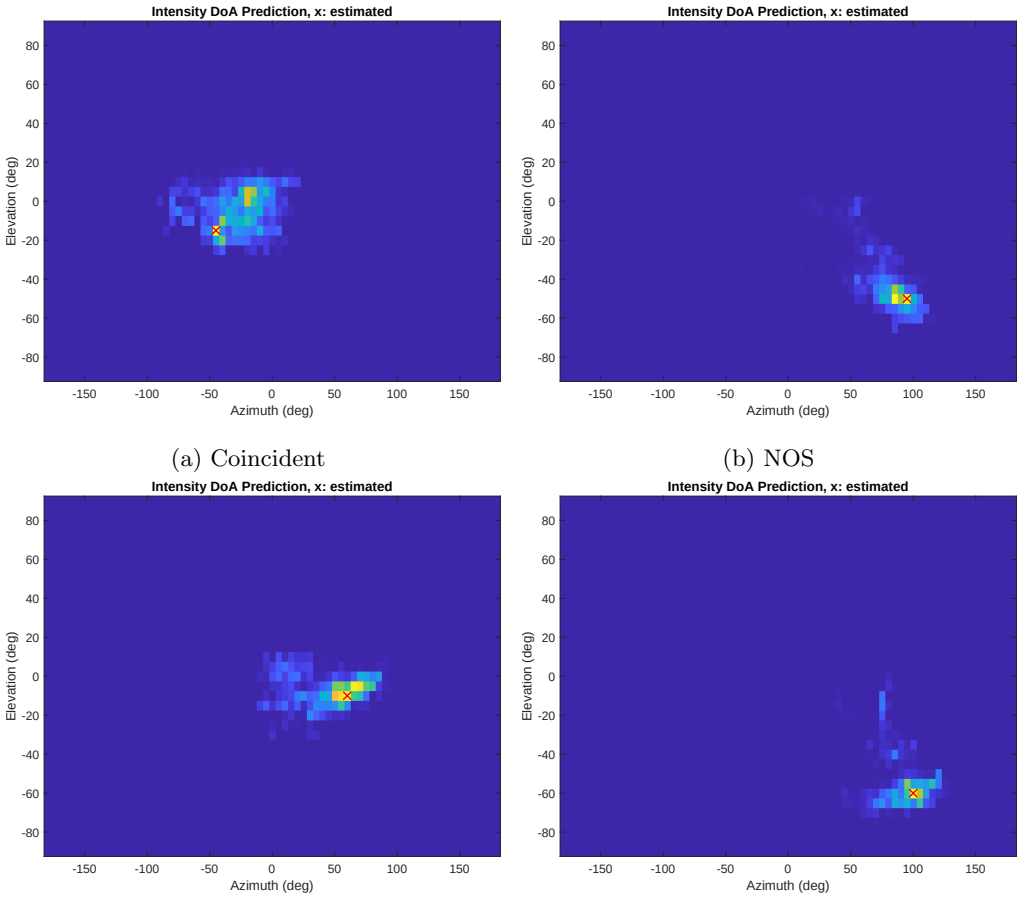
Figure 6.27: Directional grid of DOA estimates resulting from the time-sampled intensity vectors of (a) and (c) the upmixed B-format signals and (b) and (d) the ground truth B-format signals. Stereo input source is a 3s pink burst spatialised to (a), (b) $\theta = 18°$, $\phi = 18°$, and (c), (d) $\theta = 18°$, $\phi = -18°$, using IRs from the AB_omni_40 set.

azimuthal direction, with the results for elevation direction in isolation being within 25° of the DOA-Est resulting from the ground truth. In some cases, the DOA-Est for elevation appear closer to the truer intended source direction than those derived from the ground truth B-format signals. However, this could be due to the directional positions being explicitly specified during the upmix process and those time-frequency tiles being encoded as point sources. Additionally, the DOA-Est derived from the ground-truth signals may be influenced by small positional differences with respect to the intended and actual relative source and receiver positions at the IR capture stage, which would then have an impact on the positional encoding of sources using those IRs.

From these preliminary results it appears that whilst the model has begun to learn a mapping function for lateral position, it has not been able to approximate the required mapping function for sources to the rear of the receiver. It does, however, perform better than the RMSE loss values detailed in Table 6.4 would suggest. There are two possible reasons that could be hypothesised as to why the network has failed to learn a front/rear source mapping. Firstly, is that the input features do not contain the required information to adequately differentiate between front and rear source positions and different, or additional, input features are required. Secondly, is that a more suitable training strategy is required with respect to optimisation of the loss function, which would both investigate whether the model being better optimised results in more accurate estimations of frontal azimuth positions and whether better optimisation would result in the model learning a more accurate mapping function to differentiate between front and rear source positions. The results also suggest that the model has began to learn an approximate mapping function for source elevation, which results in upmixed sources being correctly positioned at either positive or negative elevation values, although results show larger error values for source positions directly above or below the receiver and exhibits, in some cases, spatial instability evidenced by large variability in DOA estimates.

To get an indication of the model's performance on unseen microphone configurations and the effect this would have on source position in a subsequent

upmix, pink noise was spatialised using IRs from the other stereo configurations listed in Table 6.2 and then upmixed to B-format. Figures 6.28, 6.29, and 6.30 show the DOA-Est for a sources at $\theta = 45°$, $\phi = -65°$, and $\theta = 90°$, $\phi = 0°$, and $\theta = 135°$, $\phi = 0°$, respectively, upmixed from Coincident, NOS, Blumlein, and AB_cardioid_40 configurations. Ground truth DOA estimates can be found for the respective source locations in Figures 6.26d, 6.23b, and 6.21d. Results show that those configurations containing appreciable ICTDs between microphone signals, such as spaced and near-coincident configurations, produce lower Total DOA error rates and appear to result in the remapping of sources in the elevation plane, which results in a Total DOA error comparable to that of the examples using the configuration on which the model was trained. This suggests that the current model has approximated a mapping function that relies more on temporal features than on features related to level in order to differentiate different elevation values as well as lateral positions.



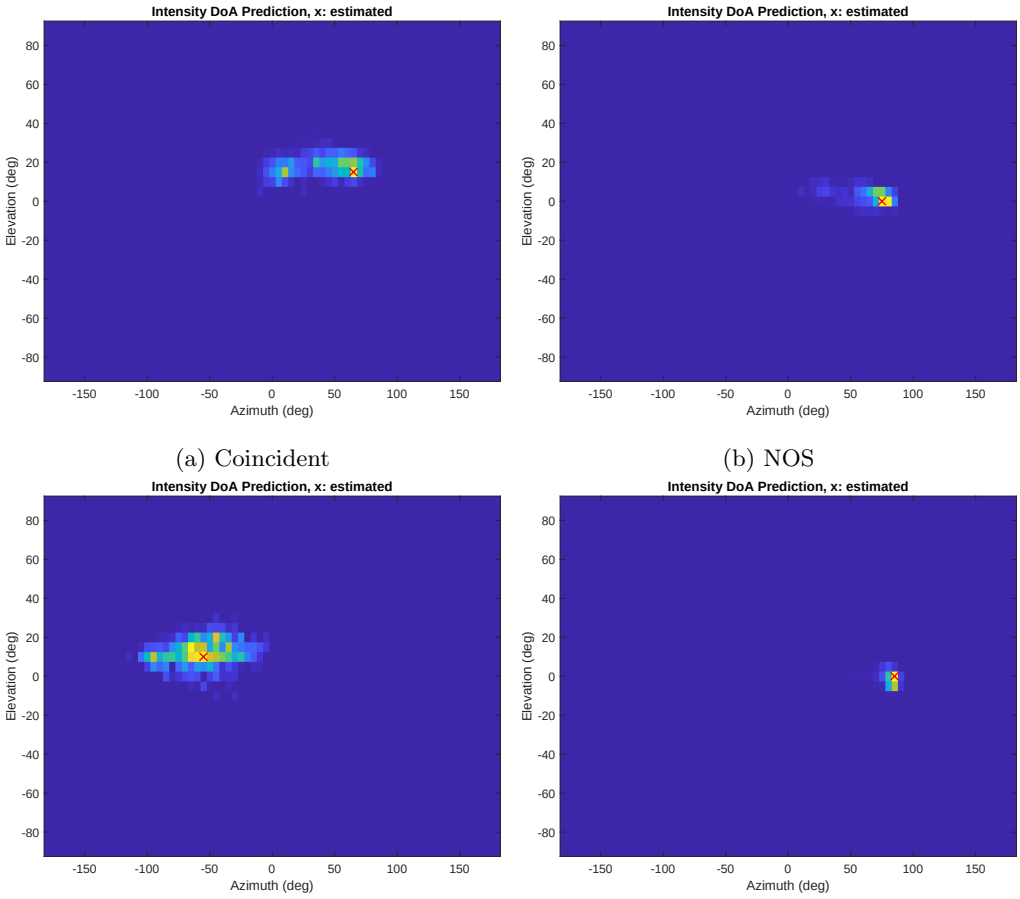
(a) Coincident

(b) NOS

(c) Blumlein

(d) AB_cardioid_40

Figure 6.28: Directional grid of DOA estimates resulting from the time-sampled intensity vectors of the upmixed B-format signals resulting from a stereo input source containing a 3s pink burst spatialised to $\theta = 45°$, $\phi = -65°$, using IRs from the (a) Coincident, (b) NOS, (c) Blumlein, and (d) AB_cardioid_40 set.



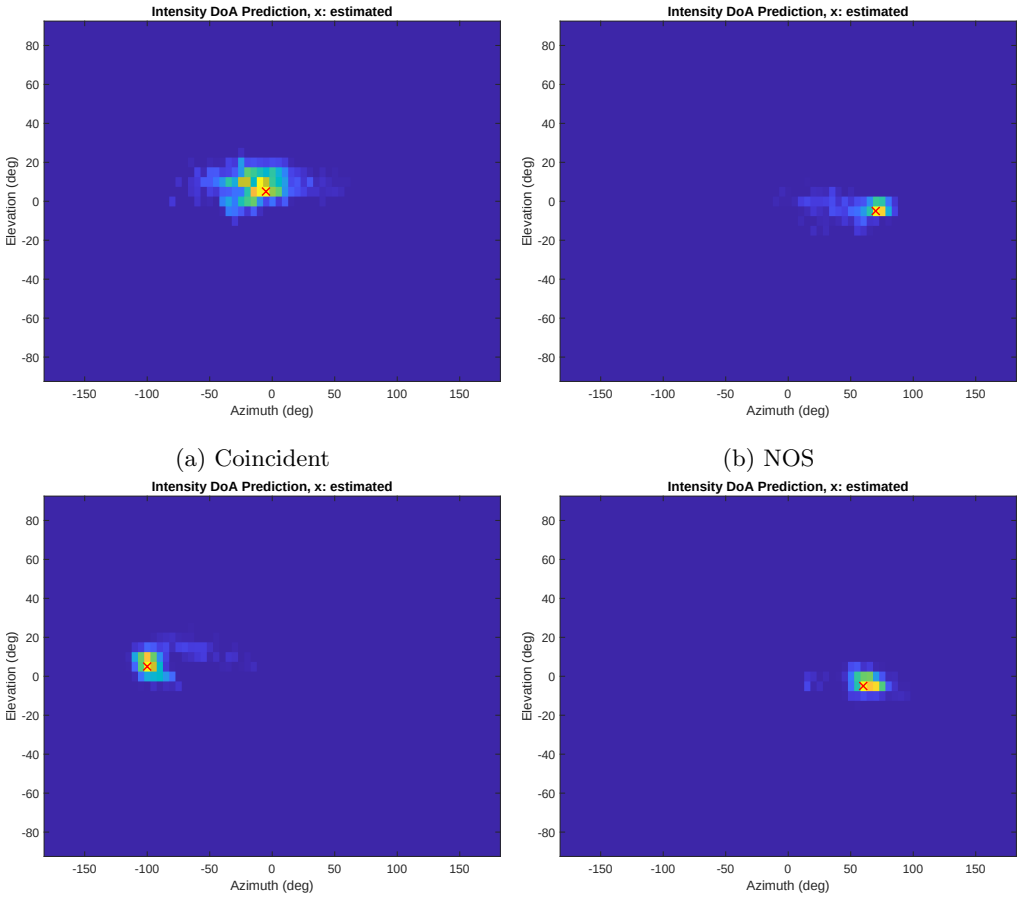
(a) Coincident

(b) NOS

(c) Blumlein

(d) AB_cardioid_40

Figure 6.29: Directional grid of DOA estimates resulting from the time-sampled intensity vectors of the upmixed B-format signals resulting from a stereo input source containing a 3s pink burst spatialised to $\theta = 90°$, $\phi = 0°$, using IRs from the (a) Coincident, (b) NOS, (c) Blumlein, and (d) AB_cardioid_40 set.



(a) Coincident

(b) NOS

(c) Blumlein

(d) AB_Cardioid_40

Figure 6.30: Directional grid of DOA estimates resulting from the time-sampled intensity vectors of the upmixed B-format signals resulting from a stereo input source containing a 3s pink burst spatialised to $\theta = 135°$, $\phi = 0°$, using IRs from the (a) Coincident, (b) NOS, (c) Blumlein, and (d) AB_cardioid_40 set.

6.9 Summary

This chapter detailed the development, investigation, and evaluation of a DNN trained to predict directional and diffuseness parameters for 360° space, using input feature vectors extracted from stereo signals. Relevant background was presented with respect to traditional methods of stereo upmixing with a specific focus on methods for directional estimations of signal components and direct-diffuse signal decomposition. Limitations of current methods were discussed including the reliance on amplitude panned material and how this could present challenges when applying current systems to stereo scenes recorded with spaced pairs. Additionally, it was highlighted that the lateral nature of current stereo directional estimation methods may result in erroneous spatial representations for sources not positioned in front of the recording array and an example was given to illustrate this.

A novel dataset of IRs was then presented that contained IRs for all loudspeakers arranged according to a 50-point Lebedev quadrature sampled sphere in both Ambisonic and stereo format. The IRs were 9 stereo configurations covering spaced, coincident, and near coincident methods, as well as spherical harmonic components up to 4th order, and 32 channels from the Eigenmike, a rigid sphere spherical microphone array. These were convolved with sound events from the NIGENs audio dataset to create equivalent stereo and Ambisonic scenes that could be used as training data for a NN. A description of DirAC was then presented which detailed the process of target feature extraction from the synthesised B-format signals using DirAC analysis to derive time-frequency directional and diffuseness parameters. Following on from this, the chosen input features were discussed along with details of the feature extraction pipeline.

The architecture for the proposed network, referred to as MuCh-Res-U-Net, was presented along with relevant background on key aspects of the architecture including its use of a U-Net backbone along with residual connections within both the encoding and decoding blocks. The original U-Net architecture was also presented as the baseline for this study. Details of the training methodology

and experimental set up were then given, including the rationale for only using a single stereo configuration in order to first develop the system within a simplified problem space. The method for hyperparameter selection was discussed as well as the optimizer, learning rate schedule, and loss function that was used for network optimisation.

Results were then presented based on the performance of the model on the validation and test sets. Whilst the results provide evidence that the general architecture and optimisation strategy has the potential to perform the chosen prediction task, several shortcomings and challenges were highlighted. The optimisation strategy did result in a reduction in the loss function, however the resulting MSE value was still high in relation to the task. Despite this, when visualising the data it became apparent that the network was able to generalise and predict the directional and diffuseness parameters relating to the time-frequency tiles associated with the sound source present in the scene. However, it generalised poorly with respect to predicting the parameters for time-frequency tiles associated with the ambient background noise. Since the ambient background noise occupied a large proportion of the spectrum, this would go some way to explaining the high loss value as each time-frequency tile is treated as an individual regression and is equally weighted within the loss function. It is suggested that a perceptually motivated audio domain loss function, such as those presented in [592], may perform better given the domain specificity of the task.

Observing the predictions of a model that was allowed to overfit, it was noted that although the overfitting caused the validation loss to increase, this was largely due to the predicted parameter values for the ambient portion of the spectrum diverging from zero and appearing to be randomly distributed in such a way that the predicted values for each time-frequency tile was further away than their ground truth. However, given the nature of the ambient and/or diffuse background noise it does not necessarily matter whether the predicted directional parameters associated with the ambient portion of the spectrum are closely aligned with the ground truth, and the reasons for this are two fold.

Firstly, the time-frequency tiles associated with the background noise are likely to be associated with higher diffuseness indexes and will therefore contain less directional energy. A limitation of the DirAC analysis results in all time-frequency tiles being allocated a given direction even if the directional energy contained them would be imperceptible to the human hearing system. The information within the time-frequency tiles associated with higher diffuseness indexes would also have a greater proportion of their signal rendered according to the chosen diffuse rendering method and less by the chosen direct component rendering method, meaning that correct predictions for those time-frequency tiles are less important and would ideally contribute less to the loss function. It therefore suggested that future work may include the investigation of a spectral loss function that is inversely weighted by the ground truth diffuseness index. Secondly, given a suitably diffuse distribution it is unlikely to cause any perceptual distortions with respect to the overall spatial image as long as the predictions for the time-frequency tiles associated with the sound sources are accurate, as humans are generally unable to differentiate between two sets of random noise assuming both sets are taken from the same distribution. It may also be possible to set user defined thresholds where any time-frequency tiles associated with a diffuseness index above the threshold is dropped from the directional parameter prediction and directional reproduction. Although, listening tests were outside the scope of this initial investigation it is acknowledged that they would be needed to verify the perceptual quality of the system and any subsequent optimisation.

Lastly, two time-frequency stereo upmixing pipelines were proposed that utilised the predicted directional and diffuseness parameters to facilitate novel methods of stereo upmixing. The novelty is in the ability for the system to predict directional parameters around a 360° space in both the horizontal and median planes, compared to traditional stereo upmixers where the spatial remapping of directional components is limited to the horizontal plane and generally to loudspeakers within the frontal portion of the array. The first upmix pipeline was based on time-frequency parametric spatial audio systems, using the DirAC pipeline as an illustrative example. The directional and diffuseness estimation

stage is replaced with our proposed network which can predict the metadata used to facilitate direct-diffuse decomposition and time-frequency directional component extraction and remapping. This being suitable for application to arbitrary 2D and 3D loudspeaker arrays. The second method utilises the predicted time-frequency directional parameters to extract and weight frequency components to target spherical harmonic components using a single omnidirectional signal from a spaced pair for spatial reconstruction. This allows for existing stereo material to be upmixed and stored in B-format in a way that approximates the spatial characteristics that would have been present at the time of recording.

Preliminary benchmark evaluations were conducted for the B-format upmix pipeline using the Total DOA error between the upmixed B-format signals and the ground truth B-format signals using pink noise spatialised with the IR dataset that was also used to synthesise the training data. It was found that while the predicted directional parameters could not successfully facilitate the spatial remapping of time-frequency components such that objects were evaluated as being placed to the rear of the spatial scene, they were able to map to both positive and negative elevation values. These results provide evidence that there exists information within stereo signals that can be used to derive height information. Results also showed that there was greater variability in DOA estimates for azimuthal values when compared to those for elevation values. When presented with sources encoded into stereo configurations unseen during model optimisation, those configurations containing appreciable ICTDs between microphone signals, such as spaced and near-coincident configurations, tended to produce lower Total DOA error rates and, in the single example examined, were still able remapped the source to an elevation value which resulted in a Total DOA error comparable to that of the examples using the configuration on which the model was trained. This suggests that the current model has approximated a mapping function that relies more on temporal features than on features related to level in order to differentiate different elevation values as well as lateral positions. Both methods aim to improve upon the current stereo upmix methods which typically create a frontally focused sound field with additional surrounding ambience. The next

chapter is the final chapter of this thesis, the work from previous chapters will be consolidated and summarised along with highlighting the contributions of this thesis to the broader field.

Chapter 7

Conclusions and Further Work

This thesis presented a body of work with the overarching goal of investigating the challenges associated with spatial audio production for IMEs and developing novel methods of spatial audio production that may assist in the design of spatial soundscapes. This final chapter will summarise the work presented in previous chapters, drawing together key findings and observations. Following this, the hypothesis proposed in Chapter 1 will be revisited and evaluated based on the work presented. Finally, areas of future work will be discussed and how they might build upon the progress made by the the work within this thesis.

7.1 Thesis Summary

Chapter 2 began by introducing the fundamentals of sound and audio signal processing. This included an overview of how sound propagates through space and a selection of sound field encoding and reproduction methods relevant to the work presented in later chapters. An introduction was then provided of digital audio signal processing methods that underpin much of the technical work in this thesis, with a particular focus given to time-frequency analysis and processing. Chapter 3 defined immersion within the context of this thesis and what is considered as an IME. Additionally, it provided an overview of the common categories of IME and the role and contributions of spatial audio to the experience of the user.

Chapter 4 presented an investigation into sound design approaches for IMEs,

focusing on the perspectives of practitioners working in the field. The methods of data collection were described in addition to an explanation of the thematic analysis framework used to interrogate the data. The generated themes were then explored and followed by a discussion framed around common topics that emerged across the themes. The analysis highlighted several areas of potential research which were then subsequently discussed with a view to informing future work both within this thesis and the wider research community.

Chapter 5 built on some of the conclusions from Chapter 4 and presented the investigation, development, and evaluation of an early stage methodology for deriving audio metadata from objects within a 2D scene and then used this to facilitate automatic panning and candidate sound effect suggestions. A description of the system architecture was given and included details on the computer vision algorithm used, the inter-frame continuity check, and object trajectory and panning derivation, concluding with a proof of concept methodology for candidate sound effects suggestion using the BBC Sound Effects archive [25] as the target repository. The obtained results indicated that, for scenes with more than one sound object, a more robust method of ensuring inter-frame continuity is required to maintain consistent positional tracking of objects over subsequent frames if accurate panning data is to be derived. In principle, it was also confirmed that using object classification is a viable method to search through sound effects repositories in order to suggest candidate sound effects files. Given recent advances in language models, the simplistic search method used as a proof of concept could easily be improved upon and several options for this were proposed.

Chapter 6 continued the investigation into machine learning approaches to sound spatialisation, with the aim of addressing another of the challenges highlighted in the results of Chapter 4. Specifically, it addressed both the perceived lack of available spatial sound effects archives and the integration of stereo content within immersive media projects utilising spatial audio. In contrast to Chapter 5, which aimed to derive directional information for specific objects, the work in this chapter adopted a parametric time-frequency approach

to facilitate the development of a novel methodology to stereo upmixing using a neural network to predict 3D parametric features, specifically time-frequency spherical coordinates, and a diffuseness index for each time-frequency tile. A novel dataset of stereo, spherical harmonic, and multi-channel IRs was presented as well as a description of the recording equipment, available microphone array formats, and sound scene synthesis. Target features were extracted from the first-order Ambisonic scenes using DirAC analysis and the input features were derived from the equivalent stereo scenes. The optimisation and evaluation pipeline was described along with details of the architecture for both the baseline model and the proposed MuCh-Res-Unet, which utilised multi-channel output and residual connections.

The performance of the model was evaluated and, whilst it performed well on the time-frequency tiles relating to the sound source present in the scene, it was unable to generalise in relation to the time-frequency tiles relating to the ambient background noise. Limitations in the DirAC analysis method used for feature extraction were discussed and potential improvements to the data processing and model optimisation strategy were suggested. Finally, two example upmix pipelines were presented, where the proposed model could be used to facilitate stereo upmixing to arbitrary loudspeaker configurations using a DirAC-style pipeline, or, using only the directional parameters, could be used to extract and weight frequency components against target spherical harmonic components to affect a stereo to B-format upmix algorithm.

7.2 Contributions to the Field

The novel contributions to the fields of sound design for IMEs and machine learning approaches to sound spatialisation are as follows:

- Chapter 4 presented the first study to investigate the defining features of IMEs as a new experience format and the challenges associated with its production from the perspective of sound design practitioners working in the field.

- The results from Chapter 5 evidenced the creative affordances that a computer vision based system could provide by deriving sound object metadata to both spatially position objects using stereo panning and search large scale sound effects repositories to recommend candidate sound effects audio files for identified sound sources within a single frame

- The dataset of IRs presented in Chapter 6 is the first to contain spatial IRs for all loudspeakers arranged according to a 50-point Lebedev quadrature sampled sphere in both Ambisonic and stereo format.. Ambisonic data is available up to 4th Order and stereo data is provided for 9 common configurations of two microphones.. This presents the first dataset suitable for synthesising training data for machine learning approaches to stereo upmix algorithms that focus on source locations sampled across a sphere at a fixed distance and where the stereo signals have not been derived from existing Ambisonic material. The dataset is available at [550].

- Chapter 6 presented evidence that it is possible to predict 360° spatial parameters from a stereo signal using a machine learning approach trained on an appropriate dataset, in this particular case based on Ambisonic and stereo IRs obtained from a 50-point Lebedev quadrature sampled sphere of loudspeakers.

- Chapter 6 further presented evidence that the predicted spatial parameters can be used to facilitate a stereo to B-format upmix approach that does not require direct/diffuse signal decomposition and where individual time-frequency components can be spatially remapped from a two-channel stereo representation to positions on a 360° sphere in the spherical harmonic domain.

7.3 Restatement of Hypothesis

The hypothesis originally stated in Chapter 1, which has informed the work presented in this thesis, is now restated as follows:

Machine Learning approaches can be used to assist in addressing challenges associated with the sound spatialisation pipline for IMEs.

The research presented in Chapters 4. 5, and 6 supports the given hypothesis as these chapters ascertain specific challenges faced by practitioners, and then through the collection of an appropriate dataset and development of machine learning approaches present assistive methods for sound spatialisation. Chapter 4 identified key features of IMEs and challenges present within the context of designing spatial audio, from the perspective of those practitioners working within the field. It also highlighted key areas where research could have a positive impact on existing workflows and practices. These included, amongst others, the automatic spatialisation of objects based on their position in a visual scene, the lack of freely available sound effects libraries, and the integration of existing stereo content with spatial audio projects. The results from this investigation went on to inform the work done in Chapters 5 and 6 which showed how machine learning could be leveraged to aid in the spatialisation of audio content. Chapter 4 presented an early-stage methodology which showed in principle that sound can be spatialised based on the data derived from existing computer vision algorithms, although further work was needed to ensure continuity between subsequent video frames for scenes containing multiple objects. Further development of this approach was decided to be outside the scope of this thesis as much of it, at the time, was a computer vision optimisation problem. Chapter 6 then presented a neural network approach to time-frequency parameter predication that has the potential to be integrated into a number of different pipelines to facilitate an approach to stereo upmixing that moves away from a frontally-bias reproduction. Two upmix pipelins are then presented that allow scenes to be upmixed for reproduction and storage without the frontally-biased representation associated with existing stereo upmixers.

The confirmation of the original hypothesis demonstrates the value of this thesis to the wider field. The findings from the various experiments and studies that make up this thesis have suggested potential ways in which this research may

be continued and/or inform other work in the future. A number of suggestions for future research will be outlined in the following section.

7.4 Future Work

Whilst the results from Chapter 4 provide valuable insight into the how professional practitioners view and approach sound design for immersive media, it is acknowledged that although all participants were highly experienced, the small sample size limits the generalisability of the findings to the wider industry. The work carried out in this chapter could therefore be extended in two obvious ways. Firstly, a larger study could be carried out to see if the results remain consistent. This work was conducted in 2020 and given the speed of progress in this area a follow up study would no doubt be useful in ascertaining how practice in this area has evolved and developed, especially given the recent advances in both generative machine learning models and the audio capabilities of game engine technology. Secondly, an ethnographic-style study could be conducted similar to that in [369], which investigated radio production practice. This would enable the collection of firsthand data on current working practices and would not rely on participants having to take time out of their schedule to participate. This was the original plan when first desgining the study however a change of approach was required due to the COVID-19 pandemic and the associated restrictions put in place. Observational studies can also mitigate some of the issues associated with self reporting. Luff et al. [362] suggests there is often a difference between what participants report they do and what they actually do. This difference is due to some activities being performed on such a regular basis that they become second nature and may not be at the forefront of a participant's thoughts when asked about the subject. Similarly, practices deemed not important by the participant can also be recorded during observation allowing for a more informed and complete representation of current working practices. Given the speed at which the application of machine learning to areas of creative practice is developing, it may be beneficial to conduct a study specifically focused how practitioners view

the current machine learning trend, how it affects their practice, and how they view this quickly developing set of automated methods.

Given that the methodology presented in Chapter 5 for automatic panning and candidate sound effects files served as a proof of concept, there are several ways this could be continued and built upon. The poor performance on scenes containing multiple objects was caused by a difficulty in maintaining continuity between frames. This difficulty resulted from using an adapted object detection system, originally intended for use on single image, to instead process sequential video frames. It is proposed that utilising more recent object detection systems that have been developed specifically for video data may mitigate the need for a custom continuity check enabling the tracking and panning data derivation for multiple objects. This could also be integrated with systems that estimate the distance of objects within scenes and applied alongside an intelligent audio mixing system whereby the predicted distance of an object is then used as input into appropriate signal processing to simulate the predicted source distance. The candidate sound effects suggestion based on the classification of objects within the scene performed relatively poorly, often failing to identify sound effects files when the tags were not an exact patch for the predicted objects label. This was caused by the search method being based on string companion between the label of the predicted object and the tags associated with sound effects file within the repository. It is proposed that given the recent rise in the use of language models, particularly transformer-based Large Language Models (LLMs), such as GPTs [599], that one could investigate the possibility of fine-tuning a LLM on a dataset of sound effects labels enabling the retrieval of sound effects that have related, but not necessarily identical, tags to that of the classification labels of the detected objects. Finally, given that the proof-of-concept was investigated using 2D scenes to generate stereo panning data, the next logical step would be to investigate the use of computer vision systems designed for object detection within 360° video to generate panning data compatible with spatial/surround sound panners.

Given the number of variables inherent in neural network design and opti-

misation it would be difficult to try and provide an exhaustive review of how the research in Chapter 6 could be built upon, so instead some areas which are considered to potentially have the most impact are suggested. One continuation of the work would be to increase the spatial resolution of the IR dataset. Whilst the dataset is comprehensive with respect to the array configurations contained within in, the spatial density is relatively sparse, sampling only 50 locations over the entire sphere. In addition to the increase in discretely sampled positions, it could also include dynamically moving sources for all configurations. One of the main limiting factors in its use as training data for upmixing is that all the sources positions are static, something which rarely occurs in natural sound scenes. It would, however, be possible to simulate moving sources by interpolating between IRs over a specified period of time.

Given the performance with respect to predicted time-frequency tiles relating to the sound source within the scene, the next step would be to see how the system performs with scenes containing overlapping sources. Appropriate training examples can be generated using the existing sound synthesis framework described in Chapter 6 and if necessary the system can be fine-tuned or a new model trained. With respect to the optimisation of the model, all predicted parameters were given equal weighting when calculating the final loss, since it is established that humans possess more acute azimuth localisation compared to elevation. Developing a weighting strategy may be beneficial in designing a perceptually informed optimisation strategy. Similarly, the use of a perceptually motivated or audio specific loss function such as those presented in [592] may improve performance when compared to the results obtained using the MSE.

Given that time-frequency tiles with smaller magnitudes and/or higher diffuseness indexes will have less energy reproduced through direct component rendering, an optimisation method could be explored to perceptually weight different regions of the spectrum with respect to their contribution to the directional parameter loss. This would be beneficial as the accuracy of the directional parameters becomes less significant for those time-frequency tiles whose directional reproduction may have limited perceptual impact on the spatial impression of the

scene. This is particularly relevant for areas of the spectrum occupied by diffuse background noise. One such method would be to weight the time-frequency tile contribution to the loss based on a perceptual measure of their loudness, such as SPL dBA. Another potential method would be to weight them according to their diffuseness index, which in effect would create a mask similar to methods used for direct-diffuse decomposition [149, 497, 498, 500, 505]. These methods could also be combined with an attention mechanism that where the model may learn to distinguish which areas of the feature map are most relevant for the given target feature.

The model was trained using data synthesised from a single stereo configuration, further investigation would be to assess its performance on unseen configurations, such as those included in the presented IR dataset. An additional model could also be trained on data containing examples from multiple stereo configurations and the performance between the two compared to see what the impact including additional configurations might be on the ability of the model to generalise.

With a baseline model developed and evaluated, an important area of further work would be to objectively and perceptually evaluate the results of the model when integrated into upmix pipelines, such as those suggested in Chapter 6 Section 6.7. This could take the form of listening tests comparing the reproduction of the original B-format signals with those of the upmixed B-format signals. Likewise, comparisons between upmixed and original signals could be evaluated through metrics such as those proposed in [592] which include signal-to-noise ratio and error-to-signal ratio amongst others. These metrics may also be suitable to replace the MSE as the loss function.

Lastly, given that spatial parameters such as those predicted by the proposed model have proven effective input features for auditory scene classification models [115], it would be interesting to investigate how results obtained using spatial parameters predicted from a stereo signal compare to those results from study which use parameters derived directly from B-format signals.

7.5 Closing Remarks

The research presented in this thesis has investigated the application of machine learning to spatial audio production for IMEs, with the aim of directly addressing some of the challenges faced by sound design practioners working in the industry. The initial qualitative work identified key challenges and highlighted multiple areas where research could have a tangible positive impact on current workflows and practices. Those related to the area of sound spatialisation were then explored and novel methods developed for the generation of object panning data, and for stereo scene upmixing that address some limitations with current stereo upmixers, such as frontal-biased representations which may cause inaccurate spatial reproduction with respect to the spatial characteristics of the original captured scene. It is also acknowledged that the methods developed in this thesis do not solve the problems they address in their entirety, however, they do offer progress that can continue to be built upon. It is hoped that the research presented in this thesis will go to inform further work in the field, as the production of high quality, realistic, spatial audio can contribute to a more immersive and engaging experience. Storytelling is one of the great human traditions and sound has, and arguably always will, play an important role in that.

List of Acronyms

Δs	Distance between loudspeakers
Δf_{ERB}	ERB critical bandwidth
2D	two-dimensional
3D	three-dimensional
3DOF	Three-Degrees-Of-Freedom
5G	Fith generation technology standard for cellular networks
6DOF	Six-Degrees-Of-Freedom
ACN	Ambisonic Channel Numbering
ADC	Analogue to Digital Converter
AE	Autoencoders
API	Application Programming Interface
AR	Augmented Reality
AudioLDM	Audio Latent Diffusion Model
BBA	Binarual-based Audio
BRIR	Binaural Room Impulse Response
BBC	British Broadcasting Corporation
BBCsfx	British Broadcasting Corporation Sound Effects Archive
BRIR	binaural room impulse response
CBA	Channel-based Audio
CD	Compact Disc
CLAP	Contrastive Language-Audio Pretraining
CNN	Convolutional Neural Network
COCO	Common Objects in COntext (dataset)

COLA	Constant Overlap Add
CPU	Computer Processing Unit
CSV	Comma-separated Values
DAC	Digital to Analogue Converter
DAFx	Digital Audio Effects
DAW	Digital Audio Workstation
dB	Decibel
DBAP	Distance-based Amplitude Panning
DCASE	Detection and Classification of Acoustic Scenes and Events
DCGAN	Deep Convolutional GAN
DFT	Discrete Fourier Transform
DIN	Deutsches Institut für Normung
DirAC	Directional Audio Coding
DNN	Deep Neural Network
DOA	Direction of Arrival
DoF	degrees of freedom
DRC	Dynamic Range Compressor
DRR	Direct-to-reverberant-ratio
DTFT	Discrete Time Fourier Transform
EBU	European Broadcasting Union
ERB	Equivalent Rectangular Bandwidth
ESS	Exponential Sine Sweep
EQ	frequency Equalisation
FFT	Fast Fourier Transform
fps	Frames Per Second
FOA	First-order Ambisonics
GANs	Generative Adversarial Networks
GPS	Global Positioning System
GPU	Graphics Processing Unit
GCC-PHAT	Generalises Cross-Correlation Phase Transform

GPT	Generative Pre-trained Transformer
HATS	head and torso simulator
HL-MRFs	Hinge-Loss Markov Random Fields
HMD	Head Mounted Display
HOA	Higher Order Ambisonics
HRIR	Head-related Impulse Response
HRTF	Head-related Transfer Function
HTC	High Tech Computer Corporation
ICLD	Inter-channel Level Difference
ICTD	Inter-channel Time Difference
iDFT	inverse Discrete Fourier Transform
iFFT	inverse Fast Fourier Transform
ILD	Interaural Level Difference
IME	Immersive Media Experience
IoU	Intersection Over Union
IPD	Interaural Phase Difference
IR	Impulse Response
ITD	Interaural Time Difference
ITD_{max}	Maximum ITD
ITU	International Telecommunications Union
JND	Just Noticeable Differences
KEMAR	Knowles Electronic Manikin for Acoustic Research
L3DAS21	Learning 3D Audio Sources 2021
LB	Left-Back
LF	Left-Front
LFE	Low Frequency Effects
LLM	Large Language Model
LSTM	Long-Short-Term-Memory
MMA	Multichannel Microphone Array
MPEG	Moving Picture Experts Group
MSE	Mean Square Error

MR	Mixed Reality
MuCh-Res-UNet	Multichannel Residual UNet
Masked-Unmasked Ratio	MUR
N3D	Three-dimensional Full Normalised
NaN	Not a Number
N/A	Not Applicable
NIGENS	Neural Information processing group GENeral sounds (database)
NN	Neural Network
NOS	Nederlandse Omroep Stichting
OBA	Object-based Audio
OLA	Overlap Add
OPSI	Optimized Phantom Source Imaging
ORTF	Office de Radiodiffusion Télévision Française
OSC	Open Sound Control
PCA	Principle Component Analysis
PHAT	Phase Transform
PSL	Probabilistic Soft Logic
RAM	Random Access Memory
RB	Right-Back
RF	Right-Front
ReLU	Rectified Linear Unit
ResNet	Residual Net
RIR	room impulse response
RMS	Root Mean Square
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
RT60	Reverb Time 60 (the time taken for reverb energy to decay by 60 dB)
SADIE	Spatial Audio Domestic Interactive Entertainment (database)

SBA	Scene-based Audio
SELD	Sound Event Localisation and Detection
SSD	Single Shot Detection
STFT	Short Time Fourier transform
SN3D	Schmidt Semi-normalised
SIL	Sound Intensity Level
SPL	Sound Pressure Level
SRIR	Spatial Room Impulse Response
TAU-SRIR DB	TAmpere University Spatial Room Impulse Response DataBase
TDOA	Time Difference of Arrival
TV	television
UK	United Kingdom
VAE	Variational Autoencoders
VBAP	Vector Base Amplitude Panning
VR	Virtual Reality
WAV	Waveform Audio File Format
WFS	Wave Field Synthesis
XR	Extended Reality

List of Symbols

$\hat{\alpha}$	Estimated panning coefficient
α	Distance exponent
β	Ambisonic format signal
Δs	Distance between loudspeakers (m)
$\Delta\theta_S^w$	Angle between the WFS loudspeaker array and the wave front to be reproduced (°)
Δ	Amount of change in a value
$\delta(t)$	dirac delta impulse
$\delta[n]$	Kronecker delta function
ϵ	Control coefficient DBAP speaker leakage
ε	time-constant for E{·}
η_{max}	Maximum learning rate
η_{min}	Minimum learning rate
η_t	Learning rate for timestep $_t$
η	Learning rate
$\Gamma\varrho$	Omnidirectional and bidirectional polar pattern coefficients)
γ	Adiabatic gas coefficient
λ	Wavelength (m)
ω	Frequency in radians/second
Φ	Starting phase of a sinusoidal wave
$\hat{\phi}$	Estimated\predicted elevation angle (°)
ϕ_l	Elevation angle of loudspeaker (°)

ϕ	Elevation angle (°)
ΨG	Generalised cross-correlation
ΨP	Frequency domain generalised cross-correlation Phase Transform
$\hat{\Psi}$	Estimated panning index
ψP	$\mathcal{F}^{-1}\{\Psi P\}$
$\psi(m, \omega_k)$	Time-frequency diffuseness index
ρ	Density of the medium ($\text{kg} \cdot \text{m}^{-3}$
$\hat{\theta}$	Estimated\predicted azimuth angle (°)
θ_s	Azimuth angle of loudspeaker s (°)
θ_0	Azimuth angle of 0 °
θ	Azimuth angle (°)
ϑ	Exponential sine sweep
i_ϑ	Inverse filter for exponential sine sweep
ξ	Hop size for block based processing
ζ	Noise present in a system
A	Measure of amplitude of a sinusoidal wave
a_n	Amplitude of nth sine harmonic
a_0	DC offset
B	Binaural signals
b_n	Amplitude of nth cosine harmonic
$\mathbf{C}$	Ambisonic re-encoding matrix
c_{gas}	Speed of sound in gas ($\text{m} \cdot \text{s}^{-1}$)
c	Speed of Sound ($\text{m} \cdot \text{s}^{-1}$)
$\mathbf{D}$	Ambisonic decoder matrix. Equivalent to $\mathbf{C}^{-1}$
$\vec{D}$	Time-averaged time-frequency direction of arrival matrix
$\vec{d}$	Direct signal component
d^ψ	Diffuse signal component
d	Virtual microphone directivity factor for calculating Ambisonic loudspeaker signals
$\vec{E}(m, \omega_k)$	Time-frequency short-time averaged energy vector

$\mathrm{E}\{\cdot\}$	Short-time averaging operation
E_{gas}	Gas equivalent Young's Modulus
E	Young's Modulus ($\mathrm{N} \cdot \mathrm{m}^{-2}$)
$\mathcal{F}$	Fast Fourier transform
$\mathcal{F}^{-1}$	Inverse Fast Fourier Transform
f_{max}	Maximum valid frequency (Hz)
f_{min}	Minimum valid frequency (Hz)
f_A	Spatial aliasing frequency (Hz)
f_n	nth frequency component
f_s	Sampling frequency (Hz)
f	Frequency (Hz)
G_L, G_R	Left and Right loudspeaker gains
G_i	Gain of the i^{th} loudspeaker
g_s	Gain of loudspeaker s
$\mathcal{H}(x)$	Neural Network input -¿ output mapping function
h	Impulse response of a given system
H_l	HRTF pair for a given loudspeaker position
H	Transfer function of a given system
I	Sound Intensity ($\mathrm{W} \cdot \mathrm{m}^{-2}$
$\vec{I}_{raw}$	unsmoothed intensity vector
$\vec{I}$	Smoothed intensity vector
I_1, I_2, I_3	FOA time-frequency intensity channel matrices containd with $\vec{I}$
I_o	Reference sound intensity value of 10^{-12} $\mathrm{W} \cdot \mathrm{m}^{-2}$
$J(A, B)$	Jaccard Index of datasets A and B
$\mathbf{k}$	wave vector
k	Wave number in radians
$\mathcal{L}$	A linear-time invariant system
L	Filter length (samples)
m	mth time block of a signal
M	Blocksize (samples)

N_{mn}^{N3D}	N3D Normalisation strategy for amplitudes of spherical harmonics
N_{mn}^{SN3D}	SN3D Normalisation strategy for amplitudes of spherical harmonics
R	Gas Constant
r	Radial distance (m)
M	Molecular mass of the gas (kg $\cdot$ mol^{-1})
Pl	Fourier domain representation of measured sound pressure at the left ear
Pr	Fourier domain representation of measured sound pressure at the right ear
P_{mn}	Legendre function
$P(m, \omega_k)$	STFT domain representation of p
$\mathbf{p}$	Position of audio object as described by a linear combination of loudspeaker vectors and gain scaling factors
p	Pressure; sound pressure (Pa)
p_{rest}	Static sound pressure (Pa)
p_o	Reference sound pressure value of 20 μ Pa
$\hat{p}$	Instantaneous sound pressure (Pa)
$\Re$	Real component of complex data
$\mathcal{STFT}$	Short-Time-Fourier-Transform
$\mathcal{STFT}^{-1}$	Inverse Short-Time-Fourier-Transform
$\mathbf{S}$	Matrix of loudspeaker vectors
$S_L, S_R(m, \omega_k)$	Time-frequency representation of left and right stereo channels
$\mathbf{s_n}$	Positional vector for nth loudspeaker
$\bar{S}_i$	Fourier representation of $\bar{s}$
$\bar{s_j}$	j^{th} Sound source
T_K	Absolute temperature (Kelvin)
T	Total time length
t	Time point in continuous time

$\vec{U}(m.\omega_k)$	STFT domain representation of particle velocity
$\vec{X}'(m,\omega_k)$	STFT domain representation of the vector of B-format pressure gradient signals
x	Time domain representation
X	Frequency domain representation
xyz	Position in Cartesian space
$Y_k(\vec{v}_l)$	Decoding coefficient for Ambisonic channel k and loudspeaker l
Y_{mn}^{σ}	spherical harmonics of order m and degree n
$\hat{y}$	Neural network prediction
$\bar{y}$	Ground truth
Z_0	Characteristic impedance of air
z	Frequency (Barks)
$*$	Convolution operator

Appendix A

Appendix A Ethical Approval Documents

THE UNIVERSITY of York

Application Form for Physical Sciences Ethics Committee Approval

Advice for applicants on completing the form

Please ensure that the information provided is:

- *Accurate and concise*
- *Clear and simple and easily understood by a lay person*
- *Free of jargon, technical terms and abbreviations*

Further advice and information can be obtained from your departmental representative on the PSEC and at: http://www.york.ac.uk/admin/aso/ethics/cttee.htm

Please return completed (typed) form to your departmental representative via email to:

elec-ethics@york.ac.uk

Title of project: AI Driven Approaches to Soundscape Design for Immersive Environments

SECTION 1 DETAILS OF APPLICANTS

Details of principal investigator (name, appointment and qualifications)

Daniel Turner – PhD Student

Names, appointments and qualifications of additional investigators *(student applicants should include their project supervisor(s) here)*

Prof. Damian Murphy – Supervisor

Location(s) of project

Genesis 6 University of York Part online survey Some surveys may be conducted at subject's place of work

SECTION 2 FUNDERS

What is the funding source(s) for the project?

EPRSC iCase Studentship BBC R&D

Please answer the following:

(i) Does the express and direct aim of the research or other activity raise ethical issues?

YES		NO	X

(ii) Is there any obvious or inevitable adaptation of research findings to ethically questionable aims?

YES		NO	X

(iii) Is the work being funded by organisations tainted by ethically questionable activities?

YES		NO	X

(iv) Are there any restrictions on academic freedoms – notably, to adapt and withdraw from ongoing research, and to publish findings?

YES		NO	X

If you answered **Yes** to any of the above, please give details below:

SECTION 3 DETAILS OF PROJECT OR OTHER ACTIVITY

Aims (100 words max)

The project seeks to explore artificial intelligence driven approaches to soundscape design for immersive environments, such as those experienced within virtual reality applications. The aim is to create new tools to assist designers in creating the audio for these experiences using current artificial intelligence technologies in novel ways. This particular study is designed to obtain information on the practices, workflows, and tools currently used by professional sound designers working in the field to contribute to the design of an artificial intelligence assistive system.

Background (250 words max)

Artificial Intelligence techniques have long been used on a variety of audio tasks including, music composition, audio production, sound synthesis, and sound event detection and recognition. One area that is yet to be comprehensively explored is the use of artificial intelligence in the generation of soundscapes for immersive environments such as those experienced in a virtual reality content.

Immersive experiences, such as virtual reality and 360 video/audio experiences are becoming much more commonplace. Many of these experiences can be recreated simply using a set of headphones and a user's mobile phone. Designing the sound for these experiences adds additional work for the sound designer over and above that of standard stereo content. In order to build tools to assist in this task a detailed understanding is required of the working practices, workflows, and current tools used by professional sound designers who produce audio for these kinds of immersive experiences. This will allow the outcome of the research to be applied in industry contexts and will also allow any outputs to be tested against current practices.

Brief outline of project/activity (250 words max)

Information will be gathered via a variety of methods including online questionnaire, semi-structured interviews, and observations. This will allow for maximum reach with respect to maximising the number of available participants and will be unrestricted in terms of geographical location.

Semi-structured interviews will use the online questionnaire content as a starting point but will allow for prompted expansion and clarification to responses where appropriate. These will be conducted either in person or via skype calls. These will be captured via audio recordings.

Observations will take place at the participant's place of work and detailed field notes will be taken including type of project being worked on, software environment used, tools used and tasks undertaken/approaches used. This may also be combined with short ad-hoc interview style questions if appropriate for clarification of noted observations i.e. give reason for using a particular method for completing a specific task. These observations will also be captured via video or audio recordings.

Study design *(if relevant – e.g. randomised control trial; laboratory-based)*

If the study involves participants, how many will be recruited?
15+

If applicable, what is the statistical power of the study, i.e. what is the justification for the number of participants needed?

SECTION 4 RECRUITMENT OF PARTICIPANTS

How will the participants be recruited?

Participants will be recruited through email call, utilising contacts with BBC R&D and within the AudioLab.

Calls will also be posted on industry specific social media such as Facebook and LinkedIn

What are the inclusion/exclusion criteria?

Aged 18+
Undertake professional sound design work i.e payment for commercial productions

Will participants be paid reimbursement of expenses?	YES [ ]	NO [X]
Will participants be paid?	YES [ ]	NO [X]

If yes, please obtain signed agreement

Will any of the participants be students?	YES [ ]	NO [**X**]

SECTION 5 DATA STORAGE AND TRANSMISSION

If the research will involve storing personal data, including sensitive data, on any of the following please indicate so and provide further details (answers only required if *personal* data is to be stored).

Manual files	**Consent forms (only name/participant number on form)**
University computers	**Name, participant number, and voice and video recording**
Home or other personal computers	**Indirect access to above through secure Google Drive**
Laptop computers, tablets	**Indirect access to above through secure Google Drive**
Website	**Password protected Qualtrics survey**

Please explain the measures in place to ensure data confidentiality, including whether encryption or other methods of anonymisation will be used.

The only place that the participants name will be displayed alongside their ID number and any categorisation data (e.g. age/gender) is on the consent form. This will be kept in a locked environment and kept confidential by the investigator (Daniel Turner).

Voice recordings will be collected on a secured password protected device and then deleted off the device once transferred to a university managed password protected google drive.

Please detail who will have access to the data generated by the study.

It is intended to publish the data and therefore anonymised data and responses will be openly accessible. Any personal data regarding participants will only be accessible by the investigators.

Please detail who will have control of and act as custodian for, data generated by the study.

Daniel Turner

Please explain where, and by whom, data will be analysed.

The data will be analysed by Daniel Turner at the AudioLab Genesis 6 University of York

Please give details of data storage arrangements, including where data will be stored, how long for, and in what form.

All data will be kept electronically on Daniel Turner's university managed Google drive under password protection. Hard copies of content forms will be scanned and electronically stored on the same Google Drive, physical copies will then be destroyed.

Processed and analysed data will be kept in a suitable form for publication and presented in thesis, publications and related works. Raw capture data will be reviewed at the of the PhD and will either be moved to secure department network drive or destroyed.

SECTION 6 CONSENT

Is written consent to be obtained?

YES	X		NO	

If yes, please attach a copy of the information for participants
https://york.qualtrics.com/jfe/form/SV_eyzNfRdAr0krlOZ
If no, please justify

Will any of the participants be from one of the following vulnerable groups?

Children under 18	YES		NO	X
People with learning difficulties	YES		NO	X
People who are unconscious or severely ill	YES		NO	X
People with mental illness	YES		NO	X
NHS patients	YES		NO	X
Other vulnerable groups (if 'yes', please give details)	YES		NO	X

If so, what special arrangements have been made for getting consent?

SECTION 7 DETAILS OF INTERVENTIONS

Indicate whether the study involves procedures which:

Involve taking bodily samples	YES		NO	X
Are physically invasive	YES		NO	X
Are designed to be challenging/disturbing (physically or psychologically)	YES		NO	X

If so, please list those procedures to which participants will be exposed:

List any potential hazards:

List any discomfort or distress:

What steps will be taken to safeguard

(i) the confidentiality of information

(ii) the specimens themselves?

What particular ethical problems or considerations are raised by the proposed study?

None directly

What do you anticipate will be the output from the study? ***Tick those that apply:***

Peer-reviewed publications	X
Non-peer-reviewed publications	
Reports for sponsor	
Confidential reports	
Presentation at meetings	X
Press releases	
Student project	X

Is there a secrecy clause to the research? YES [ ] NO [X]
If yes, please give details below

SECTION 8 SIGNATURES

The information in this form is accurate to best of my knowledge and belief and I take full responsibility for it.

I agree to advise of any adverse or unexpected events that may occur during this project, to seek approval for any significant protocol amendments and to provide interim and final reports. I also agree to advise the Ethics Committee if the study is withdrawn or not completed.

Signature of Investigator(s):

D.Turner……

Date: 20/08/19……

Responsibilities of the Principal Researcher following approval

- If changes to procedures are proposed, please notify the Ethics Committee
- Report promptly any adverse events involving risk to participants

AI driven methods for Immersive Sound Design - Consent Form

Thank you for showing interest in contributing to our study! We are interested in understanding how artificial intelligence technologies may be used to assist sound designers in the creation of immersive soundscapes. In order to do this we require an understanding of the workflows, processes, tools, and challenges encountered by producers of immersive content.

During the interview you will be asked to answer a series of questions about your experience creating and engaging with immersive content. The interview will be recorded and later transcribed. Please be assured that your responses will be kept completely confidential and anonymised.

The interview should take you around 30 minutes (maximum) to complete. Your participation in this research is voluntary. You have the right to withdraw at any point during the study prior to publication, for any reason, and without any prejudice. If you would like to contact the Principal Investigator in the study to discuss this research, please e-mail Dan Turner at djt530@york.ac.uk.

By signing below, you acknowledge that your participation in the study is voluntary, you are 18 years of age or older, and that you are aware that you may choose to terminate your participation in the study at any time and for any reason.

Interviewee Name……..
Interviewee Signature……. Date……...

Appendix B

Appendix B Survey/Interview Guide

AI driven Approaches to soundscape design for immersive environments

Q2 Welcome to the research study! We are interested in understanding how artificial intelligence technologies may be used to assist sound designers in the creation of immersive soundscapes. In order to do this we require an understanding of the workflows, processes, tool, and challenges encountered by producers of immersive content. You will be asked to answer a series of questions about your experience creating and engaging with immersive content. Please be assured that your responses will be kept completely confidential.

The study should take you around 20 minutes to complete. Your participation in this research is voluntary. You have the right to withdraw at any point during the study prior to publication, for any reason, and without any prejudice. If you would like to contact the Principal Investigator in the study to discuss this research, please e-mail Dan Turner at djt530@york.ac.uk.

By clicking the button below, you acknowledge that your participation in the study is voluntary, you are 18 years of age or older, and that you are aware that you may choose to terminate your participation in the study at any time and for any reason.

Please note that this survey will be best displayed on a laptop or desktop computer. Some features may be less compatible for use on a mobile device.

○ I consent, begin the study (1)

○ I do not consent, I do not wish to participate (2)

End of Block: Informed Consent

Start of Block: Block 2

Q3 What is your current job title?

_____

Q4 How long have you been working professionally in the industry. Professionally in this instance means being paid for your services.

- ○ <1year (1)
- ○ 1-3 years (2)
- ○ 3-5 years (3)
- ○ 5+ years (4)

Q5 What type of media do you produce content for (tick all that apply)

- ☐ Television Programmes (1)
- ☐ Radio (2)
- ☐ Streaming media i.e Youtube (3)
- ☐ Advertisement/marketing (4)
- ☐ Film (6)
- ☐ Immersive Experiences (7)
- ☐ Other (5)

Display This Question:

If What type of media do you produce content for (tick all that apply) = Other

Q25 If other, please specify

_____

Q6 What are the main programme formats or genres that you produce content for?

_____

Q8 On a day to day basis, do you work predominately with?

- ○ Sound (e.g sound design, audio editing etc) (1)
- ○ Visuals (e.g video editing, colour grading etc) (2)
- ○ Equal mix of both (3)

Q9 Please define what the term "Immersive content" means to you?

_____

_____

_____

_____

_____

Q10 Please give a brief explanation as to what you would define as key features of "immersive content".

_____

_____

_____

_____

_____

Q26 For the rest of the this survey we use the following definition for immersion.

"Immersion is a phenomenon experienced by an individual when they are in a state of deep mental involvement in which their cognitive processes (with or without sensory stimulation) cause a shift in their attentional state such that one may experience disassociation from the awareness of the physical world." (I will put in reference).

As such, immersive content is content designed to elicit a state of immersion. We are specifically interested in the technologies used in order to cause the subjective sense of being surrounded or experiencing multisensory stimulation that may then lead to a state of immersion. Examples of such technologies are binaural audio and 360° video

Q11 Do you currently or have you in the past worked on producing immersive content

- ○ Yes (1)
- ○ No (2)

End of Block: Block 2

Start of Block: Sound Design

Q21 Briefly describe some typical previous projects you have worked on.

_____

_____

_____

_____

_____

Q19 During the post-production phase are there any specific tasks that would benefit from semi-automation when working with audio content?

Semi-automation in this context could mean assistance with labour intensive tasks in order to enable greater focus on creativity.

_____

_____

_____

_____

_____

End of Block: Sound Design

Start of Block: Immersive

Q12 How long have you been working on immersive productions?

- ○ <1 year (1)
- ○ 1-3 years (2)
- ○ 3-5 years (3)
- ○ 5+ years (4)

Q13 Briefly describe any immersive content you have been involved in producing.

_____

_____

_____

_____

_____

Q14 What are the key differences, if any, between producing audio for immersive content compared to non-immersive content?

_____

_____

_____

_____

_____

Q15 What challenges, if any, do you face in creating audio for immersive content? Specifically with reference to how these might differ from non-immersive content.

_____

_____

_____

_____

_____

Q16 During the post-production phase are there any specific tasks that would benefit from semi-automation when working with immersive audio content?

Semi-automation in this context could mean assistance with labour intensive tasks in order to enable greater focus on creativity.

_____

_____

_____

_____

_____

End of Block: Immersive

Start of Block: Tools

Q17 What is your main editing environment

- Pro Tools (1)
- Adobe Suite (2)
- Cockos Reaper (3)
- Apple Logic Pro (4)
- Cubase/Nuendo (5)
- Unity (7)
- Other (6)

Display This Question:

If What is your main editing environment = Other

Q18 If other please specify

_____

_____

_____

_____

_____

Q20 Please list some of your most used plugins and a brief definition of their function

_____

_____

_____

_____

_____

End of Block: Tools

Start of Block: Observation

Q23 Alongside this survey we are also interested in conducting observations and documenting the working practices of professional sound designers at their place of work. This is something that has received little official research attention and would greatly help steer the direction of this project. If you are interested in participating indicate below and you will be contacted with more details.

- Yes (4)
- No (5)

Display This Question:

If Alongside this survey we are also interested in conducting observations and documenting the worki... = Yes

Q24 Please enter an email address below and you will be contacted about taking part in the observation portion of the study

_____

End of Block: Observation

Appendix C

Appendix C Interview metadata

Participant No.	Recording Method	Duration (HH:MM:SS)	Date
1	Zoom built-in recording functionality	00:25:52	19/02/2020
2	Telephone call & Zoom h4n audio recorder	00:36:58 minutes	05/03/2020
3	Zoom built-in recording functionality	00:39:42	16/04/2020
4	Zoom built-in recording functionality	00:33:25	20/04/2020
5	Zoom built-in recording functionality	01:19:10	04/09/2020

Table C.1: Interview Metadata

Appendix D

Appendix D IR Dataset Supplementary Information

D.1 Measurement Apparatus

Measurements were carried out in the University of York AudioLab's listening room in York, UK. A 3D array of 50 full-range Genelec loudspeakers positioned on a sphere sampled using a 50-point Lebedev quadrature sphere with a radius of 1.5m was used to reproduce the exponential sine sweeps for all configurations. Loudspeaker type identified as outlined below:

- Top: 8040A
- Ring 5 upper: 8030A
- Ring 4 upper: 8030A
- Ring 3 upper: 8030A
- Ring 2 upper: 8030A
- Horizonral Ring: 8040A
- Ring 2 lower: 8030A
- Ring 3 lower: 8030A

- Ring 4 lower: 8030A
- Ring 5 lower: 8030A
- Bottom: 8040A

The loudspeaker sound pressure levels were aligned at the centre of the array to 83dBA @ 1kHz measured on a Tenma SPL meter. The loudspeaker driving signals were calibrated for magnitude using an Earthworks M30.

D.2 Available Data

IRs are available for all 50 positions for all configurations listed in Table D. IRs are available as 24-bit PCM .wav files at 48kHz sampling rate using the following naming convention: '[mic array]_IR_[location number]_azi_[azimith in degrees]_el_[elevation in degrees].wav'

Set Name	Array Configuration	Microphone/s used	Microphone Directivity Pattern	Spacing	Orientation angle
AB_Omni_30	AB Pair	AKG C414 XLS	Omnidirectional	30 cm	Parallel
AB_Omni_40	AB Pair	AKG C414 XLS	Omnidirectional	40 cm	Parallel
AB_Cardioid_30	AB Pair	AKG C414 XLS	Cardioid	30 cm	Parallel
AB_Cardioid_40	AB Pair	AKG C414 XLS	Cardioid	40 cm	Parallel
Blumlein	Blumlein	AKG C414 XLS	Bidirectional	Coincident	90°
DIN	DIN	Rode NT5	Cardioid	20 cm near coincident	90°
NOS	NOS	Rode NT5	Cardioid	30 cm near coincident	110° near coincident
ORTF	ORTF	Rode NT5	Cardioid	17cm near coincident	90°
Eigen_SPH	Rigid Spherical Baffle	Eigenmike	up to 4th Order Spherical Harmonics	8.4cm diameter spherical array	
Eigen_raw	Rigid Spherical Baffle	Eigenmike	omnidirectional	8.4cm diameter spherical array	
Coincident	XY	Rode NT4	Cardioid	Coincident	90°

Table D.1: Details of IR sets captured including configuration, spacing, capsule angle, and microphone used.

References

[1] S. Zucchi, S. K. Fuchter, G. Salazar, and K. Alexander, "Combining immersion and interaction in XR training with 360-degree video and 3D virtual objects," in *2020 23rd International Symposium on Measurement and Control in Robotics (ISMCR)*, IEEE, 2020, pp. 1–5, ISBN: 978-1-6654-0479-2. DOI: 10.1109/ISMCR51255.2020.9263732.

[2] BBC, *BBC Soundscapes for Wellbeing aims to bring nature to everyone*, 2021. [Online]. Available: {https://www.bbc.co.uk/mediacentre/2021/soundscapes-for-wellbeing}.

[3] V. Hood, M. Knapp, and D. Griliopoulos, *Best VR games 2021: The top virtual reality games to play right now*, 2021. [Online]. Available: {https://www.techradar.com/uk/best/the-best-vr-games}.

[4] S. R. Quackenbush and J. Herre, "Mpeg standards for compressed representation of immersive audio," *Proceedings of the IEEE*, pp. 1–12, 2021. DOI: 10.1109/JPROC.2021.3075390.

[5] V Lu, J Zhang, K Logishetty, and V Khanduja, "109 The Impact of Extended Reality on Surgery: A Scoping Review," *British Journal of Surgery*, vol. 109, no. Supplement$_6$, znac269.375, Aug. 2022, ISSN: 0007-1323. DOI: 10.1093/bjs/znac269.375. eprint: https://academic.oup.com/bjs/article-pdf/109/Supplement_6/znac269.375/50772413/znac269.375.pdf. [Online]. Available: https://doi.org/10.1093/bjs/znac269.375.

[6] M. Porter and J. Heppelmann, "Why every organization needs an augmented reality strategy," *Harvard Business Review*, 2017. [Online]. Available: {https://hbr.org/2017/11/why-every-organization-needs-an-augmented-reality-strategy}.

[7] M. W. Boyce *et al.*, "Enhancing military training using extended reality: A study of military tactics comprehension," *Frontiers in Virtual Reality*, vol. 3, 2022, ISSN: 2673-4192. DOI: 10.3389/frvir.2022.754627. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frvir.2022.754627.

[8] D. Sonnenschein, *Sound design : the expressive power of music, voice, and sound effects in cinema*, eng. Studio City, Calif.: Studio City, Calif. : Michael Wiese Productions, 2001, ISBN: 9781615930159.

[9] L. Murray, *SOUND DESIGN THEORY AND PRACTICE*, 1st. Routledge, 2019.

[10] P Chueng, P Chueng, P Marsden, and P Marsden, "Designing auditory spaces: the role of expectation," *Proceedings of 10th International Conference on Human Computer Interaction*, pp. 616–620, 2003.

[11] M. C. Green and D. Murphy, "Acoustic scene classification using spatial features," *Detection and Classification of Acoustic Scenes and Events 2017*, p. 4, November 2017.

[12] Y. Aytar, C. Vondrick, and A. Torralba, "SoundNet: Learning Sound Representations from Unlabeled Video," no. Nips, 2016. arXiv: 1610.09001. [Online]. Available: http://arxiv.org/abs/1610.09001.

[13] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2015, pp. 1–6. DOI: 10.1109/MLSP.2015.7324337.

[14] A. Politis, S. Adavanne, and T. Virtanen, "A Dataset of Reverberant Spatial Sound Scenes with Moving Sources for Sound Event Localization

and Detection," 2020. arXiv: 2006.01919. [Online]. Available: http://arxiv.org/abs/2006.01919.

[15] S. Adavanne, A. Politis, and T. Virtanen, "Multichannel Sound Event Detection Using 3D Convolutional Neural Networks for Learning Interchannel Features," *Proceedings of the International Joint Conference on Neural Networks*, vol. 2018-July, 2018. DOI: 10.1109/IJCNN.2018.8489542. arXiv: 1801.09522.

[16] Y. Cao, Q. Kong, T. Iqbal, F. An, W. Wang, and M. D. Plumbley, "Polyphonic sound event detection and localization using a two-stage strategy," in *Detection and Classification of Acoustic Scenes and Events*, 2019. DOI: 10.33682/4jhy-bj81.

[17] J. D. Reiss and Ø. Brandtsegg, "Applications of Cross-Adaptive Audio Effects: Automatic Mixing, Live Performance and Everything in Between," *Frontiers in Digital Humanities*, vol. 5, 2018, ISSN: 2297-2668. DOI: 10.3389/fdigh.2018.00017. [Online]. Available: https://www.frontiersin.org/article/10.3389/fdigh.2018.00017/full.

[18] E. T. Chourdakis and J. D. Reiss, "A machine-learning approach to application of intelligent artificial reverberation," *AES: Journal of the Audio Engineering Society*, vol. 65, no. 1-2, pp. 56–65, 2017, ISSN: 15494950. DOI: 10.17743/jaes.2016.0069.

[19] S. Nercessian, *Izotope and assistive audio technology*, Jul. 2018. [Online]. Available: https://www.izotope.com/en/blog/music-production/izotope-and-assistive-audio-technology.html.

[20] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "Wavegrad: Estimating gradients for waveform generation," in *9th International Conference on Learning Representations (ICLR)*, 2021, pp. 1–15. [Online]. Available: http://arxiv.org/abs/2009.00713.

[21] A. van den Oord *et al.*, "Wavenet: A generative model for raw audio," in *9th International Speech Communication Association Speech Synthesis*

Workshop, 2016, pp. 1–15. [Online]. Available: http://arxiv.org/abs/1609.03499.

[22] H. Liu *et al.*, "Audioldm: Text-to-audio generation with latent diffusion models," *arXiv Pre-print*, Jan. 2023. [Online]. Available: http://arxiv.org/abs/2301.12503.

[23] F Schweiger *et al.*, "Tools for 6-dof immersive audiovisual content capture and production," in *International Broadcasting Convention 2021*, 2021. [Online]. Available: https://www.flir.com/products/grasshopper3-usb3/?model=GS3-U3-51S5C-C.

[24] H. Kim, L. Remaggi, A. Dourado, T. de Campos, P. J. Jackson, and A. Hilton, "Immersive audio-visual scene reproduction using semantic scene reconstruction from 360 cameras," *Virtual Reality*, vol. 26, pp. 823–838, 3 Sep. 2022, ISSN: 14349957. DOI: 10.1007/s10055-021-00594-3.

[25] BBC, "Bbc sound effects archive resource," *BBC Sound Effects Archive Resource • Research & Education Space*, 2019. [Online]. Available: http://bbcsfx.acropolis.org.uk/.

[26] D. M. Howard and J. A. Angus, *Acoustics and psychoacoustics / David M. Howard and Jamie A.S. Angus.* eng, 4th ed. Amsterdam ; London: Focal, 2009, ISBN: 9780080961873.

[27] H. Kuttruff, *Room acoustics / Heinrich Kuttruff.* eng, 5th ed. London & New York: Spon Press/Taylor & Francis, 2009, ISBN: 9786612151521.

[28] *Air - composition and molecular weight*, 2018. [Online]. Available: https://www.engineeringtoolbox.com/molecular-mass-air-d_679.html.

[29] F. A. Everest and K. C. Pohlmann, *Master handbook of acoustics*, eng, 6th edition. New York, N.Y: McGraw-Hill Education LLC, 2015, ISBN: 9780071841030.

[30] J. O. Smith III, *Mathematics of the Discrete Fourier Transform (DFT), with Audio Applications.* W3K Publishing, 2007. [Online]. Available: https://ccrma.stanford.edu/~jos/mdft/.

[31] M. Vorlander, *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality* (RWTHedition), eng. Berlin, Heidelberg: Springer Berlin / Heidelberg, 2007, ISBN: 9783540488293.

[32] *Air - composition and molecular weight*, 2023. [Online]. Available: `http://www.sengpielaudio.com/TableOfSoundPressureLevels.html`.

[33] D. Davis and C. Davis, *Sound system engineering*, eng, 2nd ed. Indianapolis, IN, USA: H.W. Sams, 1987, ISBN: 0672218577.

[34] V. Pulkki and M. Karjalainen, *Communication acoustics : an introduction to speech, audio, and psychoacoustics*, eng. Chichester, West Sussex, United Kingdom : Wiley, 2015, ISBN: 9781118866542.

[35] N. Tsingos, "Object-based audio," in A. Roginska and P. Geluso, Eds. Routledge, 2017, pp. 244–275, ISBN: 9781315707525. DOI: `10.4324/9781315707525`. [Online]. Available: `https://www.taylorfrancis.com/books/9781317480112`.

[36] T. Qu, Z. Xiao, M. Gong, Y. Huang, X. Li, and X. Wu, "Distance-dependent head-related transfer functions measured with high spatial resolution using a spark gap," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, pp. 1124 –1132, Sep. 2009. DOI: `10.1109/TASL.2009.2020532`.

[37] J Blauert, *Spatial hearing: the psychophysics of human sound localisation.* MIT Press, 1997.

[38] W. Gulick, *Hearing Physiological Acoustics, Neural Coding, and Psychoacoustics.* New York: Oxford University Press, 1989.

[39] E Zwicker and H Fastl, *Psychoacoustics – Facts and Models.* Berlin: Springer, 1999.

[40] A. J. Kolarik, B. C. J. Moore, P. Zahorik, S. Cirstea, and S. Pardhan, "Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss," *Attention, Perception, and*

Psychophysics, vol. 78, no. 2, pp. 373–395, 2016, ISSN: 1943393X. DOI: 10.3758/s13414-015-1015-1.

[41] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst, "Auditory Distance Perception in Humans : A Summary of Past and Present Research," *Acta Acustica united with Acustica*, vol. 91, no. February 2003, pp. 409–420, 2005, ISSN: 16101928.

[42] B. Wiggins, "An investigation into the real-time manipulation and control of three-dimensional sound fields," *University of Derby*, pp. 1–370, 2004. [Online]. Available: http://derby.openrepository.com/derby/handle/10545/217795.

[43] V. Pulkki, "Sound, Hearing and Perception," in *Sensory Evaluation of Sound*, N. Zacharov, Ed., Taylor & Francis Group, 2019, ch. 3, pp. 21–58, ISBN: 978-1-49-875136-0.

[44] G. von Bekesy, *Experiments in Hearing*. New York: McGraw-Hill, 1960.

[45] H. Fletcher and W. A. Munsun, "Loudness, its definition, measurement, and calculation," *The Bell System Technical Journal*, vol. 4, no. 12, pp. 337–430, 1933. DOI: 10.1016/S0016-0032(23)90506-5.

[46] H. Fletcher and W. A. Munsun, "Relation between loudness and masking," *Journal of the Acoustical Society of America*, vol. 1, no. 9, pp. 337–430, 1937. DOI: 10.1121/1.1915904.

[47] B. B Bauer and E. L. Torick, "Researches in loudness measurement," *IEEE Transactions on Audio and Electroacoustics*, vol. 3, no. 14, pp. 141–151, 1966.

[48] H. Fletcher, "Auditory patterns," *Reviews of modern Physics*, no. 12, pp. 47–65, 1940.

[49] E. Zwicker, "Subdivision of the audible frequency range into critical bands," *The Journal of the Acoustical Society of America*, no. 33, p. 248, 1961.

[50] B Moore, *Hearing*. Academic Press, 1995.

[51] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1-2, pp. 103–138, 1990. DOI: 10.1016/0378-5955(90)90170-T.

[52] S. Stevens, J. Volkman, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, vol. 3, no. 8, pp. 185–190, 1937. DOI: 10.1121/1.1915893.

[53] J Rayleigh, *The Theory of Sound*. N.Y: Dover, 1945.

[54] G. F. Kuhn, "Model for the interaural time differences in the azimuthal plane," *The Journal of the Acoustical Society of America*, vol. 62, no. 1, pp. 157–167, 1977. DOI: 10.1121/1.381498. [Online]. Available: https://doi.org/10.1121/1.381498.

[55] F. Stevens, "Strategies for Environmental Sound Measurement, Modelling, and Evaluation," PhD, University of York, 2018.

[56] B Moore, *An introduction to the psychology of hearing*, 6th. Emerald Group Publishing Limited, 1945.

[57] E. R. Hafter and J. De Maio, "Difference thresholds for interaural delay," *The Journal of the Acoustical Society of America*, vol. 1, no. 57, pp. 181–187, 1975, ISSN: 0001-4966. DOI: 10.1121/1.380412.

[58] R. C. G. Smith and S. R. Price, "Modelling of human low frequency sound localization acuity demonstrates dominance of spatial variation of interaural time difference and suggests uniform just-noticeable differences in interaural time difference," *PLOS ONE*, vol. 9, no. 2, pp. 1–9, Feb. 2014. DOI: 10.1371/journal.pone.0089033. [Online]. Available: https://doi.org/10.1371/journal.pone.0089033.

[59] G. Kearney, "Auditory Scene Synthesis using Virtual Acoustic Recording and Reproduction," Ph.D. dissertation, University of Dublin, 2010, p. 368.

[60] T. Weiping, H. Ruimin, W. Heng, and C. Wenqin, "Measurement and analysis of just noticeable difference of interaural level difference cue," *2010 International Conference on Multimedia Technology, ICMT 2010*, pp. 5–7, 2010. DOI: 10.1109/ICMULT.2010.5630980.

[61] J. Blauert, "Sound localization in the median plane," *Acta Acustica united with Acustica*, vol. 22, Nov. 1969.

[62] F. L. Wightman and D. J. Kistler, "Resolution of front–back ambiguity in spatial hearing by listener and source movement," *The Journal of the Acoustical Society of America*, vol. 105, no. 5, pp. 2841–2853, 1999, ISSN: 0001-4966. DOI: 10.1121/1.426899. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.426899.

[63] D. R. Begault, A. S. Lee, E. M. Wenzel, and M. R. Anderson, "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 *Audio Engineering Society Convention 108*, Feb. 2000. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=9204.

[64] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *The Journal of the Acoustical Society of America*, vol. 104, no. 5, pp. 3048–3058, 1998, ISSN: 0001-4966. DOI: 10.1121/1.423886. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.423886.

[65] H. Wallach, "The role of head movements and vestibular and visual cues in sound localization," *Journal of Experimental Psychology*, vol. 27, no. 4, pp. 339–368, 1940, ISSN: 00221015. DOI: 10.1037/h0054629.

[66] W. R. Thurlow, J. W. Mangels, and P. S. Runge, "Head movements during sound localization," *Journal of the Acoustical Society of America*, vol. 2, no. 42, pp. 489–493, 1967. DOI: 10.1121/1.1910605.

[67] W. R. Thurlow and P. S. Runge, "Effect of induced head movements on localization of direction of sounds," *Journal of the Acoustical Society of America*, vol. 2, no. 42, pp. 480–488, 1967.

[68] G. Kearney, L. Xujia, A. Manns, and M. Gorzel, "Auditory Distance Perception with Static and Dynamic Binaural Rendering," in *AES 57th International Conference*, 2015, pp. 1–8, ISBN: 9781942220015.

[69] L. Thresh, C. Armstrong, and G. Kearney, "A Direct Comparison of Localisation Performance When Using First, Third and Fifth Order Ambisonics For Real Loudspeaker And Virtual Loudspeaker Rendering," *AES 143rd Convention*, pp. 1–9, 2017.

[70] M. Noisternig, A. Sontacchi, T. Musil, and R. Höldrich, "a 3D Ambisonic Based Binaural Sound Reproduction System," in *AES 24th International Conference on Multichannel Audio*, Banf, 2003, pp. 1–5. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=12314.

[71] E. H. A. Langendijk and A. W. Bronkhorst, "Contribution of spectral cues to human sound localization," *The Journal of the Acoustical Society of America*, vol. 112, no. 4, pp. 1583–1596, 2002, ISSN: 0001-4966. DOI: 10.1121/1.1501901. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.424945http://asa.scitation.org/doi/10.1121/1.1501901.

[72] C. Armstrong, L. Thresh, D. Murphy, and G. Kearney, "A perceptual evaluation of individual and non-individual hrtfs: A case study of the sadie ii database," *Applied Sciences*, vol. 8, no. 11, 2018, ISSN: 2076-3417. DOI: 10.3390/app8112029. [Online]. Available: https://www.mdpi.com/2076-3417/8/11/2029.

[73] W. Bronkhorst Adelbert, "Localization of real and virtual sound sources," *Journal of the Acoustical Society of America*, vol. 98, no. 5, pp. 2542–2553, 2005.

[74] G. Kearney and T. Doyle, "An hrtf database for virtual loudspeaker rendering," in *Audio Engineering Society Convention 139*, 2015. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=17980.

[75] C. W. Pike, "Evaluating the Perceived Quality of Binaural Technology," Ph.D. dissertation, University of York, 2019.

[76] P. Zahorik, "Assessing auditory distance perception using virtual acoustics," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1832–1846, 2002, ISSN: 0001-4966. DOI: 10.1121/1.1458027.

[77] G. Kearney, M. Gorzel, H. Rice, and F. Boland, "Distance perception in interactive virtual acoustic environments using first and higher order ambisonic sound fields," *Acta Acustica united with Acustica*, vol. 98, pp. 61–71, 1 2012, ISSN: 16101928. DOI: 10.3813/AAA.918492.

[78] P. Cocran, J. Throop, and W. E. Simpson, "Estimation of Distance of a Source of Sound," *The American journal of psychology*, vol. 81, no. 2, pp. 198–206, 1968.

[79] W. E. Simpson and L. D. Stanton, "Head movement does not facilitate perception of the distance of a source of sound.," *The American journal of psychology*, vol. 86, no. 1, pp. 151–9, 1973, ISSN: 0002-9556. DOI: 10.2307/1421856. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/4742382.

[80] G. Miller, "Differential sensitivity to white noise," *The Journal of the Acoustical Society of America*, vol. 19, no. 4, pp. 609–619, 1947.

[81] W. Jesteadt, C. Wier, and D. Green, "Intensity discrimination as a function of frequency and sensation level," *The Journal of the Acoustical Society of America*, vol. 61, pp. 169–177, 1977.

[82] R. Riesz, "The relationship between loudness and the minimum perceptible increment of intensity," *The Journal of the Acoustical Society of America*, vol. 4, pp. 211–216, 1933.

[83] A. S. Edwards, "Accuracy of Auditory Depth Perception," *The Journal of General Psychology*, vol. 52, no. 2, pp. 327–329, 1955, ISSN: 0022-1309. DOI: 10.1080/00221309.1955.9920247. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/00221309.1955.9920247.

[84] M. A. Akeroyd, S. Gatehouse, and J. Blaschke, "The detection of differences in the cues to distance by elderly hearing-impaired listeners," *The*

Journal of the Acoustical Society of America, vol. 121, no. 2, pp. 1077–1089, 2007, ISSN: 0001-4966. DOI: 10.1121/1.2404927. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.2404927.

[85] D. Ashmead, D. LeRoy, and R. Odom, "Perception of the relative distances of nearby sound sources," *Attention and Psychophysics*, vol. 47, pp. 326–331, 1990.

[86] D. R. Begault, "Preferred sound intensity increase for sensation of half distance," *Perceptual and Motor Skills*, vol. 72, pp. 1019–1029, 1991, ISSN: 0031-5125. DOI: 10.2466/pms.1991.72.3.1019.

[87] P. Zahorik and F. L. Wightman, "Loudness constancy with varying sound source distance," *Nature Neuroscience*, vol. 4, no. 1, pp. 78–83, 2001, ISSN: 10976256. DOI: 10.1038/82931.

[88] C. F. Altmann, K. Ono, A. Callan, M. Matsuhashi, T. Mima, and H. Fukuyama, "Environmental reverberation affects processing of sound intensity in right temporal cortex," *European Journal of Neuroscience*, vol. 38, no. 8, pp. 3210–3220, 2013, ISSN: 0953816X. DOI: 10.1111/ejn.12318.

[89] P. Zahorik, "Direct-to-reverberant energy ratio sensitivity," *The Journal of the Acoustical Society of America*, vol. 112, no. 5, pp. 2110–2117, 2002, ISSN: 0001-4966. DOI: 10.1121/1.1506692.

[90] D. H. Mershon and J. N. Bowers, "Absolute and relative cues for the auditory perception of egocentric distance," *Perception*, vol. 8, no. 3, pp. 311–322, 1979, ISSN: 03010066. DOI: 10.1068/p080311.

[91] D. H. Mershon and L. E. King, "Intensity and reverberation as factors in the auditory perception of egocentric distance," *Perception & Psychophysics*, vol. 18, no. 6, pp. 409–415, 1975, ISSN: 00315117. DOI: 10.3758/BF03204113.

[92] N. Kopčo and B. G. Shinn-Cunningham, "Effect of stimulus spectrum on distance perception for nearby sources," *The Journal of the Acoustical*

Society of America, vol. 130, no. 3, pp. 1530–1541, 2011, ISSN: 0001-4966. DOI: 10.1121/1.3613705. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.3613705.

[93] A. W. Bronkhorst and T. Houtgast, "Auditory distance perception in rooms," *Nature*, no. 397, pp. 517–205, 1999, ISSN: 00047554.

[94] A. J. Kolarik, S. Cirstea, and S. Pardhan, "Discrimination of virtual auditory distance using level and direct-to-reverberant ratio cues," *The Journal of the Acoustical Society of America*, vol. 134, no. 5, pp. 3395–3398, 2013, ISSN: 0001-4966. DOI: 10.1121/1.4824395. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.4824395.

[95] E. Larsen, N. Iyer, C. R. Lansing, and A. S. Feng, "On the minimum audible difference in direct-to-reverberant energy ratio," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 450–461, 2008, ISSN: 0001-4966. DOI: 10.1121/1.2936368. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.2936368.

[96] J. Jetzt, "Critical distance measurement of rooms from the sound energy spectral response," *The Journal of the Acoustical Society of America*, vol. 65, pp. 1024–1211, 1979.

[97] D. S. Brungart, N. I. Durlach, and W. M. Rabinowitz, "Auditory localization of nearby sources. II. Localization of a broadband source," *The Journal of the Acoustical Society of America*, vol. 106, no. 4, pp. 1956–1968, 1999, ISSN: 0001-4966. DOI: 10.1121/1.427943. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.427943.

[98] A. D. Little, D. H. Mershon, and P. H. Cox, "Spectral content as a cue to perceived auditory distance.," *Perception*, vol. 21, no. 3, pp. 405–416, 1992, ISSN: 03010066. DOI: 10.1068/p210405.

[99] R. A. Butler, E. T. Levy, and W. D. Neff, "Apparent distance of sounds recorded in echoic and anechoic chambers," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 6, no. 4, pp. 745–750, 1980, ISSN: 00961523. DOI: 10.1037/0096-1523.6.4.745.

[100] W. O. Brimijoin, A. W. Boyd, and M. A. Akeroyd, "The contribution of head movement to the externalization and internalization of sounds," *PLoS ONE*, vol. 8, no. 12, pp. 1–12, 2013, ISSN: 19326203. DOI: 10.1371/journal.pone.0083068.

[101] V. Best, R. Baumgartner, M. Lavandier, P. Majdak, and N. Kopčo, "Sound externalization: A review of recent research," *Trends in Hearing*, vol. 24, p. 2 331 216 520 948 390, 2020, PMID: 32914708. DOI: 10.1177/2331216520948390. eprint: https://doi.org/10.1177/2331216520948390. [Online]. Available: https://doi.org/10.1177/2331216520948390.

[102] A. Reeves, "Electric signalling system.," pat. 54 023, 1942.

[103] F. Stevens and D. Murphy, "Spatial impulse response measurement in an urban environment," in *AES 55th International Conference*, Helsinki, Finland, 2014.

[104] M. Schroeder, "Integrated-impulse method measuring sound decay without using impulses," *Journal of the Acoustical Society of America*, vol. 66, pp. 497–500, 1979. DOI: 10.1121/1.383103.

[105] C Dunn and M. Hawksford, "Integrated-impulse method measuring sound decay without using impulses," *Journal of the Audio Engineering Society*, vol. 41, pp. 314–335, 1993.

[106] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," Nov. 2000.

[107] g.-b. stan, j.-j. embrechts, and d. archambeau, "Comparison of different impulse response measurement techniques," *journal of the audio engineering society*, vol. 50, no. 4, pp. 249–262, 2002.

[108] G. Loy, *Musimathics, Volume 2: The Mathematical Foundations of Music*. The MIT Press, 2007, ISBN: 9780262122856. [Online]. Available: http://www.jstor.org/stable/j.ctt5hhm8g (Accessed Mar. 6, 2023).

[109] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex fourier series," *Mathematics of Computation*, vol. 50, no. 4, pp. 297–301, 1965.

[110] J. O. Smith III, *SPECTRAL AUDIO SIGNAL PROCESSING*. W3K Publishing, 2011. [Online]. Available: https://ccrma.stanford.edu/~jos/sasp/.

[111] W. Zhang, P. N. Samarasinghe, H. Chen, and T. D. Abhayapala, *Surround by sound: A review of spatial audio recording and reproduction*, May 2017. DOI: 10.3390/app7050532.

[112] S Devonport and R Foss, "The distribution of ambisonic and point source rendering to ethernet avb speakers," 2019. DOI: 10.22032/dbt.39936. [Online]. Available: https://plugins.iem.at/.

[113] EBU, *Types of audio*, 2019. [Online]. Available: https://adm.ebu.io/background/audio_types.html.

[114] G. Dickins and R. Kennedy, "Towards Optimal Soundfield Representation," in *106th Convention of the Audio Engineering Society*, Munich, Germany, 1999.

[115] M. C. Green, "Environmental Sound Monitoring Using Machine Listening and Spatial Audio," Ph.D. dissertation, University of York, 2021, pp. 1–267.

[116] B. Rafaely, *Fundamentals of Spherical Array Processing* (Springer Topics in Signal Processing), eng, 2015th ed. Berlin, Heidelberg: Springer Berlin / Heidelberg, 2015, vol. 8, ISBN: 9783662456637.

[117] B. Xie, "Spatial Sound-History, Principle, Progress and Challenge," *Chinese Journal of Electronics*, vol. 29, no. 3, pp. 397 –416, 2020. DOI: 10.1049/cje.2020.02.016.

[118] D. self, *Audio engineering* (Newnes know it all series). Amsterdam ; London: Newnes/Elsevier, 2009, ISBN: 9781856175265.

[119] "Multichannel stereophonic sound system with and without accompanying picture," International Telecommunications Union, Geneva, CH, Tech. Rep., 1994.

[120] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "Mpeg-h 3d audio - the new standard for coding of immersive spatial audio," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, pp. 770–779, 5 2015.

[121] H. Fantel, "100 years ago: The beginning of stereo," *The New York Times*, Jan. 4, 1981. [Online]. Available: `https://www.nytimes.com/1981/01/04/arts/100-years-ago-the-beginning-of-stereo.html` (Accessed Mar. 9, 2023).

[122] A. Blumlein, "Improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems," pat. 394 325, 1931.

[123] P. Geluso, "Stereo," in *Immersive Sound: The Art and Science of Binaural and Multi-Channel Audio*, A. Roginska and P. Geluso, Eds., New York, NY: Routledge, 2017, pp. 63–87.

[124] H. Hacihabiboğlu, E. D. Sena, Z. Cvetković, J. Johnston, and J. O. Smith, "Perceptual spatial audio recording, simulation, and rendering: An overview of spatial-audio techniques based on psychoacoustics," *IEEE Signal Processing Magazine*, vol. 34, pp. 36–54, 3 May 2017, ISSN: 10535888. DOI: `10.1109/MSP.2017.2666081`.

[125] D. M. Huber and R. E. Runstein, *Modern Recording Techniques* (Audio Engineering Society Presents), eng. Taylor and Francis, 2017, ISBN: 9781138954373.

[126] F. Rumsey and T. McCormick, *Sound and Recording: Applications and Theory*, eng. Oxford: Taylor & Francis Group, 2014, ISBN: 9780415843409.

[127] B. B. Bauer, "Phasor Analysis of Some Stereophonic Phenomena," *The Journal of the Acoustical Society of America*, vol. 33, no. 11, pp. 1536–1539, 1961, ISSN: 0001-4966. DOI: `10.1121/1.1908492`. [Online]. Available: `http://asa.scitation.org/doi/10.1121/1.1908492`.

[128] D. M. Leakey, "Some Measurements on the Effects of Interchannel Intensity and Time Differences in Two Channel Sound Systems," *The Journal of the Acoustical Society of America*, vol. 31, no. 7, pp. 977–986, 1959, ISSN: 0001-4966. DOI: 10.1121/1.1907824. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.1907824.

[129] M. Frank, "Phantom Sources using Multiple Loudspeakers in the Horizontal," PhD, University of Music and Performing Arts Graz, 2013.

[130] S. Mikrofone, *Decca tree set*, 2020. [Online]. Available: .https://schoeps.de/en/products/stereo/sets/decca-tree-set.html.

[131] M Williams and G Le Du, "Multichannel microphone array design," in *Proceedings of the 108th Audio Engineering Society Convention*, Paris, France: Audio Engineering Society, 2000.

[132] H Wittek and G Theile, "Development and application of a stereophonic multichannel recording technique for 3D Audio and VR," in *143rd Audio Engineering Society Convention*, New York, NY, USA, 2017.

[133] R. Bleidt, A. Borsum, H. Fuchs, and S. M. Weiss, "Object-based audio: Opportunities for improved listening experience and increased listener involvement," *SMPTE Motion Imaging Journal*, vol. 124, no. 5, pp. 1–13, 2015. DOI: 10.5594/j18579.

[134] C. Robinson, N. Tsingos, and S. Mehta, "Scalable format and tools to extend the possibilities of cinema audio," *SMPTE Motion Imaging Journal*, vol. 121, no. 8, 2015.

[135] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *Journal of the Audio Engineering Society*, vol. 6, no. 45, pp. 456–466, 1997.

[136] T. Lossius, P Baltazar, and T. de la Hogue, "DBAP - distance-based amplitude panning," in *2009 International Computer Music Conference*, Montreal, QC, Canada, 2009.

[137] F. Zotter and M. Frank, "Amplitude panning using vector bases," in *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*. Springer Cham, 2019, pp. 41–52. DOI: 10.1007/978-3-030-17207-7_3.

[138] D. Kostadinov, J. D. Reiss, and V. Mladenov, "Evaluation of distance based amplitude panning for spatial audio," Institute of Electrical and Electronics Engineers (IEEE), Dec. 2010, pp. 285–288. DOI: 10.1109/icassp.2010.5495933.

[139] A. Berkhout, "A holographic approach to acoustic control," *Journal of the Audio Engineering Society*, vol. 36, pp. 977–995, 1988.

[140] T. Sporer, K. Brandenburg, S. Brix, and C. Sladeczek, "Wave field synthesis," in A. Roginska and P. Geluso, Eds. Routledge, Oct. 2017, pp. 311–332, ISBN: 9781315707525. DOI: 10.4324/9781315707525. [Online]. Available: https://www.taylorfrancis.com/books/9781317480112.

[141] H. Wittek, "Perceptual Differences Between Wave Field Synthesis and Stereophony," PhD, University of Surrey, 2007.

[142] E. Corteel, K. Nguyen, O. Warusfel, T. Caulkins, and R. Pellegrini, "Objective and subjective comparison of electrodynamic and map loudspeakers for wave field synthesis," in *30th International Conference of the Audio Engineering Society*, 2007.

[143] A. Franck, A. Grä"fe, T. Korn, and Strauß, "Reproduction of moving sound sources by wave field synthesis: An analysis of artifacts," in *32nd International Conference of the Audio Engineering Society*, 2007.

[144] S. Spors, "Spatial aliasing artifacts produced by linear loudspeaker arrays used for wave field synthesis," in *Proceedings of the 2nd International Symposium on Communications, Control and Signal Processing (ISCCSP)*, IEEE Signal Processing Society, 2006.

[145] S. Spors, "Extension of an analytic secondary source selection criterion for wave field synthesis," in *Proceedings of the 123rd Audio Engineering Society Convention*, AES, 2007.

[146] F. Melchior, C. Sladeczek, D. de Vries, and B. Frö"hlich, "User-dependent optimization of wave field synthesis for directive sound fields," in *Proceedings of the 124th Convention of the Audio Engineering Society*, AES, 2008.

[147] "Digital Audio Compression (ac-4) Standard part 2: Immersive and Personalized Audio," ESTI, Tech. Rep., 2015.

[148] S. Füg, A. Hölzer, C. Borß, C. Ertel, M. Kratschmer, and J. Plogsties, "Design, coding and processing of metadata for object-based interactive audio," *Proceedings of the 137th Convention of the Audio Engineering Society (2014)*, pp. 1–12, 2014.

[149] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal of the Audio Engineering Society*, vol. 46, no. 6, pp. 458–466, 2007. [Online]. Available: `http://www.mendeley.com/research/spatial-sound-reproduction-directional-audio-coding/#`.

[150] V. Pulkki, "Directional audio coding in spatial sound reproduction and stereo upmixing," *28th International Conference: The Future of Audio Technology–Surround and Beyond*, pp. 1–7, 2006. [Online]. Available: `http://www.aes.org/e-lib/browse.cfm?elib=13847`.

[151] J. Ahonen, V. Pulkki, and T. Lokki, "Teleconference application and b-format microphone array for directional audio coding," *Proceedings of the AES International Conference*, pp. 1–10, 2007.

[152] m. a. gerzon, "Periphony: With-height sound reproduction," *journal of the audio engineering society*, vol. 21, no. 1, pp. 2–10, 1973.

[153] M. A. Gerzon, "Criteria for evaluating surround-sound systems," *Journal of the Audio Engineering Society*, vol. 25, no. 6, pp. 400–408, 1977.

[154] M. A. Gerzon, "Design of Ambisonic decoders for multispeaker surround sound," in *58th Convention of the Audio Engineering Society*, New York, USA, 1977.

[155] P. B. Fellgett, "Ambisonic reproduction of directionality in surround-sound systems," *Nature*, vol. 252, pp. 234–238, 1975.

[156] M. A. Gerzon, "The design of precisely coincident microphone arrays for stereo and surround sound," in *50th Convention of the Audio Engineering Society*, London, UK, 1975. DOI: 10.1017/CBO9781107415324.004..

[157] P. Craven and M. A. Gerzon, "Coincident microphone simulation covering three dimensional space and yielding various directional outputs," pat. US4042779A, 1975.

[158] m. a. gerzon, "The design of precisely coincident microphone arrays for stereo and surround sound," *journal of the audio engineering society*, 1975.

[159] *Soundfield sps200 software controlled microphone*, 2023. [Online]. Available: https://www.soundfield.com/#/products/sps200.

[160] C. Armstrong and G. Kearney, "Ambisonics understood," in *3D Audio*, J. Paterson and H. Lee, Eds., New York, NY: Routledge, 2021, pp. 99–129.

[161] T. McKenzie, "High Frequency Reproduction in Binaural Ambisonic Rendering," Ph.D. dissertation, University of York, 2019, p. 370. [Online]. Available: http://etheses.whiterose.ac.uk/26445/.

[162] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*, 10th. Washington D.C: Dover Publications, 1972.

[163] D. Malham, "Stereo Signal Decomposition and Upmixing to Surround and 3D Audio," MPhil, University of York, 2003.

[164] M. Chapman *et al.*, "A Standard for interchange of Ambisonic signal sets," in *Proceedings of the Ambisonic Symposium*, Graz, Austria, 2009.

[165] mH Acoustics, *em32 Eigenmike microphone array release notes (v17.0*, NA. [Online]. Available: {https://mhacoustics.com/sites/default/files/ReleaseNotes.pdf}.

[166] J.-M. Jot, V. Larcher, and J.-M. Pernaux, "A comparative study of 3-d audio encoding and rendering techniques," in *16th International Conference: Spatial Sound Reproduction*, 2013. [Online]. Available: http://www.audiogroup.web.fh-koeln.de/FILES/AIA-DAGA2013_HRIRs.pdf.

[167] A. Farina, R. Glasgal, E. Armelloni, and A. Torger, "Ambiophonic principles for the recording and reproduc- tion of surround sound for music," in *AES 19th International Conference*, 2001.

[168] F. Zotter and M. Frank, *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*, 1st ed. Springer Cham, 2019, ISBN: 9783030172060. DOI: 10.1007/978-3-030-17207-7_1. [Online]. Available: http://link.springer.com/10.1007/978-3-030-17207-7_1.

[169] B. Wiggins, "THE GENERATION OF PANNING LAWS FOR IRREGULAR SPEAKER ARRAYS USING HEURISTIC METHODS," in *AES 31st International Conference*, London, UK, 2007, pp. 1–14.

[170] A. J. Heller, E. Benjamin, and R. Lee, "Design of Ambisonic Decoders for Irregular Arrays of Loudspeakers by Non-Linear Optimization," in *129th Audio Engineering Society Convention*, San Francisco, CA, USA, 2010.

[171] V Bruschi, S Nobili, S Cecchi, and F Piazza, "An Innovative Method for Binaural Room Impulse Responses Interpolation," in *Proceedings of the 148th Audio Engineering Society Convention*, Online: AES, 2020. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=20802.

[172] C Masterson, G Kearney, and F Boland, "Acoustic Impulse Response Interpolation for Multichannel Systems Using Dynamic Time Warping," in *Proceedings of the 35th AES International Conference: Audio for Games*, London, UK: AES, 2009. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=15188.

[173] j. merimaa and v. pulkki, "Spatial impulse response rendering i: Analysis and synthesis," *journal of the audio engineering society*, vol. 53, no. 12, pp. 1115–1127, 2005.

[174] D. T Murphy and S Shelly, "Openair: An interactive auralization web resource and database," in *129th Audio Engineering Society Convention*, San Francisco, CA, USA, 2010.

[175] G Kaasik, N Näveri, H Möller, and Piskarskas, "Openair: An interactive auralization web resource and database," in *Proceesings of the Institute of Acoustics*, vol. 40, 2018.

[176] H. Møller, "Fundamentals of binaural technology," *Applied Acoustics*, vol. 36, no. 4, pp. 171–218, 1992, ISSN: 0003682X. DOI: 10.1016/0003-682X(92)90046-U.

[177] B. Atal, "Apparent sound source translator," pat. 3 236 949, 1966.

[178] Kö"ring, J. and Schmitz, A., "Simplifying cancellation of cross-talk for playback of head-related recordings in a two-speaker system," *Acta Acustica United with Acustica*, vol. 79, no. 3, pp. 221–232, 1993.

[179] B. Rafaely *et al.*, *Spatial audio signal processing for binaural reproduction of recorded acoustic scenes-review and challenges*, 2022. DOI: 10.1051/aacus/2022040.

[180] S. Nagel and P. Jax, "Dynamic binaural cue adaption," in *16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, IEEE, 2018.

[181] F. Brinkmann, A. Lindau, and S. Weinzierl, "A high resolution head-related transfer function database including different orientations of head above the torso," in *AIA-DAGA 2013 Conference on Acoustics*, 2013.

[182] I. Engel, D. Alon, P. Robinson, and R. Mehra, "The effect of generic headphone compensation on binaural renderings," in *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*, 2019.

[183] A. Lindau and F. Brinkmann, "Perceptual evaluation of headphone compensation in binaural synthesis based on non-individual recordings," *Journal of the Audio Engineering Society*, vol. 60, no. 1-2, pp. 54–62, 2012.

[184] D. Pralong and S. Carlile, "The role of individualized headphone calibration for the generation of high fidelity virtual auditory space," *The Journal of the Acoustical Society of America*, vol. 100, no. 6, pp. 3785–3793, 1996.

[185] P. Majdak, P. Balazs, and B. Laback, "Multiple exponential sweep method for fast measurement of head-related transfer functions," *Journal of the Audio Engineering Society*, vol. 55, pp. 623–636, 2007.

[186] B. Bernschü"tz, "A spherical far field hrir/hrtf compilation of the neumann ku 100," in *Fortschritte der Akustik - AIA-DAGA 2013*, 2013. [Online]. Available: `http://www.audiogroup.web.fh-koeln.de/FILES/AIA-DAGA2013_HRIRs.pdf`.

[187] H. Møller, "Binaural technique: do we need individual recordings," *Journal of the Audio Engineering Society*, vol. 44, no. 6, pp. 451–469, 1996.

[188] E. Wenzel, M. Arruda, D. Kistler, and F. Wightman, "Localization using nonindividualzed head-related transfer function," *The Journal of the Acoustical Society of America*, vol. 94, no. 1, pp. 111–123, 1993. DOI: `10.1121/1.407089`. [Online]. Available: `https://doi.org/10.1121/1.407089`.

[189] D. Begault and E. Wenzel, "Headphone localization of speech," *Human Factors*, vol. 35, no. 2, pp. 361–376, 1993. DOI: `10.1177/001872089303500210`. [Online]. Available: `https://doi.org/10.1177/001872089303500210`.

[190] D. Begault, E. Wenzel, and M. Anderson, "Direct comparison of the impact of head tracking, reverberation, and individualised head-related transfer functions on the spatial perception of a virtual speech source," *Journal of The Audio Engineering Society*, vol. 49, no. 10, pp. 904–916, 2001.

[191] A. McKeag and D. McGrath, "Sound field format to binaural decoder with head-tracking," in *6th Australian Regional Convention of the Audio Engineering Society*, 1996.

[192] C. Armstrong, "Improvements in the measurement and optimisation of head related transfer functions for binaural ambisonics," 2019.

[193] E. Wenzel and S. Foster, "Perceptual consequences of interpolating head-realted transfer functions during spatial synthesis," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 1996.

[194] A. Avni, J. Ahrens, M. Geier, S. Spors, H. Wierstorf, and B. Rafaely, "Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 2711–2721, 2013. DOI: 10.1121/1.4795780.

[195] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, eng. MIT Press, 2016.

[196] F. Chollet, *Deep Learning with Python*, eng. Manning, 2021, ISBN: 9781617296864.

[197] U. Zölzer, *DAFX : digital audio effects*, eng. Chichester: Wiley, 2011, ISBN: 9781119991298.

[198] T. Wilmering, D. Moffat, A. Milo, and M. B. Sandler, *A history of audio effects*, Feb. 2020. DOI: 10.3390/app10030791.

[199] M. A. Ramírez, E. Benetos, and J. D. Reiss, "Deep learning for black-box modeling of audio effects," *Applied Sciences (Switzerland)*, vol. 10, 2 Jan. 2020, ISSN: 20763417. DOI: 10.3390/app10020638.

[200] D. Reed, "A perceptual assistant to do sound equalization," ACM, Jan. 2000, pp. 212–218, ISBN: 1581131348. DOI: 10.1145/325737.325848. [Online]. Available: https://dl.acm.org/doi/10.1145/325737.325848.

[201] B. A. Kolasinski, "A framework for automatic mixing using timbral similarity measures and genetic optimization," May 2008. [Online]. Available: www.aes.org..

[202] N. Jillings and R. Stables, "Automatic masking reduction in balance mixes using evolutionary computing," Oct. 2017. [Online]. Available: https://tech.ebu.ch/docs/r/r128.pdf.

[203] P. Aichinger, A. Sontacchi, and B. Schneider-Stickler, "Describing the —transparency of mixdowns: The masked-to-unmasked-ratio," 2011.

[204] J. Scott, M. Prockup, E. M. Schmidt, and Y. E. Kim, "Automatic multi-track mixing using linear dynamical systems," 2011. [Online]. Available: http://music.ece.drexel.edu/research/AutoMix.

[205] J. Scott and Y. E. Kim, "Analysis of acoustic features for automated multi-track mixing," 2011. [Online]. Available: http://music.ece.drexel.edu/research/AutoMix.

[206] E. T. Chourdakis, L. Ward, M. Paradis, and J. D. Reiss, "Modelling experts' decisions on assigning narrative importances of objects in a radio drama mix," Sep. 2019. [Online]. Available: http://marsyas.info/downloads/datasets.html.

[207] S. Hershey *et al.*, "Cnn architectures for large-scale audio classification," 2017, pp. 131–135, ISBN: 9781509041176.

[208] J. Gemmeke *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," 2017, pp. 131–135, ISBN: 9781509041176.

[209] A. L. Benito and J. D. Reiss, "Intelligent multitrack reverberation based on hinge-loss markov random fields," Jun. 2017. [Online]. Available: http://www.aes.org/e-lib.

[210] D. Sheng and G. Fazekas, "A feature learning siamese model for intelligent control of the dynamic range compressor," Jul. 2019, ISBN: 9781728120096. [Online]. Available: http://www.ieee.org/publications.

[211] G. Koch, "Siamese neural networks for one-shot image recognition," Feb. 2015.

[212] M. A. M. Ramírez, "Deep learning for audio effects modeling," Nov. 2022.

[213] T. Schmitz and J.-J. Embrechts, "Nonlinear real-time emulation of a tube amplifier with a long short term memory neural-network," May 2018. [Online]. Available: http://www.aes.org/e-lib.

[214] Z. Zhang, E. Olbrych, J. Bruchalski, T. J. McCormick, and D. L. Livingston, "A vacumm-tube guitar amplifier model using long/short term memory networks," 2018.

[215] E.-P. Damskägg, L. Juvela, E. Thuillier, and V. Valimaki, "Deep learning for tube amplifier emulation," IEEE, May 2019, pp. 471–475, ISBN: 978-1-4799-8131-1. DOI: 10.1109/ICASSP.2019.8682805. [Online]. Available: https://ieeexplore.ieee.org/document/8682805/.

[216] A. Wright, E.-P. Damskägg, L. Juvela, and V. Välimäki, "Real-time guitar amplifier emulation with deep learning," *Applied Sciences*, vol. 10, no. 3, 2020, ISSN: 2076-3417. DOI: 10.3390/app10030766. [Online]. Available: https://www.mdpi.com/2076-3417/10/3/766.

[217] M. A. M. Ramírez and J. D. Reiss, "End-to-end equalization with convolutional neural networks," Sep. 2018.

[218] M. A. M. Ramirez and J. D. Reiss, "Modeling nonlinear audio effects with end-to-end deep neural networks," IEEE, May 2019, pp. 171–175, ISBN: 978-1-4799-8131-1. DOI: 10.1109/ICASSP.2019.8683529. [Online]. Available: https://ieeexplore.ieee.org/document/8683529/.

[219] S. H. Hawley, B. Colburn, and S. I. Mimilakis, "Profiling audio compressors with deep neural networks," Oct. 2019. [Online]. Available: http://www.aes.org/e-lib.

[220] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI*, May 2015, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28. [Online]. Available: http://link.springer.com/10.1007/978-3-319-24574-4_28.

[221] T. T. Lim, R. A. Yeh, Y. Xi, M. N. Do, and M. Hasegwa-Johnson, "Time-frequency networks for audio super-resolution," IEEE, 2018.

[222] S. Vasquez and M. Lewis, *Melnet: A generative model for audio in the frequency domain*, 2019. arXiv: 1906.01083 [eess.AS].

[223] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," *CoRR*, vol. abs/1711.00937, 2017. arXiv: 1711.00937. [Online]. Available: http://arxiv.org/abs/1711.00937.

[224] C. Donahue, J. J. McAuley, and M. S. Puckette, "Synthesizing audio with generative adversarial networks," *CoRR*, vol. abs/1802.04208, 2018. arXiv: 1802.04208. [Online]. Available: http://arxiv.org/abs/1802.04208.

[225] A. Barahona-Rios and S. Pauletto, "Synthesising knocking sound effects using conditional wavegan," in *Proceedings of the 17th Sound and Music Computing Conference*, Jun. 2020.

[226] J. H. Engel, K. K. Agrawal, S. Chen, I. Gulrajani, C. Donahue, and A. Roberts, "Gansynth: Adversarial neural audio synthesis," *CoRR*, vol. abs/1902.08710, 2019. arXiv: 1902.08710. [Online]. Available: http://arxiv.org/abs/1902.08710.

[227] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," *CoRR*, vol. abs/1601.06759, 2016. arXiv: 1601.06759. [Online]. Available: http://arxiv.org/abs/1601.06759.

[228] A. van den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu, "Conditional image generation with pixelcnn decoders," *CoRR*, vol. abs/1606.05328, 2016. arXiv: 1606.05328. [Online]. Available: http://arxiv.org/abs/1606.05328.

[229] P. Dutilleux, "An implementation of the "algorithme à trous" to compute the wavelet transform," in J.-M. Combes, A. Grossman, and P. Tchamitchian, Eds. Springer, 1989, pp. 289–304.

[230] S. Mehri *et al.*, "SampleRNN: An unconditional end-to-end neural audio generation model," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=SkxKPDv5xl.

[231] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N.

Lawrence, and K. Weinberger, Eds., vol. 27, Curran Associates, Inc., 2014. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

[232] A. Radford, L. Metz, and S. Chintala, *Unsupervised representation learning with deep convolutional generative adversarial networks*, 2016. arXiv: 1511.06434 [cs.LG].

[233] C. Y. Lee, A. Toffy, G. J. Jung, and W.-J. Han, "Conditional wavegan," *arXiv preprint arXiv:1809.10636*, Sep. 2018.

[234] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *CoRR*, vol. abs/1411.1784, 2014. arXiv: 1411.1784. [Online]. Available: http://arxiv.org/abs/1411.1784.

[235] L. Yang *et al.*, "Diffusion models: A comprehensive survey of methods and applications," *ACM Comput. Surv.*, vol. 56, no. 4, 2023, ISSN: 0360-0300. DOI: 10.1145/3626235. [Online]. Available: https://doi.org/10.1145/3626235.

[236] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. A. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," *CoRR*, vol. abs/2105.06337, 2021. arXiv: 2105.06337. [Online]. Available: https://arxiv.org/abs/2105.06337.

[237] S. Kim, H. Kim, and S. Yoon, *Guided-tts 2: A diffusion model for high-quality adaptive text-to-speech with untranscribed data*, 2022. arXiv: 2205.15370 [cs.SD].

[238] D. Yang *et al.*, "Diffsound: Discrete diffusion model for text-to-sound generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1720–1733, 2023. DOI: 10.1109/TASLP.2023.3268730.

[239] H. Liu *et al.*, "Audioldm: Text-to-audio generation with latent diffusion models," in *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[240] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095969.

[241] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *CoRR*, vol. abs/2010.05646, 2020. arXiv: 2010.05646. [Online]. Available: https://arxiv.org/abs/2010.05646.

[242] J. Francombe, T. Brookes, and R. Mason, "Evaluation of spatial audio reproduction methods (Part 1): Elicitation of perceptual differences," *AES: Journal of the Audio Engineering Society*, vol. 65, no. 3, pp. 198–211, 2017.

[243] J. Williams, S. Shepstone, and D. Murphy, "Understanding immersion in the context of films with spatial audio," in *Audio Engineering Society Conference: AES 2022 International Audio for Virtual and Augmented Reality Conference*, Audio Engineering Society, 2022.

[244] S. Agrawal, A. Simon, S. Bech, K. Bærenstein, and S. Forchhammer, "Defining Immersion: Literature Review and Implications for Research on Immersive Audiovisual Experiences," *AES 147th Convention*, pp. 1–11, 2019.

[245] C. Eaton and H. Lee, "Quantifying Factors of Auditory Immersion in Virtual Reality," in *International conference on Immersive and Interaction Audio*, York, 2019.

[246] L Ermi and F Mäyrä, "Fundamental components of the gameplay experience: Analysing immersion," 2005.

[247] F Biocca and B Delaney, "Immersive virtual reality technology," in *Communication in the Age of Virtual Reality*, Lawrence Erlbaum Associates, Inc, 1995.

[248] A. McMahan, "Immersion, engagement, and presence: A method for analyzing 3-d video games," in *The Video Game Theory Reader*, January 2003, M. J. Wolf and B. Perron, Eds., 1st, Routledge, 2003, ch. 3, pp. 67–86, ISBN: 9781135205195. DOI: 10.4324/9780203700457-10.

[249] E Adams and A Rollings, *Fundamentals of Game Design*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc, 2006.

[250] A Thon J-N, "Immersion revisited: On the value of a contested concept," in *Extending Experiences. Structure, Analysis and Design of Computer Game Player Experience*, Rovaniemi, Finland: Lapland University Press, 2008.

[251] M. Ryan, *Narrative as Virtual Reality: Immersion and Interactivity in Literature and Electronic Media*. Baltimore, MD, USA: The John Hopkinds Univerity Press, 2003.

[252] BBC R&D, *Sounding Special: Doctor Who in Binaural Sound*, https://www.bbc.co.uk/rd/blog/201 05-doctor-who-in-binaural-sound, 2020.

[253] M. Slater, "A note on presence terminology," 2003.

[254] J. J. Ruscella and M. F. Obeid, "A taxonomy for immersive experience design," in *Proceedings of 2021 7th International Conference of the Immersive Learning Research Network, iLRN 2021*, Institute of Electrical and Electronics Engineers Inc., May 2021, ISBN: 9781734899528. DOI: 10.23919/iLRN52045.2021.9459328.

[255] M. Firth, R. Bailey, and C. Pike, *Binaural EBU ADM Renderer*, 2020. [Online]. Available: https://www.bbc.co.uk/rd/blog/2020-10-ear-next-generation-audio-software-tools.

[256] Facebook, *Facebook360*. [Online]. Available: https://facebook360.fb.com/.

[257] Google, *Google arcore*. [Online]. Available: https://developers.google.com/ar.

[258] S. H.-W. Chuah, "Why and who will adopt extended reality technology? literature review, synthesis, and future research agenda," *SSRN Electronic Journal*, 2018, ISSN: 1556-5068. DOI: 10.2139/ssrn.3300469. [Online]. Available: https://www.ssrn.com/abstract=3300469.

[259] R. Azuma, "A survey of augmented reality," *Presence: Teleoperators and Virtual Environments*, vol. 6, pp. 355–385, 1997.

[260] T. Caudell and D. Mizell, "Augmented Reality: an application of heads-up display technology to manual manufacturing processes," in *Proceedings of 25th Hawaii International Conference on System Sciences*, Hawaii, USA, 1992, pp. 659–669.

[261] N. Inc and T. P. Company, *Pokèmon go*, 2020. [Online]. Available: https://pokemongolive.com/?hl=en.

[262] D Drascic and P Milgram, "Perceptual issues in augmented reality," *Stereoscopic Displays and Virtual Reality*, vol. 2653, pp. 123–135, 1996.

[263] M. C. tom Dieck, T. H. Jung, and P. A. Rauschnabel, "Determining visitor engagement through augmented reality at science festivals: An experience economy perspective," *Computers in Human Behavior*, vol. 82, pp. 44–53, May 2018, ISSN: 07475632. DOI: 10.1016/j.chb.2017.12.043.

[264] R. Doerner, W. Broll, B. Jung, P. Grimm, M. Göbel, and R. Kruse, "Introduction to Virtual and Augmented Reality," in *Virtual and Augmented Reality (VR/AR)*, R. Doerner, W. Broll, P. Grimm, and B. Jung, Eds., Springer, 2022, ch. 1, pp. 1–38, ISBN: 978-3-030-79061-5.

[265] Cubitts, *The speculator*, 2020. [Online]. Available: https://cubitts.com/pages/the-speculator.

[266] IKEA, *Ikea place*, 2020. [Online]. Available: https://www.ikea.com/au/en/customer-service/mobile-apps/say-hej-to-ikea-place-pub1f8af050.

[267] S. Inc., *Snapchat*, 2020. [Online]. Available: https://www.snapchat.com/en-GB.

[268] Google, *Live view in google maps*, 2020. [Online]. Available: https://arvr.google.com/ar/.

[269] StoryFutures Academy, *Storytrails*, 2022. [Online]. Available: https://story-trails.com/.

[270] A. Härmä, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, and J. Hiipakka, "Augmented reality audio for mobile and weable appliances, volume = , year = 2004," *Journal of the Audio Engineering Society*, no. 52, pp. 618–693,

[271] DARKFIELD, *Darkfield radio: Immersive audio experiences at home*, 2023. [Online]. Available: https://www.darkfield.org/radio.

[272] DARKFIELD, *Darkfield*, 2023. [Online]. Available: https://www.darkfield.org/.

[273] DARKFIELD, *Visitors*, 2020. [Online]. Available: https://www.darkfield.org/dfradio-season-1#visitors.

[274] ECHOES, *Ghost walk*, 2023. [Online]. Available: https://echoes.xyz/project/project-template-duplicate.

[275] Bose, *Bose frames*, 2022. [Online]. Available: https://www.bose.co.uk/en_gb/products/certified_refurbished/refurbished_audio_sunglasses/bose-frames-tenor-fr.html#v=bose_frames_tenor_fr_black_row.

[276] Playlines, *Consequences ar: London showcase*, 2019. [Online]. Available: https://playlines.net/portfolio/consequences-ar-london-showcase/.

[277] Meta, *Ray-ban — meta*, 2023. [Online]. Available: https://www.meta.com/gb/smart-glasses/.

[278] E. Ltd, *Maverick*, 2023. [Online]. Available: https://www.everysight.com/maverick.

[279] Google, *Ar glasses experience*, 2023. [Online]. Available: https://arvr.google.com/.

[280] Google, *Google lens*, 2023. [Online]. Available: https://lens.google/intl/en-GB/.

[281] S Mann, "Mediated Reality," *Linux Journal*, vol. 59, no. 5, 1999.

[282] stare at the wall, *Stare at the wall x megaverse: Flood*, 2023. [Online]. Available: https://www.stareatthewall.studio/stareatthewall-x-megaverse.

[283] Vicon, *Vicon*, 2023. [Online]. Available: https://www.vicon.com/.

[284] R. Doerner and F. Steinicke, "Percpetual Aspects of VR," in *Virtual and Augmented Reality (VR/AR)*, R. Doerner, W. Broll, P. Grimm, and B. Jung, Eds., Springer, 2022, ch. 2, pp. 39–70, ISBN: 978-3-030-79061-5.

[285] H. Vive, *Vive focus 3*, 2023. [Online]. Available: https://www.vive.com/uk/product/vive-focus3/specs/.

[286] R. Monica and J. Aleotti, "Evaluation of the oculus rift s tracking system in room scale virtual reality," *Virtual Reality*, vol. 26, pp. 1335–1345, 4 Dec. 2022, ISSN: 1359-4338. DOI: 10.1007/s10055-022-00637-3.

[287] Meta, *Meta quest 2*, 2023. [Online]. Available: https://www.meta.com/gb/quest/products/quest-2/.

[288] Vive, *Htc vive cosmos*, 2023. [Online]. Available: https://www.vive.com/uk/product/vive-cosmos-elite-headset/overview/.

[289] Meta, *Meta quest pro*, 2023. [Online]. Available: https://www.meta.com/gb/quest/quest-pro/.

[290] Varjo, *Varjo vr-3*, 2023. [Online]. Available: https://varjo.com/products/vr-3/.

[291] PicoXR, *Pico neo3 link*, 2023. [Online]. Available: https://www.picoxr.com/uk/products/neo3-link.

[292] J. Khalili, *Htc vive focus 3 review*, 2021. [Online]. Available: https://www.techradar.com/reviews/htc-vive-focus-3.

[293] Vicon, *Vicon pulsar accessories*, 2023. [Online]. Available: https://m.facebook.com/Vicon/posts/10156615076109177/?_se_imp=2LkYelJHi9hYdA0s5.

[294] Varjo, *Varjo aero. reach new heights in virtual reality*, 2023. [Online]. Available: https://varjo.com/products/aero/.

[295] Steam, *Valve index*, 2023. [Online]. Available: https://store.steampowered.com/valveindex.

[296] PicoXR, *Pico 4*, 2023. [Online]. Available: https://www.picoxr.com/uk/products/pico4.

[297] Meta, *Unwrap mixed reality with meta quest 3*, 2023. [Online]. Available: https://www.meta.com/gb/quest/quest-3/.

[298] H. Vive, *Vive xr elite*, 2023. [Online]. Available: https://www.vive.com/uk/product/vive-xr-elite/overview/.

[299] Microsoft, *Hololens 2*, 2023. [Online]. Available: https://www.microsoft.com/en-us/hololens/hardware#Document%20experiences.

[300] Varjo, *Varjo xr-4*, 2023. [Online]. Available: https://varjo.com/products/xr-4/.

[301] Varjo, *Varjo xr-3*, 2023. [Online]. Available: https://varjo.com/products/xr-3/.

[302] C. Flavián, S. Ibáñez-Sánchez, and C. Orús, "The impact of virtual, augmented and mixed reality technologies on the customer experience," *Journal of Business Research*, vol. 100, pp. 547–560, Jul. 2019, ISSN: 01482963. DOI: 10.1016/j.jbusres.2018.10.050.

[303] P Milgram, H Takemura, A Utsumi, and K. F, "Augmented Reality: a class of displays on the reality-virtuality continuum," in *Proc SPIE*, 1990.

[304] S. Rokhsaritalemi, A. Sadeghi-Niaraki, and S. M. Choi, "A review on mixed reality: Current trends, challenges and prospects," *Applied Sciences (Switzerland)*, vol. 10, 2 Jan. 2020, ISSN: 20763417. DOI: 10.3390/app10020636.

[305] P. Reverie, *Prox & reverie: Where worlds collide*, 2022. [Online]. Available: https://proxandreverie.com/.

[306] X. Stories, *Xr stories*, 2023. [Online]. Available: https://xrstories.co.uk/about/.

[307] W. Bros., *Warner bros. discovery*, 2023. [Online]. Available: https://wbd.com/.

[308] C. E. Team, *Interchange: A multi-reality gateway*, 2022. [Online]. Available: https://www.youtube.com/watch?v=vzgtubZUgtE.

[309] J. Collins, H. Regenbrecht, and T. Langlotz, "Visual coherence in mixed reality: A systematic enquiry," *Presence*, vol. 26, no. 1, pp. 16–41, 2017. DOI: 10.1162/PRES_a_00284.

[310] J. Mateer, "Directing for cinematic virtual reality: How the traditional film director's craft applies to immersive environments and notions of presence," *Journal of Media Practice*, vol. 18, pp. 14–25, 1 Jan. 2017, ISSN: 1468-2753. DOI: 10.1080/14682753.2017.1305838.

[311] Youtube, *Youtube*, 2023. [Online]. Available: https://www.youtube.com/.

[312] BBC, *Click 360*, 2023. [Online]. Available: https://www.youtube.com/watch?v=c3zeH-YEHkE.

[313] F. Nielsen, "Surround video: A multihead camera approach," *Visual Computer*, vol. 21, pp. 92–103, 1-2 Feb. 2005, ISSN: 01782789. DOI: 10.1007/s00371-004-0273-z.

[314] R. Muller *et al.*, *Workflow post-production*, 2023. [Online]. Available: https://www.youtube.com/watch?v=c3zeH-YEHkE.

[315] S. S. Lab, *An interactive vr film: Afterlife*, 2023. [Online]. Available: http://afterlife-vr.com/.

[316] J.-L. Sinclair, *Principles of Game Audio and Sound Design; Sound Design and Audio Implementation for Interactive and Immersive Media*. 2020.

[317] G Zdanowicz and S. Bambrick, *The Game Audio Strategy Guide: A Practical Course*, 1st. Taylor and Francis, 2019.

[318] I. Salselas and R. Penha, "The role of sound in inducing storytelling in immersive environments," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Sep. 2019, pp. 191–198, ISBN: 9781450372978. DOI: 10.1145/3356590.3356619.

[319] C. Popp and D. Murphy, "Dynamic binaural cue adaption," in *2022 AES Conference on Audio for Virtual and Augmented Reality*, AES, Aug. 2022.

[320] M. Gröhn, T. Lokki, and T. Takala, "Comparison of auditory, visual, and audiovisual navigation in a 3d space," *ACM Transactions on Applied Perception*, vol. 2, pp. 564–570, 4 Oct. 2005, ISSN: 1544-3558. DOI: 10.1145/1101530.1101558. [Online]. Available: https://dl.acm.org/doi/10.1145/1101530.1101558.

[321] B. N. Walker and J. Lindsay, "Navigation performance with a virtual auditory display: Effects of beacon sound, capture radius, and practice," *Human Factors*, vol. 48, no. 2, pp. 265–278, 2006, PMID: 16884048. DOI: 10.1518/001872006777724507. eprint: https://doi.org/10.1518/001872006777724507. [Online]. Available: https://doi.org/10.1518/001872006777724507.

[322] K. Allain *et al.*, "An audio game for training navigation skills of blind children," in *IEEE 2nd VR Workshop on Sonic Interactions for Virtual Environments*, 2015.

[323] O. Koskela and K. Tuuri, "Investigating metaphors of musical involvement: Immersion, flow, interaction and incorporation," in *Proceedings of the Audio Mostly 2018 on Sound in Immersion and Emotion*, ser. AM'18, Wrexham, United Kingdom: Association for Computing Machinery, 2018, ISBN: 9781450366090. DOI: 10.1145/3243274.3243293. [Online]. Available: https://doi.org/10.1145/3243274.3243293.

[324] S. D. Lipscomb and S. M. Zehnder, "Immersion in the virtual environment: The effect of a musical score on the video gaming experience," *Journal of PHYSIOLOGICAL ANTHROPOLOGY and Applied Human Science*, vol. 23, pp. 337–343, 6 2004, ISSN: 1345-3475. DOI: 10.2114/jpa.23.337.

[Online]. Available: http://www.jstage.jst.go.jp/article/jpa/23/6/23_6_337/_article.

[325] A. J. Cohen, "Music as a Source of Emotion in Film," in *Music and Emotion: Theory and Research*, P. Juslin and J. Sloboda, Eds., 1st, Oxford University Press, 2001, pp. 249–279.

[326] G. O., *Virtual Art: From Illusion to Immersion*. Cambridge, Mass: MIT Press, 2003.

[327] C. Summers and M. Jesse, "Creating immersive and aesthetic auditory spaces in virtual reality," vol. 2017-April, Institute of Electrical and Electronics Engineers Inc., Apr. 2017, pp. 1–6, ISBN: 9781538604595. DOI: 10.1109/SIVE.2017.7938144.

[328] J. Vachon, "Avoiding tedium - fighting repetition in game audio," in *35th International Conference: Audio for Games*, Audio Engineering Society, 2009.

[329] S. Greuter and A. Nash, "Game asset repetition," in *ACM International Conference Proceeding Series*, vol. 02-03-December-2014, Association for Computing Machinery, Dec. 2014, ISBN: 9781450327909. DOI: 10.1145/2677758.2677782.

[330] N. Paterson, K. Naliuka, S. K. Jensen, T. Carrigy, M. Haahr, and F. Conway, "Spatial audio and reverberation in an augmented reality game sound design," in *AES 40th International Conference*, Oct. 2010.

[331] P. Bala, R. Masu, V. Nisi, and N. Nunes, ""when the elephant trumps": A comparative study on spatial audio for orientation in 360◦ videos," in *Conference on Human Factors in Computing Systems - Proceedings*, Association for Computing Machinery, May 2019, ISBN: 9781450359702. DOI: 10.1145/3290605.3300925.

[332] F. Cuadrado, I. Lopez-Cobo, T. Mateos-Blanco, and A. Tajadura-Jiménez, "Arousing the sound: A field study on the emotional impact on children of arousing sound design and 3d audio spatialization in an audio story,"

Frontiers in Psychology, vol. 11, May 2020, ISSN: 16641078. DOI: 10.3389/fpsyg.2020.00737.

[333] M. Gospodarek, A. Genovese, D. Dembeck, C. Brenner, A. Roginska, and K. Perlin, "Sound design and reproduction techniques for co-located narrative vr experiences," in *147th Audio Engineering Society Convention*, Audio Engineering Society, 2019.

[334] J. Ott, A.-S. Tutescu, N. Wienböker, J. Rosenbauer, and T. Görne, "Spatial audio production for immersive fulldome projections," in *5th International Conference on Spatial Audio (ICSA)*, 2019. DOI: 10.22032/dbt.39974.

[335] X. Amatriain *et al.*, "Experiencing audio and music in a fully immersive environment," in *International Symposium on Computer Music Modelling and Retrieval. Sense of Sounds*, Springer Berlin Heidelberg, 2008, pp. 380–400. DOI: 10.1007/978-3-540-85035-9_27. [Online]. Available: http://link.springer.com/10.1007/978-3-540-85035-9_27.

[336] W. E. Garity and N. A. Hawkins, "Fantasound," *Journal of the Society of Motion Picture Engineers*, vol. 37, no. 7, 1941. DOI: 10.5594/J12890.

[337] F. Rumsey, *Spatial Audio*. Focal Press, 2001.

[338] T. Holman, *Surround Sound: Up and running*. Focal Press, 2008.

[339] M. Kerins, *Beyond Dolby (stereo): Cinema in the digital sound age*. Indiana University Press, 2011.

[340] M. Lopez, G. Kearney, and K. Hofstädter, "Seeing films through sound: Sound design, spatial audio, and accessibility for visually impaired audiences," *British Journal of Visual Impairment*, vol. 40, pp. 117–144, 2 May 2022, ISSN: 17445809. DOI: 10.1177/0264619620935935.

[341] H. A. Witkin, S. Wapner, and T. Leventhal, "Sound localization with conflicting visual and auditory cues.," *Journal of Experimental Psychology*, vol. 43, pp. 58–67, 1 1952, ISSN: 0022-1015. DOI: 10.1037/h0055889.

[342] *Dive into sound reimagined with dolby atmos*, 2020. [Online]. Available: https://www.dolby.com/en-gb/technologies/dolby-atmos/.

[343] M. Lalwani, *For vr to be truly immersive, it needs convincing sound to match*, 2016. [Online]. Available: https://www.engadget.com/2016-01-22-vr-needs-3d-audio.html.

[344] L. T. Nielsen *et al.*, "Missing the point: An exploration of how to guide users' attention during cinematic virtual reality," in *Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology*, ser. VRST '16, Munich, Germany: Association for Computing Machinery, 2016, 229–232, ISBN: 9781450344913. DOI: 10.1145/2993369.2993405. [Online]. Available: https://doi.org/10.1145/2993369.2993405.

[345] M. Gödde, F. Gabler, D. Siegmund, and A. Braun, "Cinematic narration in vr – rethinking film conventions for 360 degrees," in *Virtual, Augmented and Mixed Reality: Applications in Health, Cultural Heritage, and Industry*, J. Y. Chen and G. Fragomeni, Eds., Cham: Springer International Publishing, 2018, pp. 184–201, ISBN: 978-3-319-91584-5.

[346] A. Sheikh, A. Brown, Z. Watson, and M. Evans, "Directing attention in 360-degree video," in *IBC 2016 Conference*, Institution of Engineering and Technology, 2016, ISBN: 9781785613432. DOI: 10.1049/ibc.2016.0029.

[347] S. Rothe and H. Hußmann, "Guiding the viewer in cinematic virtual reality by diegetic cues," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10850 LNCS, Springer Verlag, 2018, pp. 101–117, ISBN: 9783319952697. DOI: 10.1007/978-3-319-95270-3_7.

[348] L. Matney, "Facebook just bought vr audio company two big ears and is making their tech free to developers," *TechCrunch*, May 2016.

[349] J. Paterson and O. Kadel, "Immersive audio post-production for 360º video: Workflow case studies," in *International Conference on Immersive and Interactive Audio*, Mar. 2019. [Online]. Available: http://www.aes.org/e-lib.

[350] G. Zdanowicz and S. Bambrick, "The game audio strategy guide : A practical course," eng, in New York : Routledge, Taylor & Francis Group, 2020, ch. 2, ISBN: 9781138498334.

[351] L. Murray, *Sound Design Theory and Practice: Working with Sound*, eng, 1st ed. Milton: Routledge, 2019, ISBN: 9781138125407.

[352] C. Baume, M. D. Plumbley, and J. Ćalić, "Use of audio editors in radio production," *138th Audio Engineering Society Convention 2015*, vol. 1, pp. 144–153, 2015.

[353] C. Baume, M. D. Plumbley, D. Frohlich, and J. Ćalić, "Paperclip: A digital pen interface for semantic speech editing in radio production," *AES: Journal of the Audio Engineering Society*, vol. 66, pp. 241–252, 4 Apr. 2018, ISSN: 15494950. DOI: 10.17743/jaes.2018.0006.

[354] L. Ward, M. Glancy, S. Bowman, and M. Armstrong, "The impact of new forms of media on production tools and practices," 2020.

[355] C. Cieciura, M. Glancy, and P. J. Jackson, "Producing personalised object-based audio-visual experiences: An ethnographic study," in *Proceedings of the 2023 ACM International Conference on Interactive Media Experiences*, ser. IMX '23, Nantes, France: Association for Computing Machinery, 2023, 71–82, ISBN: 9798400700286. DOI: 10.1145/3573381.3596156. [Online]. Available: https://doi.org/10.1145/3573381.3596156.

[356] A. A. C. B. Reis, "Immersive media, social change, and creativity: A framework for designing collaborative 360° video productions," University of Porto, Jan. 2021.

[357] H. Esmaeili, H. Thwaites, and P. C. Woods, "Workflows and challenges involved in creation of realistic immersive virtual museum, heritage, and tourism experiences: A comprehensive reference for 3d asset capturing," in *2017 13th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, 2017, pp. 465–472. DOI: 10.1109/SITIS.2017.82.

[358] D. Candusso, "The immersive cinematic sound space: Audience perspectives," in *Proceedings of the 13th Annual Australian Screen Production Education and Research Association (ASPERA) Conference*, S Kerrigan, K Dooley, and B Frankham, Eds., Australia: Australian Screen Production Education and Research Association (ASPERA), 2016, pp. 1–15. [Online]. Available: https://eprints.qut.edu.au/197011/.

[359] D. Candusso, "Designing a sonic landscape: A practice-led approach to creating 3-d sound space for screen," *Fusion*, pp. 144–161, 10 Dec. 2016.

[360] S. Bhangu, F. Provost, and C. Caduff, "Introduction to qualitative research methods – part i," *Perspectives in Clinical Research*, vol. 14, p. 39, 1 2023, ISSN: 2229-3485. DOI: 10.4103/picr.picr_253_22.

[361] N. Surawy-Stepney, F. Provost, S. Bhangu, and C. Caduff, "Introduction to qualitative research methods: Part 2," *Perspectives in Clinical Research*, vol. 14, p. 95, 2 2023, ISSN: 2229-3485. DOI: 10.4103/picr.picr_37_23.

[362] P. Luff, J Hindmarsh, and C Heath, *Workplace Studies: Recovering Work Practice and Informing System Design.* Cambridge, UK: Cambridge University Press, 2000.

[363] E. Denny and A. Weckesser, "How to do qualitative research?" *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 129, pp. 1166–1167, 7 Jun. 2022, ISSN: 1470-0328. DOI: 10.1111/1471-0528.17150.

[364] J. McCambridge, J. Witton, and D. R. Elbourne, "Systematic review of the hawthorne effect: New concepts are needed to study research participation effects," *Journal of Clinical Epidemiology*, vol. 67, pp. 267–277, 3 Mar. 2014, ISSN: 08954356. DOI: 10.1016/j.jclinepi.2013.08.015.

[365] M. L. Ranney, Z. F. Meisel, E. K. Choo, A. C. Garro, C. Sasson, and K. M. Guthrie, "Interview-based qualitative research in emergency care part ii: Data collection, analysis and results reporting," *Academic Emergency Medicine*, vol. 22, pp. 1103–1112, 9 Sep. 2015, ISSN: 1069-6563. DOI: 10.1111/acem.12735.

[366] J. W. Creswell, *Research design : qualitative, quantitative, and mixed methods approaches / John W. Creswell.* eng, 3rd ed. Thousand Oaks, CA ; London: Sage Publications, 2008, ISBN: 9781412965569.

[367] V. Braun, V. Clarke, E. Boulton, L. Davey, and C. McEvoy, "The online survey as a qualitative research tool," *International journal of social research methodology*, vol. 24, no. 6, pp. 641–654, 2021.

[368] M. Debenham, "Computer mediated communication and disability support: Addressing barriers to study for undergraduate distance learners with long-term health problems," The Open University, 2001.

[369] C. Baume, M. D. Plumbley, and J. Calic, "Use of audio editors in radio production," *Journal of The Audio Engineering Society*, 2015.

[370] C. McNamara, *General guidelines for conducting interviews*, 2009. [Online]. Available: {https://management.org/businessresearch/interviews.htm}.

[371] N. C. Jenn, "Designing a questionnaire.," *Malaysian family physician : the official journal of the Academy of Family Physicians of Malaysia*, vol. 1, pp. 32–5, 1 2006, ISSN: 1985-207X.

[372] Y. Baruch, "Response Rate in Academic Studies - A Comparative Analysis," *Human Relations*, vol. 52, no. 4, 1999.

[373] QSR International Pty Ltd, *NVivo (released in March 2020)*, 2020. [Online]. Available: {http://www.s3a-spatialaudio.org/wp-content/uploads/2019/10/userdoc-0.12.0.pdf}.

[374] Qualtrics, *Qualtrics (2020), Provo, UT, United States*, 2020. [Online]. Available: {http://www.s3a-spatialaudio.org/wp-content/uploads/2019/10/userdoc-0.12.0.pdf}.

[375] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, 2006, ISSN: 14780887. DOI: 10.1191/1478088706qp063oa.

[376] D. R. Thomas, "A General Inductive Approach for Analyzing Qualitative Evaluation Data," *American Journal of Evaluation*, vol. 27, no. 2, pp. 237–246, 2006.

[377] L Ermi and F Mäyrä, "Fundamental components of the gameplay experience: Analysing immersion," 2005.

[378] G Calleja, "Revising Immersion: A conceptual Model for the Analysis of Digital Game Involvement," in *3rd Digital Games Research Association International Conference: Situated Play*, 2007.

[379] A. McArthur, "Disparity in horizontal correspondence of sound and source positioning: The impact on spatial presence for cinematic VR," *AES Conference on Audio for Virtual and Augmented Reality*, 2016.

[380] M. A. M. Izhar, M. Volino, A. Hilton, and P. Jackson, "Tracking sound sources for object-based spatial audio in 3d audio-visual production," *Forum Acusticum*, pp. 2051–2058, 2020. DOI: 10.48465/fa.2020.0884. [Online]. Available: https://hal.science/hal-03235364.

[381] R. Mahler, "Approximate multisensor cphd and phd filters," in *13th International Conference on Information Fusion*, 2010, pp. 1–8.

[382] D. Berghi, A. Hilton, and P. J. Jackson, "Visually supervised speaker detection and localization via microphone array," IEEE, Oct. 2021, pp. 1–6, ISBN: 978-1-6654-3288-7. DOI: 10.1109/MMSP53017.2021.9733678. [Online]. Available: https://ieeexplore.ieee.org/document/9733678/.

[383] C. Gan, H. Zhao, P. Chen, D. D. Cox, and A. Torralba, "Self-supervised moving vehicle tracking with stereo sound," 2019.

[384] A. B. Vasudevan, D. Dai, and L. Gool, "Semantic object prediction and spatial sound super-resolution with binarual sounds," 2020.

[385] P. Coleman, A. Franck, P. J. Jackson, R. J. Hughes, L. Remaggi, and F. Melchior, "Object-based reverberation for spatial audio," vol. 65, Audio Engineering Society, Jan. 2017, pp. 66–77. DOI: 10.17743/jaes.2016.0059.

[386] P. Coleman, A. Franck, D. Menzies, and P. J. B. Jackson, "Object-based reverberation encoding from first-order ambisonic rirs," May 2017. [Online]. Available: http://www.aes.org/e-lib.

[387] D. Brungart, "Preliminary model of auditory distance perception for nearby sources," *Computational models of auditory function*, pp. 83–96, 2001.

[388] Y. Suzuki, S. Takane, H.-Y. Kim, and T. Sone, "A modeling of distance perception based on an auditory parallax model," *The Journal of the Acoustical Society of America*, vol. 103, no. 5_Supplement, pp. 3083–3083, 1998.

[389] H.-Y. Kim, Y. Suzuki, S. Takane, and T. Sone, "Control of auditory distance perception based on the auditory parallax model," *Applied Acoustics*, vol. 62, no. 3, pp. 245–270, 2001.

[390] A. Kan, C. Jin, and A. van Schaik, "A psychophysical evaluation of near-field head-related transfer functions synthesized using a distance variation function," *The Journal of the Acoustical Society of America*, vol. 125, pp. 2233–2242, 4 Apr. 2009, ISSN: 0001-4966. DOI: 10.1121/1.3081395.

[391] Z.-W. Chen, G.-Z. Yu, B.-S. Xie, and S.-Q. Guan, "Calculation and analysis of near-field head-related transfer functions from a simplified head-neck-torso model," *Chinese Physics Letters*, vol. 29, no. 3, p. 034 302, 2012.

[392] J. Xu, X. Wang, M. Zhang, C. Yang, and G. Gao, "Binaural sound source distance reproduction based on distance variation function and artificial reverberation," in *MultiMedia Modeling: 23rd International Conference, MMM 2017, Reykjavik, Iceland, January 4-6, 2017, Proceedings, Part II 23*, Springer, 2017, pp. 101–111.

[393] S. Spagnol, E. Tavazzi, and F. Avanzini, "Distance rendering and perception of nearby virtual sound sources with a near-field filter model," *Applied Acoustics*, vol. 115, pp. 61–73, 2017, ISSN: 0003-682X. DOI: https://doi.

org/10.1016/j.apacoust.2016.08.015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X16302407.

[394] M. Zhang, Y. Qiao, X. Wu, and T. Qu, "Distance-dependent modeling of head-related transfer functions," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 276–280. DOI: 10.1109/ICASSP.2019.8683756.

[395] S.-W. Jeon, Y.-C. Park, and D. H. Youn, "Auditory distance rendering based on icpd control for stereophonic 3d audio system," *IEEE Signal Processing Letters*, vol. 22, no. 5, pp. 529–533, 2015. DOI: 10.1109/LSP.2014.2363455.

[396] M.-V. Laitinen, A. Walther, J. Plogsties, and V. Pulkki, "Auditory distance rendering using a standard 5.1 loudspeaker layout," *Audio Engineering Society Convention 139*, pp. 1–7, 2015. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=17991.

[397] M. Yiwere and E. J. Rhee, "Sound source distance estimation using deep learning: An image classification approach," *Sensors (Switzerland)*, vol. 20, no. 1, 2020, ISSN: 14248220. DOI: 10.3390/s20010172.

[398] G. Bologni, R. Heusdens, and J. Martinez, "Acoustic Reflectors Localization from Stereo Recordings Using Neural Networks," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 1–5, ISBN: 978-1-7281-7605-5. DOI: 10.1109/ICASSP39728.2021.9414473. [Online]. Available: https://ieeexplore.ieee.org/document/9414473/.

[399] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2016-May, pp. 405–409, 2016, ISSN: 15206149. DOI: 10.1109/ICASSP.2016.7471706.

[400] B.-S. Xie, "Recovery of individual head-related transfer functions from a small set of measurementsa)," *The Journal of the Acoustical Society of America*, vol. 132, no. 1, pp. 282–294, Jul. 2012, ISSN: 0001-4966. DOI: 10.1121/1.4728168. eprint: https://pubs.aip.org/asa/jasa/article-pdf/132/1/282/15299258/282_1_online.pdf. [Online]. Available: https://doi.org/10.1121/1.4728168.

[401] M. Pezzoli, F. Antonacci, and A. Sarti, *Implicit neural representation with physics-informed neural networks for the reconstruction of the early part of room impulse responses*, 2023. arXiv: 2306.11509 [eess.AS].

[402] T.-H. Tan, Y.-T. Lin, Y.-L. Chang, and M. Alkhaleefah, "Sound source localization using a convolutional neural network and regression model," *Sensors*, vol. 21, no. 23, 2021, ISSN: 1424-8220. DOI: 10.3390/s21238031. [Online]. Available: https://www.mdpi.com/1424-8220/21/23/8031.

[403] P. W. Anderson and P. Zahorik, "Auditory/visual distance estimation: Accuracy and variability," *Frontiers in Psychology*, vol. 5, no. SEP, pp. 1–11, 2014, ISSN: 16641078. DOI: 10.3389/fpsyg.2014.01097.

[404] S. Kraft and U. Zölzer, "Stereo signal separation and upmixing by mid-side decomposition in the frequency-domain," *DAFx 2015 - Proceedings of the 18th International Conference on Digital Audio Effects*, pp. 1–6, 2015.

[405] S. Y. Park, C. J. Chun, and H. K. Kim, "Subband-based upmixing of stereo to 5.1-channel audio signals using deep neural networks," *2016 International Conference on Information and Communication Technology Convergence, ICTC 2016*, pp. 377–380, 2015 2016. DOI: 10.1109/ICTC.2016.7763500.

[406] J. Choi and J.-H. Chang, "Exploiting Deep Neural Networks for Two-to-Five Channel Surround Decoder," *Journal of the Audio Engineering Society*, vol. 68, no. 12, pp. 938–949, 2021, ISSN: 15494950. DOI: 10.17743/jaes.2020.0020.

[407] A. Walther and C. Faller, "Direct-ambient decomposition and upmix of surround signals," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 277–280, 2011. DOI: 10.1109/ASPAA.2011.6082279.

[408] M. V. Laitinen and V. Pulkki, "Converting 5.1 audio recordings to B-format for directional audio coding reproduction," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 61–64, 2011, ISSN: 15206149. DOI: 10.1109/ICASSP.2011.5946328.

[409] M. V. Laitinen, "Converting two-channel stereo signals to B-format for directional audio coding reproduction," *137th Audio Engineering Society Convention 2014*, pp. 314–319, 2014.

[410] P. Morgado, Y. Li, and N. Vasconcelos, "Learning representations from audio-visual spatial alignment," 2020.

[411] J. He, "Spatial audio reproduction using primary ambient extraction," PhD, Nanyang Technological University, 2016.

[412] BBC Academy, *Spatial audio: Where do i start?* 2020. [Online]. Available: https://www.bbc.com/academy-guides/spatial-audio-where-do-i-start.

[413] J. Janai, F. Guney, A. Behl, and A. Geiger, "Computer Vision for Autonomous Vehicles: Problems, Datasets and State of the Art," *Foundations and Trends® in Computer Graphics and Vision*, vol. 12, no. 1-3, pp. 1–308, 2020. DOI: 10.1561/0600000079.

[414] D. Acharya, K. Khoshelham, and S. Winter, "Real-time detection and tracking of pedestrians in CCTV images using a deep convolutional neural network," in *Proceedings of the 4th Annual Conference of Research@Locate*, Sydney, Australia, 2017, pp. 31–36.

[415] E. A. Torres-gallegos, F. Orduña-bustamante, and F. Arámbula-cosío, "Personalization of head-related transfer functions (HRTF) based on automatic photo-anthropometry and inference from a database," *Applied*

Acoustics, vol. 97, pp. 84–95, 2015, ISSN: 0003-682X. DOI: 10.1016/j.apacoust.2015.04.009. [Online]. Available: http://dx.doi.org/10.1016/j.apacoust.2015.04.009.

[416] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning : A Survey and Taxonomy," *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE,*, vol. 41, no. 2, pp. 423–443, 2019.

[417] Y. Kajihara, S. Dozono, and N. Tokui, "Imaginary Soundscape : Cross-Modal Approach to Generate Pseudo Sound Environments," *Workshop on Machine Learning for Creativity and Design (NIPS 2017)*, no. Nips, pp. 1–3, 2017. [Online]. Available: https://nips2017creativity.github.io/doc/Imaginary{_}Soundscape.pdf.

[418] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually Indicated Sounds," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2016, pp. 2405–2413. arXiv: 1512.08512. [Online]. Available: http://arxiv.org/abs/1512.08512http://ieeexplore.ieee.org/document/7780633/.

[419] J. Huang *et al.*, "Speed/accuracy trade-offs for modern convolutional object detectors," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017, pp. 3296–3305, 2017. arXiv: arXiv:1611.10012v3.

[420] M. Abadi *et al.*, *TensorFlow: Large-scale machine learning on heterogeneous systems*, Software available from tensorflow.org, 2015. [Online]. Available: https://www.tensorflow.org/.

[421] L. Vladimirov, *Detect objects using your webcam*, 2018. [Online]. Available: https://tensorflow-object-detection-api-tutorial.readthedocs.io/en/tensorflow-1.14/camera.html.

[422] Tensorflow, *Tensorflow 1 detection model zoo*, 2018. [Online]. Available: https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/tf1_detection_zoo.md.

[423] T. Y. Lin *et al.*, "Microsoft COCO: Common objects in context," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8693 LNCS, no. PART 5, pp. 740–755, 2014. arXiv: arXiv:1405.0312v3.

[424] M. Zhu and M. Liu, "Mobile Video Object Detection with Temporally-Aware Feature Maps," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 5686–5695, 2018. arXiv: 1711.06368.

[425] K. Kang *et al.*, "T-CNN: Tubelets with Convolutional Neural Networks for Object Detection from Videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2896–2907, 2018. arXiv: 1604.02532.

[426] P. Jaccard, "The Distribution of Flora in the Alpine zone," *The New Phytologist*, vol. 11, no. 2, 1912. DOI: 10.1111/j.1469-8137.1912.tb05611.x.

[427] H. Rezatofighi, N. Tsoi, J. Gwak, I. Reid, and S. Savarese, "Generalized Intersection over Union : A Metric and A Loss for Bounding Box Regression," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA: IEEE, 2019, ISBN: 9781728132938. DOI: 10.1109/CVPR.2019.00075. [Online]. Available: https://ieeexplore.ieee.org/document/8953982.

[428] A. Rosebrock, *Intersection over union (iou) for object detection*, 2016.

[429] B. B. Corporation, *Remarc license*, 2023. [Online]. Available: https://sound-effects.bbcrewind.co.uk/licensing.

[430] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, pp. 1733–1746, 10 2015, ISSN: 15209210. DOI: 10.1109/TMM.2015.2428998.

[431] M. Wright, "Open sound control 1.0 specification," 2002. [Online]. Available: http://opensoundcontrol.org/spec-1_0.

[432] Cockos, *Reaper — audio production without limits*, 2019. [Online]. Available: https://www.reaper.fm/.

[433] F. Iida, Y. Minekawa, J. Rummel, and A. Seyfarth, "Toward a human-like biped robot with compliant legs," *Robotics and Autonomous Systems*, vol. 57, no. 2, pp. 139–144, 2009, ISSN: 09218890. DOI: 10.1016/j.robot.2007.12.001.

[434] W. D. Hairston, M. T. Wallace, J. W. Vaughan, B. E. Stein, J. L. Norris, and J. A. Schirillo, "Visual localization ability influences cross-modal bias," *Journal of Cognitive Neuroscience*, vol. 15, pp. 20–29, 1 Jan. 2003, ISSN: 0898-929X. DOI: 10.1162/089892903321107792.

[435] C. E. Jack and W. R. Thurlow, "Effects of degree of visual association and angle of displacement on the "ventriloquism" effect," *Perceptual and Motor Skills*, vol. 37, pp. 967–979, 3 Dec. 1973, ISSN: 0031-5125. DOI: 10.1177/003151257303700360.

[436] H. Stenzel, J. Francombe, and P. J. Jackson, "Limits of perceived audio-visual spatial coherence as defined by reaction time measurements," *Frontiers in Neuroscience*, vol. 13, no. MAY, pp. 1–17, 2019, ISSN: 1662453X. DOI: 10.3389/fnins.2019.00451.

[437] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, and T. K. Kim, "Multiple object tracking: A literature review," *Artificial Intelligence*, vol. 293, p. 103 448, 2021, ISSN: 00043702. DOI: 10.1016/j.artint.2020.103448. arXiv: 1409.7618. [Online]. Available: https://doi.org/10.1016/j.artint.2020.103448.

[438] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[439] J. Pennington, R. Socher, and C. D. Manning, "GloVe : Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. [Online]. Available: https://aclanthology.org/D14-1162.

[440] M. Al-Faris, J. Chiverton, D. Ndzi, and A. I. Ahmed, "A review on computer vision-based methods for human action recognition," *Journal of Imaging*, vol. 6, no. 6, 2020, ISSN: 2313-433X. DOI: 10.3390/jimaging6060046. [Online]. Available: https://www.mdpi.com/2313-433X/6/6/46.

[441] L. Xie, F. Lee, L. Liu, K. Kotani, and Q. Chen, "Scene recognition: A comprehensive survey," *Pattern Recognition*, vol. 102, p. 107 205, 2020, ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2020.107205. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S003132032030011X.

[442] A. M.V. and D. M. Khan, "Recent trends on object detection and image classification: A review," in *2020 International Conference on Computational Performance Evaluation (ComPE)*, 2020, pp. 427–435. DOI: 10.1109/ComPE49325.2020.9200080.

[443] X. Wu, D. Sahoo, and S. C. Hoi, "Recent advances in deep learning for object detection," *Neurocomputing*, vol. 396, pp. 39–64, 2020, ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2020.01.085. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231220301430.

[444] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[445] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[446] W. Liu *et al.*, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, Springer, 2016, pp. 21–37.

[447] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[448] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[449] C. Li *et al.*, "Yolov6: A single-stage object detection framework for industrial applications," *arXiv preprint arXiv:2209.02976*, 2022.

[450] S. Wu, X. Li, and X. Wang, "Iou-aware single-stage object detector for accurate localization," *Image and Vision Computing*, vol. 97, p. 103 911, 2020.

[451] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[452] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[453] R. Girshick, "Fast r-cnn," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448. DOI: 10.1109/ICCV.2015.169.

[454] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[455] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," *Advances in neural information processing systems*, vol. 29, 2016.

[456] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask scoring r-cnn," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6409–6418.

[457] L. Jiao *et al.*, "New generation deep learning for video object detection: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 8, pp. 3195–3215, 2022. DOI: 10.1109/TNNLS.2021.3053249.

[458] S. K. Pal, A. Pramanik, J. Maiti, and P. Mitra, "Deep learning in multi-object detection and tracking: State of the art," *Applied Intelligence*, vol. 51, pp. 6400–6429, 9 Sep. 2021, ISSN: 0924-669X. DOI: 10.1007/s10489-021-02293-7.

[459] J. Beal, E. Kim, E. Tzeng, D. H. Park, A. Zhai, and D. Kislyuk, "Toward transformer-based object detection," *CoRR*, vol. abs/2012.09958, 2020. arXiv: 2012.09958. [Online]. Available: https://arxiv.org/abs/2012.09958.

[460] Y. Xu, X. Zhou, S. Chen, and F. Li, "Deep learning for multiple object tracking: A survey," *IET Computer Vision*, vol. 13, pp. 355–368, 4 Jun. 2019, ISSN: 1751-9632. DOI: 10.1049/iet-cvi.2018.5598.

[461] S. Sun, N. Akhtar, H. Song, A. Mian, and M. Shah, "Deep affinity network for multiple object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 104–119, 2021. DOI: 10.1109/TPAMI.2019.2929520.

[462] B. Mirzaei, H. Nezamabadi-Pour, A. Raoof, and R. Derakhshani, "Small object detection and tracking: A comprehensive review.," *Sensors (Basel,*

Switzerland), vol. 23, 15 Aug. 2023, ISSN: 1424-8220. DOI: 10.3390/s23156887.

[463] L. Zhang and L. van der Maaten, "Preserving structure in model-free tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 4, pp. 756–769, 2014. DOI: 10.1109/TPAMI.2013.221.

[464] J. Xie, E. Stensrud, and T. Skramstad, "Detection-based object tracking applied to remote ship inspection," *Sensors*, vol. 21, no. 3, 2021, ISSN: 1424-8220. DOI: 10.3390/s21030761. [Online]. Available: https://www.mdpi.com/1424-8220/21/3/761.

[465] H. Bai, W. Cheng, P. Chu, J. Liu, K. Zhang, and H. Ling, "Gmot-40: A benchmark for generic multiple object tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 6719–6728.

[466] R. Sanchez-Matilla, F. Poiesi, and A. Cavallaro, "Online multi-target tracking with strong and weak detections," in H. G. and H. Jégou, Eds. 2016, pp. 84–99. DOI: 10.1007/978-3-319-48881-3_7.

[467] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 3464–3468. DOI: 10.1109/ICIP.2016.7533003.

[468] L. Chen, H. Ai, C. Shang, Z. Zhuang, and B. Bai, "Online multi-object tracking with convolutional neural networks," in *2017 IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 645–649. DOI: 10.1109/ICIP.2017.8296360.

[469] Q. Chu, W. Ouyang, H. Li, X. Wang, B. Liu, and N. Yu, "Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4846–4855. DOI: 10.1109/ICCV.2017.518.

[470] L. Leal-Taixe, C. Canton-Ferrer, and K. Schindler, "Learning by tracking: Siamese cnn for robust target association," in *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2016.

[471] S. Schulter, P. Vernaza, W. Choi, and M. Chandraker, "Deep network flow for multi-object tracking," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2730–2739. DOI: 10.1109/CVPR.2017.292.

[472] R. Kasturi *et al.*, "Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 319–336, 2009. DOI: 10.1109/TPAMI.2008.57.

[473] Y. Li, C. Huang, and R. Nevatia, "Learning to associate: Hybridboosted multi-target tracker for crowded scene," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2953–2960. DOI: 10.1109/CVPR.2009.5206735.

[474] L. Leal-Taixé, A. Milan, K. Schindler, D. Cremers, I. D. Reid, and S. Roth, "Tracking the trackers: An analysis of the state of the art in multiple object tracking," *CoRR*, vol. abs/1704.02781, 2017. arXiv: 1704.02781. [Online]. Available: http://arxiv.org/abs/1704.02781.

[475] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "MOTChallenge 2015: Towards a benchmark for multi-target tracking," *arXiv:1504.01942 [cs]*, Apr. 2015, arXiv: 1504.01942. [Online]. Available: http://arxiv.org/abs/1504.01942.

[476] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "MOT16: A benchmark for multi-object tracking," *arXiv:1603.00831 [cs]*, Mar. 2016, arXiv: 1603.00831. [Online]. Available: http://arxiv.org/abs/1603.00831.

[477] P. Dendorfer *et al.*, "MOT20: A benchmark for multi object tracking in crowded scenes," *CoRR*, vol. abs/2003.09003, 2020. arXiv: 2003.09003. [Online]. Available: https://arxiv.org/abs/2003.09003.

[478] Q. Wang, Y. Zheng, P. Pan, and Y. Xu, "Multiple object tracking with correlation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 3876–3886.

[479] A. Sadeghian, A. Alahi, and S. Savarese, "Tracking the untrackable: Learning to track multiple cues with long-term dependencies," *CoRR*, vol. abs/1701.01909, 2017. arXiv: 1701.01909. [Online]. Available: http://arxiv.org/abs/1701.01909.

[480] J. Zhu, H. Yang, N. Liu, M. Kim, W. Zhang, and M.-H. Yang, "Online multi-object tracking with dual matching attention networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[481] S. Tang, B. Andres, M. Andriluka, and B. Schiele, "Multi-person tracking by multicut and deep matching," in *Computer Vision – ECCV 2016 Workshops*, G. Hua and H. Jégou, Eds., Cham: Springer International Publishing, 2016, pp. 100–111, ISBN: 978-3-319-48881-3.

[482] P. Sun *et al.*, "Transtrack: Multiple-object tracking with transformer," *CoRR*, vol. abs/2012.15460, 2021. arXiv: 2012.15460. [Online]. Available: https://arxiv.org/abs/2012.15460.

[483] K. H. Tran *et al.*, *Z-gmot: Zero-shot generic multiple object tracking*, 2023. arXiv: 2305.17648 [cs.CV].

[484] A. Dave, T. Khurana, P. Tokmakov, C. Schmid, and D. Ramanan, "TAO: A large-scale benchmark for tracking any object," *CoRR*, vol. abs/2005.10356, 2020. arXiv: 2005.10356. [Online]. Available: https://arxiv.org/abs/2005.10356.

[485] Q. Fan, C.-K. Tang, and Y.-W. Tai, "Few-shot video object detection," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., Cham: Springer Nature Switzerland, 2022, pp. 76–98, ISBN: 978-3-031-20044-1.

[486] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[487] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[488] J. Ferryman and A. Shahrokni, "Pets2009: Dataset and challenge," in *2009 Twelfth IEEE international workshop on performance evaluation of tracking and surveillance*, IEEE, 2009, pp. 1–6.

[489] J.-P. Jodoin, G.-A. Bilodeau, and N. Saunier, "Tracking all road users at multimodal urban traffic intersections," *IEEE transactions on intelligent transportation systems*, vol. 17, no. 11, pp. 3241–3251, 2016.

[490] A. Pramanik, S. K. Pal, J. Maiti, and P. Mitra, "Granulated rcnn and multi-class deep sort for multi-object detection and tracking," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 6, no. 1, pp. 171–181, 2022. DOI: 10.1109/TETCI.2020.3041019.

[491] L. H. Li *et al.*, "Grounded language-image pre-training," *CoRR*, vol. abs/2112.03857, 2021. arXiv: 2112.03857. [Online]. Available: https://arxiv.org/abs/2112.03857.

[492] J. Cao, J. Pang, X. Weng, R. Khirodkar, and K. Kitani, *Observation-centric sort: Rethinking sort for robust multi-object tracking*, 2023. arXiv: 2203.14360 [cs.CV].

[493] K. Jo, J. Im, J. Kim, and D.-S. Kim, "A real-time multi-class multi-object tracker using yolov2," in *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, 2017, pp. 507–511. DOI: 10.1109/ICSIPA.2017.8120665.

[494] B. Lee, E. Erdenee, S. Jin, and P. Rhee, "Multi-class multi-object tracking using changing point detection," *CoRR*, vol. abs/1608.08434, 2016. arXiv: 1608.08434. [Online]. Available: http://arxiv.org/abs/1608.08434.

[495] S. Finnie, F.-l. Zhang, and T. Rhee, "Visual object tracking in spherical 360° videos: A bridging approach," in *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, 2020, pp. 1–6. DOI: 10.1109/IVCNZ51579.2020.9290549.

[496] A. Delforouzi, S. A. H. Tabatabaei, K. Shirahama, and M. Grzegorzek, "A polar model for fast object tracking in 360-degree camera images," *Multimedia Tools and Applications*, vol. 78, pp. 9275–9297, 7 Apr. 2019, ISSN: 1380-7501. DOI: 10.1007/s11042-018-6525-0.

[497] S. Kraft, "Stereo Signal Decomposition and Upmixing to Surround and 3D Audio," PhD, Hamburg, 2022. DOI: 10.24405/14379. [Online]. Available: https://openhsu.ub.hsu-hh.de/handle/10.24405/14379.

[498] C. Avendano and J. M. Jot, "A frequency-domain approach to multichannel upmix," *Journal of the Audio Engineering Society*, vol. 52, no. 7-8, pp. 740–749, 2004, ISSN: 15494950.

[499] S. Kraft and U. Zölzer, "Low-complexity stereo signal decomposition and source separation for application in stereo to 3D upmixing," in *140th Audio Engineering Society International Convention 2016, AES 2016*, 2016.

[500] C. Avendano and J. Jot, "Frequency domain techniques for stereo to multichannel upmix," in *22nd International Audio Engineering Society Conference on Virtual, Synthetic and Entertainment Audio*, Audio Engineering Society, 2002.

[501] C. J. Chun, Y. G. Kim, J. Y. Yang, and H. K. Kim, "Real-time conversion of stereo audio to 5.1 channel audio for providing realistic sounds," *International Journal of Signal Processing, Image processing and Pattern Recognition*, vol. 2, pp. 85–94, 4 2009.

[502] M. R. Bai, H. Hsu, and J.-C. Wen, "Spatial sound field synthesis and upmixing based on the equivalent source method," *The Journal of the Acoustical Society of America*, vol. 135, no. 1, pp. 269–282, 2014, ISSN: 0001-4966. DOI: 10.1121/1.4835815.

[503] J. Usher, "Design criteria for high quality upmixers," in *28th AES International Conference on The Future of Audio Technology–Surround and Beyond*, 2006, pp. 1–13.

[504] C. Faller, L. Altmann, J. Levison, and M. Schmidt, "Multi-channel ring upmix," in *134th Audio Engineering Society Convention 2013*, 2013, pp. 736–741, ISBN: 9781627485715.

[505] C. Avendano, "Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2003.

[506] M. Goodwin and J. Jot, "Primary-ambient signal decomposition and vector-based localisation for spatial audio coding and enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2007.

[507] M. S. Vinton, M. F. Davis, and C. Q. Robinson, "Signal models and upmixing techniques for generating multichannel audio," *Proceedings of the AES International Conference*, pp. 1–12, 2011.

[508] M. Vinton, D. McGrath, C. Robinson, and P. Brown, "Next generation surround decoding and upmixing for consumer and professional applications," *Proceedings of the AES International Conference*, vol. 2015-Janua, pp. 1–9, 2015.

[509] M. A. Gerzon, "Optimal Reproduction Matrices for Multispeaker Stereo," in *91st Audio Engineering Society Convention*, 1991.

[510] R. Dressler, "Dolby surround pro logic decoder principles of operation," *Dolby White Paper*, 1982.

[511] R. Dressler, "Dolby surround pro logic ii decoder principles of operation," *Dolby White Paper*, 2000.

[512] R. Irwan and R. Aarts, "Two-to-five channel sound processing," *Journal of the Audio Engineering Society*, vol. 11, no. 50, pp. 915–926, 2002.

[513] R. Irwan and R. Aarts, "A method to convert stereo to multi-channel sound," in *19th International Conference of the Audio Engineering Society*, 2001.

[514] J. Usher and J. Benesty, "Fundamental components of the gameplay experience: Analysing immersion," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2141–2150, 2007.

[515] J. Merimaa, M. Goodwin, and J. Jot, "Correlation-based ambience extraction from stereo recordings," in *123rd Audio Engineering Society Convention*, Audio Engineering Society, 2007.

[516] Y. Li and P. Driessen, "An unsupervised adaptive filtering approach of 2-to-5 channel upmix," in *119th Audio Engineering Society Convention*, Audio Engineering Society, 2005.

[517] C. Faller, "Multiple-loudspeaker playback of stereo signals," *Journal of the Audio Engineering Society*, vol. 11, no. 54, pp. 1051–1064, 2006.

[518] S. Jeon, Y. Park, S. Lee, and D. Youn, "Robust Representation of Spatial Sound in Stereo-to-Multichannel Upmix," in *128th Audio Engineering Society Conference*, 2010.

[519] J. He, E. Tan, and W. Gan, "Linear estimation based primary-ambient extraction for stereo audio signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 2, pp. 505–517, 22 2014.

[520] J. He, E. Tan, and W. Gan, "Primary-ambient extraction using ambient spectrum estimation for immersive spatial audio reproduction," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 9, pp. 1431–1444, 23 2015.

[521] K. M. Ibrahim and M. Allam, "Primary-ambient source separation for upmixing to surround sound systems," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2018-April, pp. 431–435, 2018, ISSN: 15206149. DOI: 10.1109/ICASSP.2018.8461459.

[522] E. Vickers, "Frequency-domain two-to-three-channel upmix for center channel derivation and speech enhancement," in *127th Audio Engineering Society Convention*, Audio Engineering Society, 2009.

[523] K. M. Ibrahim and M. Allam, "Primary-ambient extraction in audio signals using adaptive weighting and principal component analysis," in *13th Sound and Music Computing Conference (SMC)*, 2016.

[524] Blue Ripple Sound, *O3A Upmixers*, https://www.blueripplesound.com/products/o3a-upmixers, 2020.

[525] NUDEN Audio, *Halo Upmix*, https://nugenaudio.com/haloupmix, 2020.

[526] Penteo, *Penteo for ambisonics*. [Online]. Available: `https://www.perfectsurround.com/ambisonics`.

[527] H. Cardew, "A Proposed Stereophonic to B-Format Up-Mix Algorithm Using Multilevel Thresholding," Masters, University of Derby, 2016.

[528] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, Oct. 2019. [Online]. Available: `https://dcase.community/documents/workshop2019/proceedings/DCASE2019Workshop_Adavanne_46.pdf`.

[529] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection," Nov. 2021, pp. 125–129. [Online]. Available: `http://arxiv.org/abs/2106.06999`.

[530] A. Politis *et al.*, "A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," Nov. 2022, pp. 125–129.

[531] A. Politis *et al.*, *Sound Event Localization and Detection Evaluated in Real Spatial Sound Scenes: Task Description*, 2022. [Online]. Available: {`https://dcase.community/challenge2022/task-sound-event-`

localization-and-detection-evaluated-in-real-spatial-sound-scenes}.

[532] M. Green and D. Murphy, "Eigenscape: A database of spatial acoustic scene recordings," *Applied Sciences*, vol. 7, p. 1204, 12 2017, ISSN: 2076-3417. DOI: 10.3390/app7111204. [Online]. Available: http://www.mdpi.com/2076-3417/7/11/1204.

[533] M. C. Green, S. Adavanne, D. Murphy, and T. Virtanen, "Acoustic scene classification using higher-order ambisonic features," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, vol. 2019-Octob, pp. 328–332, 2019, ISSN: 19471629. DOI: 10.1109/WASPAA.2019.8937282.

[534] E. Guizzo *et al.*, "L3das21 challenge: Machine learning for 3d audio signal processing," in *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*, 2021, pp. 1–6. DOI: 10.1109/MLSP52302.2021.9596248.

[535] A. Politis, S Adavanne, and T. Virtanen, *Tau spatial room impulse response database (tau-srir db)*, 2022. DOI: 10.5281/zenodo.6408611. [Online]. Available: https://zenodo.org/records/6408611.

[536] T. Lübeck, J. Arend, and C. Pörschmann, "A high-resolution spatial room impulse response database," in *Proceedings of the 47th DAGA*, Aug. 2021.

[537] B. Bernschütz, C. Pörschmann, S. Spors, and S. Weinzierl, "Entwurf und aufbau eines variablen sphärischen mikrofonarrays für forschungsanwendungen in raumakustik und virtual audio," in *Proceedings of the 36th DAGA*, 2010.

[538] P. Lecomte, P. Gauthier, C. Langrenne, A. Berry, and A. Garcia, "A fifty-node lebedev grid and its applications to ambisonics," *Journal of the Audio Engineering Society*, vol. 64, pp. 868–881, 2016. DOI: 10.17743/jaes.2016.0036..

[539] M. Cobos, J. Ahrens, K. Kowalczyk, and A. Politis, "An overview of machine learning and other data-based methods for spatial audio capture,

processing, and reproduction," *Eurasip Journal on Audio, Speech, and Music Processing*, vol. 2022, 1 2022, ISSN: 16874722. DOI: 10.1186/s13636-022-00242-x.

[540] R. Gao and K. Grauman, "2.5d visual sound," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[541] GoPro, *All cameras*, 2023. [Online]. Available: https://gopro.com/en/gb/shop/cameras.

[542] D. Sound, *Free space xlr binaural microphone*, 2023. [Online]. Available: https://3diosound.com/collections/microphones/products/free-space-xlr-binaural-microphone.

[543] A. Hirway, Y. Qiao, and N. Murray, "A qoe and visual attention evaluation on the influence of spatial audio in 360 videos," in *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, 2020, pp. 345–350. DOI: 10.1109/AIVR50618.2020.00071.

[544] A. Hirway, Y. Qiao, and N. Murray, "Spatial audio in 360° videos: Does it influence visual attention?" In *Proceedings of the 13th ACM Multimedia Systems Conference*, ser. MMSys '22, Athlone, Ireland: Association for Computing Machinery, 2022, 39–51, ISBN: 9781450392839. DOI: 10.1145/3524273.3528179. [Online]. Available: https://doi.org/10.1145/3524273.3528179.

[545] I. Trowitzsch, J. Taghia, Y. Kashef, and K. Obermayer, "The nigens general sound events database," *eprint arXiv:1902.08314*, pp. 1–5, Feb. 2019. [Online]. Available: http://arxiv.org/abs/1902.08314.

[546] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, pp. 2421–2424, 5 Nov. 2006, ISSN: 0001-4966. DOI: 10.1121/1.2229005.

[547] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *Proceedings of the 21st ACM international conference on Multimedia*, ACM, Oct. 2013, pp. 411–412, ISBN: 9781450324045. DOI: 10.1145/2502081.2502245.

[548] mh Acoustics, *Eigenunits®: Vst plugins for macos and windows*, English, version 2, mh Acoustics, Oct. 2019, 18 pp.

[549] T. McKenzie, L. McCormack, and C. Hold, "Dataset of spatial room impulse responses in a variable acoustics room for six degrees-of-freedom rendering and analysis," *ArxIV*, Nov. 2021. [Online]. Available: http://arxiv.org/abs/2111.11882.

[550] D. Turner and D. Murphy, *Dataset of stereo and multi-channel IRs for a 50-point Lebedev quadrature.* Version V.1.0, May 2023. DOI: 10.5281/zenodo.7990195. [Online]. Available: https://doi.org/10.5281/zenodo.7990195.

[551] A. Politis, T. Pihlajamäki, and V. Pulkki, "Parametric spatial audio effects," in *15th International Conference on Digital Audio Effects (DAFx-12)*, Sep. 2012.

[552] V. Pulkki, A. Politis, T. Pihlajamaki, and M.-V. Laitinen, "Spatial sound scene synthesis and manipulation for virtual reality and audio effects," in *Parametric Time-Frequency Domain Spatial Audio*, V. Pulkki, S. Delikaris-Manias, and A. Politis, Eds., 1st, John Wiley & Sons, 2018, pp. 347–361.

[553] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *The Journal of the Acoustical Society of America*, vol. 116, pp. 3075–3089, 5 Nov. 2004, ISSN: 0001-4966. DOI: 10.1121/1.1791872.

[554] V. Pulkki and C. Faller, "Directional audio coding: Filterbank and stft-based design," May 2006. [Online]. Available: www.aes.org..

[555] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *13th International Conference on Ar-*

tificial Intelligence and Statistics (AISTATS), 2010. [Online]. Available: http://www.iro.umontreal..

[556] Y.-Y. Yang *et al.*, "Torchaudio: Building blocks for audio and speech processing," *arXiv preprint arXiv:2110.15018*, 2021.

[557] C. Knapp and G. C Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, 4 1976.

[558] J. Choi and J. H. Chang, "Convolutional neural network-based direction-of-arrival estimation using stereo microphones for drone," *2020 International Conference on Electronics, Information, and Communication, ICEIC 2020*, 2020. DOI: 10.1109/ICEIC49074.2020.9051364.

[559] R. Lee, M.-S. Kang, B.-H. Kim, K.-H. Park, S. Q. Lee, and H.-M. Park, "Sound source localization based on gcc-phat with diffuseness mask in noisy and reverberant environments," *IEEE Access*, vol. 8, pp. 7373–7382, 2020. DOI: 10.1109/ACCESS.2019.2963768.

[560] N. Jillings, A. Clifford, and J. D. Reiss, "Performance optimization of gcc-phat for delay and polarity correction under real world conditions," in *134th Audio Engineering Society Convention*, May 2013. [Online]. Available: www.aes.org..

[561] V. S. Kadandale, J. F. Montesinos, G. Haro, and E. Gomez, "Multi-channel u-net for music source separation," *IEEE 22nd International Workshop on Multimedia Signal Processing, MMSP 2020*, 2020. DOI: 10.1109/MMSP48831.2020.9287108.

[562] Z. Zhang, Q. Liu, Y. Wang, and S. Member, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2017.

[563] L. Ou and Y. Chen, "Acoustic bandwidth extension by audio deep residual u-net," in *2022 5th International Conference on Information Communication and Signal Processing, ICICSP 2022*, Institute of Electrical and

Electronics Engineers Inc., 2022, pp. 549–554, ISBN: 9781665485890. DOI: 10.1109/ICICSP55539.2022.10050671.

[564] W. Choi, M. Kim, J. Chung, D. Lee, and S. Jung, "Investigating u-nets with various intermediate blocks for spectrogram-based singing voice separation," in *21st International Society for Music Information Retrieval*, 2020. [Online]. Available: http://arxiv.org/abs/1912.02591.

[565] D. Stoller, S. E. Spotify, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," 2018. [Online]. Available: https://github.com/f90/.

[566] X. Song, Q. Kong, X. Du, and Y. Wang, "Catnet: Music source separation system with mix-audio augmentation," *CoRR*, vol. abs/2102.09966, 2021. arXiv: 2102.09966. [Online]. Available: https://arxiv.org/abs/2102.09966.

[567] S. Venkatesh, "Deep learning for audio segmentation and intelligent remixing," University of Plymouth, Dec. 2022. DOI: 10.24382/778. [Online]. Available: http://hdl.handle.net/10026.1/20092http://dx.doi.org/10.24382/778.

[568] T. Nakamura, S. Kozuka, and H. Saruwatari, "Time-domain audio source separation with neural networks based on multiresolution analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1687–1701, 2021. DOI: 10.1109/TASLP.2021.3072496.

[569] F. Henkel, R. Kelz, and G. Widmer, "Audio-conditioned u-net for position estimation in full sheet images," in *International Workshop on Reading Music Systems 2019 (WoRMS)*, Oct. 2019. [Online]. Available: http://arxiv.org/abs/1910.07254.

[570] D. Y. Wu, Y. H. Chen, and H. Y. Lee, "Vqvc+: One-shot voice conversion by vector quantization and u-net architecture," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2020-October, International Speech Communication

Association, 2020, pp. 4691–4695. DOI: 10.21437/Interspeech.2020-1443.

[571] R. Li, D. Pu, M. Huang, and B. Huang, "Unet-tts: Improving unseen speaker and style transfer in one-shot voice cloning," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8327–8331. DOI: 10.1109/ICASSP43922.2022.9746049.

[572] E. Moliner and V. Välimäki, "A two-stage u-net for high-fidelity denoising of historical recordings," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2022-May, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 841–845, ISBN: 9781665405409. DOI: 10.1109/ICASSP43922.2022.9746977.

[573] M. R. Saddler, A. Francl, J. Feather, K. Qian, Y. Zhang, and J. H. McDermott, *Speech denoising with auditory models*, 2021. arXiv: 2011.10706 [eess.AS].

[574] X. Xu, H. Zhou, Z. Liu, B. Dai, X. Wang, and D. Lin, "Visually informed binaural audio generation without binaural audios," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 15 485–15 494.

[575] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *33rd Conference on Neural Information Processing Systems (NeurIPS)*, Dec. 2019. [Online]. Available: http://arxiv.org/abs/1912.01703.

[576] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[577] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, "Visualizing the loss landscape of neural nets," in *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, 2018. [Online]. Available: https://github.com/tomgoldstein/loss-landscape.

[578] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access*, vol. 9, pp. 82 031–82 057, 2021. DOI: 10.1109/ACCESS.2021.3086020.

[579] G. Kim, D. K. Han, and H. Ko, "Specmix : A mixed sample data augmentation method for training with time-frequency domain features," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 1, International Speech Communication Association, 2021, pp. 6–10, ISBN: 9781713836902. DOI: 10.21437/Interspeech.2021-103.

[580] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. Interspeech 2015*, 2015, pp. 3586–3589. DOI: 10.21437/Interspeech.2015-711.

[581] N. Srivastava, G. Hinton, A. Krizhevsky, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," 2014, pp. 1929–1958.

[582] L. Lu, Y. YeonjongSu, and G. Em Karniadakis, "Dying relu and initialization: Theory and numerical examples," *Communications in Computational Physics*, vol. 28, no. 5, pp. 1671–1706, 2020, ISSN: 1991-7120. DOI: https://doi.org/10.4208/cicp.OA-2020-0165. [Online]. Available: http://global-sci.org/intro/article_detail/cicp/18393.html.

[583] L. Nanni, S. Brahnam, M. Paci, and S. Ghidoni, "Comparison of different convolutional neural network activation functions and methods for building ensembles for small to midsize medical data sets," *Sensors*, vol. 22, 16 Aug. 2022, ISSN: 14248220. DOI: 10.3390/s22166129.

[584] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," 2013.

[585] L. Biewald, *Experiment tracking with weights and biases*, Software available from wandb.com, 2020. [Online]. Available: https://www.wandb.com/.

[586] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," in *5th International Conference on Learning Representations (ICLR)*, Apr. 2017. [Online]. Available: http://arxiv.org/abs/1608.03983.

[587] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7.

[588] P. Seetharaman, G. Wichern, B. Pardo, and J. L. Roux, "Autoclip : Adaptive gradient clipping for source separation networks," in *2020 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING*, IEEE, 2020, ISBN: 9781728166629.

[589] J. Zhang, T. He, S. Sra, and A. Jadbabaie, "Why gradient clipping accelerates training: A theoretical justification for adaptivity," in *8th International Conference on Learning Representations (ICLR)*, May 2020. [Online]. Available: http://arxiv.org/abs/1905.11881.

[590] NVIDIA, *Geforce rtx 3090 family*, 2023. [Online]. Available: https://www.nvidia.com/en-gb/geforce/graphics-cards/30-series/rtx-3090-3090ti/.

[591] T. Wolf, *Training neural nets on larger batches: Practical tips for 1-gpu multi-gpu & distributed setups*, 2018. [Online]. Available: https://medium.com/huggingface/training-larger-batches-practical-tips-on-1-gpu-multi-gpu-distributed-setups-ec88c3e51255.

[592] C. J. Steinmetz and J. D. Reiss, "Auraloss: Audio focused loss functions in PyTorch," in *Digital Music Research Network One-day Workshop (DMRN+15)*, 2020.

[593] S. Arık, H. Jun, and G. Diamos, "Fast spectrogram inversion using multi-head convolutional neural networks," *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 94–98, 2019. DOI: 10.1109/LSP.2018.2880284.

[594] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP*, Oct. 2020. [Online]. Available: http://arxiv.org/abs/1910.11480.

[595] P. Manocha, A. Finkelstein, R. Zhang, N. J. Bryan, G. J. Mysore, and Z. Jin, "A differentiable perceptual audio metric learned from just noticeable differences," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2020-October, International Speech Communication Association, 2020, pp. 2852–2856. DOI: 10.21437/Interspeech.2020-1191.

[596] S. Hughes and G. Kearney, "Moving virtual source perception in 2d space," *Proc. of Audio Engineering Society Conference on Audio for Virtual and Augmented Reality*, pp. 1–9, 2016. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=18492.

[597] A. Politis, *Spherical array processing*, https://github.com/polarch/Spherical-Array-Processing, 2021.

[598] S. Tervo, "Direction estimation based on sound intensity vectors," in *2009 17th European Signal Processing Conference*, 2009, pp. 700–704.

[599] OpenAI, "Gpt-4 technical report," 2023.