Three-Dimensional Geometry Inference of Convex and Non-Convex Rooms using Spatial Room Impulse Responses

Michael James Lovedee-Turner

PhD Thesis

University of York
Electronic Engineering
September 2019

Abstract

This thesis presents research focused on the problem of geometry inference for both convex- and non-convex-shaped rooms, through the analysis of spatial room impulse responses. Current geometry inference methods are only applicable to convex-shaped rooms, requiring between 6–78 discretely spaced measurement positions, and are only accurate under certain conditions, such as a first-order reflection for each boundary being identifiable across all, or some subset of, these measurements. This thesis proposes that by using compact microphone arrays capable of capturing spatiotemporal information, boundary locations, and hence room shape for both convex and non-convex cases, can be inferred, using only a sufficient number of measurement positions to ensure each boundary has a first-order reflection attributable to, and identifiable in, at least one measurement. To support this, three research areas are explored. Firstly, the accuracy of direction-of-arrival estimation for reflections in binaural room impulse responses is explored, using a state-of-the-art methodology based on binaural model fronted neural networks. This establishes whether a two-microphone array can produce accurate enough direction-of-arrival estimates for geometry inference. Secondly, a spherical microphone array based spatiotemporal decomposition workflow for analysing reflections in room impulse responses is explored. This establishes that simultaneously arriving reflections can be individually detected, relaxing constraints on measurement positions. Finally, a geometry inference method applicable to both convex and more complex non-convex shaped rooms is proposed. Therefore, this research expands the possible scenarios in which geometry inference can be successfully applied at a level of accuracy comparable to existing work, through the use of commonly used compact microphone arrays. Based on these results, future improvements to this approach are presented and discussed in detail.

Contents

List of Figures

List of Tables

Acknowledgements

Foremost, I would like to thank my supervisor, Professor Damian Murphy, for allowing me to explore my research into geometry inference. His continuous guidance, support, and enthusiasm throughout has been a major benefit during this PhD. I would like to extend my gratitude to the rest of my supervision team: Marc Green, Dr. Amelia Gully, Dr. Frank Stevens, Dr. Joe Rees-Jones, Dr. Becky Gradwell-Vos, Dr. Catriona Cooper, Dr. Duncan Williams, and Yang Fu for listening to and providing feedback on my research. A special thank you to Andrew Chadwick for helping with equipment and building measurement systems.

I would like to extend a thank you to my colleagues, and good friends, that I shared an office with, Marc Green and Kat Young, who provided motivation, support, discussion, and numerous much needed coffee and cake breaks. I must also extend thanks to the rest of the AudioLab, who have provided motivation, support, and audio related discussions over the past few years.

I must thank my two good friends, Matt Darling and Reece Loake, who have always been there throughout this journey. Your support, conversation, and our numerous gaming sessions have provided much needed diversions and relaxation these past four years.

This work could never have been achieved without the unwavering support and love from my family. I must extend special thanks to both my parents, Andrew Lovedee-Turner and Christine Lovedee-Turner, for their continuous support, and for allowing me to experience living in Hong Kong, whilst writing my thesis. Without their support none of this would have been possible.

Author Deceleration

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 this, or any other, University. All sources are acknowledged as References.

The following publications form the basis of the work presented in this thesis:

Published

Lovedee-Turner, M., & Murphy, D. (2018). "Application of Machine Learning for the Spatial Analysis of Binaural Room Impulse Responses". Applied Sciences (Switzerland), 8(1). Available: http://doi.org/10.3390/app8010105 [Accessed: March 28, 2018]

Lovedee-Turner, M., & Murphy, D. (2019) "3D Reflector Localisation and Room Geometry Estimation using Spherical Microphone Arrays". Journal of the Acoustical Society of America. Available: https://doi.org/10.1121/1.5130569 [Accessed: Feb. 17, 2018]

'... the geometry of the place was all wrong. One could not be sure that the sea and the ground were horizontal, hence the relative position of everything else seemed phantasmally variable.'

The Call of Cthulhu by H. P. Lovecraft

Part I

Introduction

Chapter 1

Introduction

As sound propagates through an enclosed space, inevitably it will be incident upon, and interact with, any reflective boundaries or surfaces present. These interactions result in some proportion of the energy contained within the propagating sound being reflected back into the space. This process repeats until all the acoustic energy of the signal has been lost, and the room returns to its original steady state. These reflections are characteristic of a given room, and convey information to the listener about size, shape, and ultimately the sound of the environment. In acoustics it is common to measure this characteristic response of the room using an impulse-like broadband excitation signal, producing what is referred to as the *room impulse response*.

The room impulse response typically refers to a monophonic recording of the acoustic response of a room when using an omnidirectional source and receiver, and it is a superposition of the direct source-to-receiver sound component, early reflections produced through limited interactions with the most significant boundaries or surfaces in the space, and a densely-distributed and exponentially decaying reverberant field. The resulting measured room impulse response is defined by the location, shape and acoustic properties of the reflective boundaries or surfaces, together with the source and receiver positions. Therefore, the measured room impulse response for a given room not only conveys information about the acoustic properties of the given room, but also, the location of any boundaries in the room, as the time-of-arrival of any reflections is directly related to the propagation distance of the sound. Therefore, from one or more Room Impulse Responses (RIRs), attempts can be made to infer the size and position of reflective boundaries and consequently the shape of a given room, and this process is referred to as *geometry inference*..

Geometry inference has potential applications in various aspects of acoustics and signal processing research, where normally *a priori* knowledge of a room's boundary locations would be required, which is not possible when implemented within consumer technology. In acoustics consultancy, geometry inference can be used as a means of deriving key reflection in a given environment, providing data that can be used when acoustically treating the room. The geometric model of a room can be used to simulate the acoustic conditions, and consequently the Spatial Room Impulse Responses (SRIRs), for different source and receiver positions within the environment. Geometry inference in this context can be used to generate a room model, which subsequently can be used to generate additional SRIRs throughout the environment. This has potential applications in interactive media such as video games where SRIRs can be used to produce a more realistic rendering of an acoustic scene, producing an immersive experience for the player. In smart home-devices, knowledge of the surrounding environment, and therefore geometry inference, can be used as a means of enhancing speech recognition through source separation and dereverberation as seen in [1–3]. Furthermore, geometry inference can be applied to robotics as a means of providing real-time information about a robot's surrounding environment and its current and previous position [4]. Finally, in the context of virtual and augmented reality, geometry inference can be used to track a user's position within an environment or produce more robust methods for spatial audio rendering by evaluating a user's loudspeaker setup and listening environment, which subsequently can be accounted for when rendering a virtual auditory environment [5], so producing an ideally more immersive user experience. From these applications, it is evident that removing *a priori* knowledge of an environment, it is of paramount importance to arrive at a method for geometry inference that is universally robust to rooms of different shape, size, complexity, and measurement conditions.

At present, geometry inference methods are only accurate for simple convex rooms where all interior angles are less than 180°, such as cuboid-shaped rooms, which limits the application of these methods. Rooms come in different shapes, sizes, and levels of complexity, and as such, geometry inference methods need to be accurate at estimating the geometry for these different room conditions. The focus of this thesis is to develop and establish a geometry inference method applicable to both convex- and non-convex-shaped rooms. Contrary to existing work, which generally considers the problem of boundary localisation for cuboid-shaped rooms, the proposed research will present an end-to-end method for boundary localisation, boundary validation, and room shape estimation, that is tested using different convex- and non-convex-shaped rooms. This will be achieved through the use of a spatiotemporal decomposition-based reflection detection method for compact microphone arrays; an image-source based boundary detection

method; a geometry validation process, and a room shape inference method.

The research presented in this thesis will expand on existing work into direction-of-arrival estimation for reflections in binaural room impulse responses, reflection detection, and geometry inference.

1.1 Hypothesis

The hypothesis that informs and guides the work in this thesis is as follows:

Given a compact microphone array and a sufficient number of spatial room impulse responses to ensure a first-order reflection is detectable for each boundary, accurate boundary estimation, and consequently, room shape estimation, can be achieved for both convex- and non-convex-shaped rooms.

1.1.1 Description of Hypothesis

Geometry Inference

Geometry inference is the problem of determining the location of reflective boundaries within a given enclosed space based on reflections detected within a room impulse response. There are two aspects to this area of research, determining the location and positioning of the reflective boundaries, which is the focus of the majority of prior work, and the subsequent inference of the shape of the room from these identified boundaries. The key challenges with geometry inference are to develop a robust end-to-end system which does not require strict assumptions about a room's shape or the number of reflections, can attempt inference of reflection paths for higher-order reflections, and can reduce the impact of false-positive detections or incorrectly inferred reflection paths. In the context of this thesis, a successful result is defined as comparable accuracy to that previously achieved for cuboid-shaped rooms in the literature, that is an average difference in the position of the estimated boundaries with respect to the ground-truth position of the boundaries of between 1.7 cm – 24.5 cm [6–9].

Compact Microphone Arrays

A compact microphone in the context of this thesis is used to refer to a microphone array with a diameter less than or equal to 18 cm, that can be used to estimate the direction-of-arrival of auditory events within the surrounding sound field. The work presented in this thesis will use the KE-

MAR 45BC [10] and KU100 [11] binaural dummy head microphones, and the EigenMike EM32 32-channel spherical microphone array, as previously used for direction-of-arrival analysis of reflections in [12]. For a microphone arrays to be considered viable for geometry inference, an ideally low error in direction-of-arrival estimation is desirable, as any error in direction-of-arrival results in an equivalent angular error in boundary position relative to the desired boundary.

First-order reflections

A first-order reflection is one that is produced by a propagating sound wave interacting with a single reflective boundary. In the context of geometry inference, first-order reflections are used to determine the location of a reflective boundary based on the reflection's time-of-arrival or time- and direction-of-arrival. Therefore, it is imperative that each boundary within a given enclosed space has a first-order reflection detectable in at least one spatial room impulse response. This constraint defines the number of measurement positions needed to infer the shape of the room, with more measurement positions needed for more complex, non-convex-shaped rooms.

1.2 Novel Contributions

The research presented in this thesis has resulted in the following novel contributions to the field:

- The application of a binaural model fronted neural network for direction-of-arrival estimation of reflections in binaural room impulse responses, as opposed to a continuous speech signal, considering a cascade-forward neural network topology.

- A method for spatiotemporal decomposition based reflection detection and analysis using spherical microphone arrays, capable of detecting simultaneously arriving reflections at the microphone array as discrete events.

- A method for geometry inference, room shape estimation, and boundary validation , applicable to both convex- and non-convex-shaped rooms.

- Validation of the proposed methods considering an objective analysis of results from rooms of different shape, size, and complexity, expanding the range of current state-of-the-art geometry inference test scenarios.

1.3 Thesis Layout

Chapter 2 introduces the fundamental acoustic and audio signal processing principles upon which the rest of this thesis is founded. This includes acoustic propagation, acoustic reflection, the room impulse response, the image-source method, binaural signals, neural networks, acoustic beamforming, and spherical harmonics.

Chapter 3 introduces the concept of time- and direction-of-arrival estimation for reflections present in (spatial) room impulse responses, and discusses the current state-of-the-art methodology. These methods are often a prerequisite to geometry inference, as the information extracted for each reflection is directly relatable to the boundaries present in a given measurement environment. This chapter also discusses the drawbacks of these methods that this thesis aims to address.

Chapter 4 introduces geometry inference, and discusses prior work in geometry inference methodology. Key limitations are discussed, particularly the assumption of convexity made by these methods, which is one of the main contribution of this thesis.

Chapter 5 discusses the development of a method for estimating the direction-of-arrival for reflections in binaural room impulse responses. This method is tested using both the direct sound and reflections measured using two different binaural dummy head microphones and two different loudspeakers, testing the generalisability of the method to different measurement setups. This establishes whether a compact microphone array consisting of two-channels, that is capable of capturing three-dimensional spatial information, can be used to produce sufficiently accurate estimates of direction-of-arrival for use in geometry inference.

Chapter 6 presents a spatiotemporal decomposition based reflection detection method for use with spherical microphone arrays. The method is validated using four sets of measurements, two simulated and two real-world. To assess the accuracy of the method, an implementation of the image-source method is used to compute the expected time-of-arrival for candidate reflections within the spatial room impulse response. The accuracy of the proposed method is then compared to implementations of two state-of-the-art reflection detection methods, based on circular-variance local maxima [14] and dynamic time warping matching pursuit [15].

Chapter 7 introduces the novel image-source reversion, room shape estimation, and boundary validation method. This chapter presents objective analysis of the performance across three

scenarios. Scenario One presents test cases for seven sets of CATT-Acoustic [16] simulated spatial room impulse response for six rooms, of different size, shape, and complexity, including two non-convex cases. Scenario Two tests the variability of the proposed method across 33 sets of measurement position combinations for two different non-convex L-shaped rooms, again simulated using CATT-Acoustic. Finally, Scenario Three tests the robustness of the method to real-world conditions, using measurements obtained from a cuboid-shaped room.

Chapter 8 summarises the results of the thesis, reconsiders the hypothesis that has been stated in this chapter, and looks to future research based on the results presented in this thesis, and some of the further questions that have emerged as a consequence.

Part II

Literature Review

Chapter 2

Conceptual Foundations

2.1 Introduction

In this chapter the fundamental concepts that underpin the work discussed later in this thesis will be introduced. Starting by defining the properties of sound propagation, interactions with reflective boundaries, and the room impulse response in Section 2.2. In Section 2.3 the image-source model as proposed by Allen and Berkley [17] is defined, which is the conceptual basis of the geometry inference method proposed later in this thesis. Section 2.4 outlines the basics of binaural audio Finally, Section 2.5 introduces spatial room impulse responses in the context of spherical microphone arrays, and defines spherical harmonics, which forms the basis by which spatiotemporal decomposition of spatial room impulse responses is performed.

2.2 Fundamentals of Acoustics

2.2.1 Sound Propagation

Sound waves are represented as the displacement of particles within a medium from their mean position [18], and these fluctuations in pressure, when transmitted through and amplified by the human auditory auditory system, represent what is referred to as sound. Sound waves propagate outwards from a point-of-origin by locally displacing molecules present within a medium (solid, liquid, or gas), producing points of compression (high pressure) and rarefaction (low pressure) [18–21]. The distances between points of compression and rarefaction, the wavelength λ of the sound, define the frequency f of the sound, with shorter wavelengths defining higher frequencies and longer wavelengths resulting in low-frequencies as [20],

$$f = \frac{c}{\lambda} \tag{2.1}$$

where c is the speed of sound. The amplitude of the sound wave at a given point in space (x, y, z) is directly proportional to the change in pressure $p(x, y, z)$ from the mediums resting pressure $\rho_0(x, y, z)$ as [19, p. 9],

$$p(x, y, z) = \rho(x, y, z) - \rho_0(x, y, z) \tag{2.2}$$

where $\rho(x, y, z)$ is the instantaneous pressure at a given point in space. Given the acoustic pressure, the amplitude of the sound wave can be quantified as sound pressure level (SPL) in decibels (dB), a logarithmic scale using the ratio between the acoustic pressure p and the threshold of hearing p_{ref} (20 μPa at 1 kHz [19]), as,

$$\text{SPL} = 20\log_{10}\frac{p}{p_{ref}}(dB) \tag{2.3}$$

The propagation of a sound wave through a medium can be described using a differential equation, relating the time (t) and spatially varying pressure p to the speed of sound in the medium c, referred to as the wave equation, which for one-dimension is expressed as [19, p. 10-13],

$$\frac{\delta^2 p}{\delta t^2} = c^2 \frac{\delta^2 p}{\delta x^2} \tag{2.4}$$

The general solution to this differential equation, as proposed by d'Alambert, can be expressed using two twice-differentiable arbitrary functions defining the right p_r and left p_l going components of the wave in the x direction with velocity c as [18],

$$p(x, t) = p_r(ct - x) + p_l(ct + x) \tag{2.5}$$

Furthermore, the time-dependant displacement of the particles within the medium, particle ve-

locity v, can be defined from (2.5) as [18],

$$v(x,t) = \frac{1}{Z_0}[p_r(ct-x) + p_l(ct+x)] \tag{2.6}$$

where Z_0 is the acoustic impedance of the medium, which quantifies the medium's resistance to the flow of acoustic energy [18]. Given the periodic nature of sound waves, it is common to adopt a complex harmonic function representing $p_r(ct-x)$, with $p_l(ct+x)$ set to zero [18, 19], as

$$p(x,t) = \widehat{p}e^{i(\omega t - kx)} \tag{2.7}$$

where $\widehat{p}$ is the pressure amplitude, k is the wave number, ω is the angular frequency, $i = \sqrt{-1}$, and $e^{(.)}$ denotes the exponential (full derivation of these equations available in [18, 19]).

As a sound propagates in the free-field the acoustic pressure will be attenuated based on the inverse-square law. The inverse-square law states that, within a free-field, as sound propagates outwards from the point-of-origin, the intensity of the sound will be attenuated by the square of the distance [20] .

Furthermore, the speed at which sound travels, c, within a medium is not constant, and varies with respect to both temperature and humidity calculated as,

$$c = \sqrt{\kappa \frac{p_{air}}{\xi}} + 0.6 * T \tag{2.8}$$

where p_{air} is the air pressure, ξ is the air density which will vary with humidity as a result of the increasing/decreasing number of water molecules present in the air, κ is the adiabatic exponent (the ratio of heat capacity at constant pressure to heat capacity at constant volume) which for air is 1.4, and T is the temperature in centigrade [18]. The speed of sound at $20°$c and 0% humidity is 343.36 m/s, increasing the humidity to 50% yields a speed of sound of 343.99 m/s.

2.2.2 Acoustic Reflection

If sound propagates through an enclosed space rather than the free-field it will be incident upon and interact with a boundary. Upon this interaction part of the sound wave's energy will be absorbed into the boundary, and either converted to heat or transmitted through the boundary, while the rest is reflected back into the space [18] as seen in Figure 2.1. The amount of energy absorbed upon incidence with a boundary is dependant upon the material the boundary is made of, and is defined as a boundary's absorption coefficient α. These acoustic properties are frequency dependant, and hence for different materials the quantity of the sound wave's energy being absorbed will vary with frequency.

Assuming a surface is perfectly rigid the sound energy will be reflected specularly according to Snell's law, where the angle of incidence θ_i relative to the normal of the plane or surface equals the angle of reflection θ_r relative to this plane's normal as seen in Figure 2.2 [20]. The magnitude and phase changes introduced as a result of interactions with a boundary can therefore be expressed, in two-dimensions, using the reflection factor for a boundary R_f by expressing (2.6) and (2.7), from [19], as,

$$p_i(x,y,t) = \widehat{p}e^{i(\omega t - kx\,\cos(\theta_i) - ky\,\sin(\theta_i))} \tag{2.9}$$

$$v_i(x,y,t) = \frac{\widehat{p}}{Z_0}e^{i(\omega t - kx\,\cos(\theta_i) - ky\,\sin(\theta_i))} \tag{2.10}$$

$$p_r(x,y,t) = \widehat{p}R_f e^{i(\omega t - kx\,\cos(\theta_r) - ky\,\sin(\theta_r))} \tag{2.11}$$

$$v_r(x,y,t) = \frac{-R_f\widehat{p}}{Z_0}e^{i(\omega t - kx\,\cos(\theta_r) - ky\,\sin(\theta_r))} \tag{2.12}$$

$$\tag{2.13}$$

where $p_i(x,y,t)$ and $v_i(x,y,t)$ are the incident pressure and particle velocity respectively, $p_r(x,y,t)$ and $v_r(x,y,t)$ are the reflected pressure and particle velocity respectively, and the reflection factor R_f is related to the acoustic impedance of the wall Z as [19],

$$R_f = \frac{Z\cos(\theta_r) - Z_0}{Z\cos(\theta_r) + Z_0} \tag{2.14}$$

Therefore, the reflection factor for the boundary is directly linked to the absorption coefficient for the boundary such that [18, 19],

$$\alpha = 1 - |R_f|^2 \tag{2.15}$$

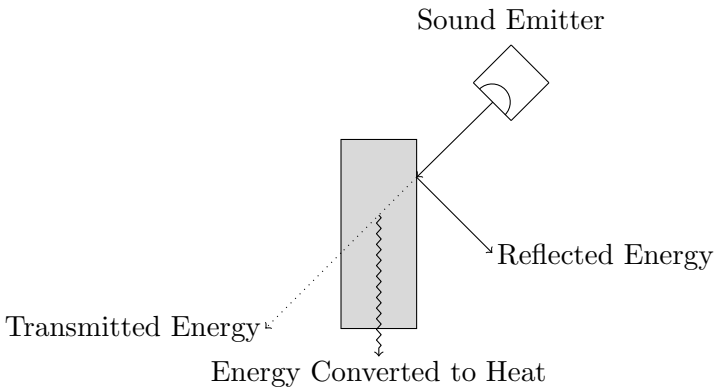

Figure 2.1: Collision of propagating sound wave with room boundary, showing reflected sound energy and absorbed sound energy.

The boundary material and shape will also have an impact on the direction in which the reflected energy travels. If the boundary surface is parabolic in shape, the reflected sound will be focused to a point dependant upon the angle of incidence. Similarly to parabolic surfaces, concave surfaces focus the reflected sound to a point, however, the precision of this point varies depending on the shape of the surface [20]. A convex shaped boundary will cause the reflected energy from the sound wavefront to be scattered across numerous different directions [20]. An example of reflections of these types of surfaces can be seen in Figure 2.3.

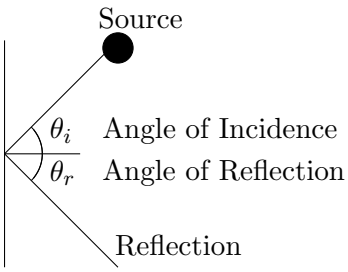

Figure 2.2: Specular reflection from a flat surface, where angle of incidence is equal to the angle of reflection

While specular reflections are key to the work detailed in this thesis, other reflection types exist. Reflection scattering occurs in the case of a boundary with a non-smooth surface and results in the reflected sound energy being scattered outwards from the boundary across all directions relative to the point of incidence and the boundary. Scattering is also frequency-dependent, and the

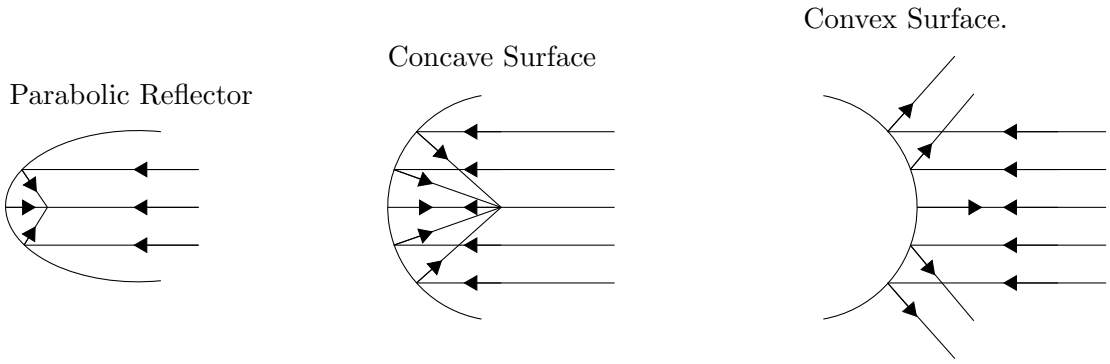

Figure 2.3: Example reflection patterns for parabolic, concave, and convex surfaces

quantity of reflected energy that is scattered for each frequency band is defined as the scattering coefficient for the surface in question [20]. Diffraction occurs when a propagating sound wave passes around the edge of a boundary of finite length, resulting in the propagation path bending around the object [20] as seen in Figure 2.4. As with absorption and scattering, diffraction is frequency-dependent, and relative to the size of the diffracting object [20]. As obstacle size increases, so does the wavelength at which diffraction occurs, and therefore diffraction is more commonly observed at lower frequencies [20].

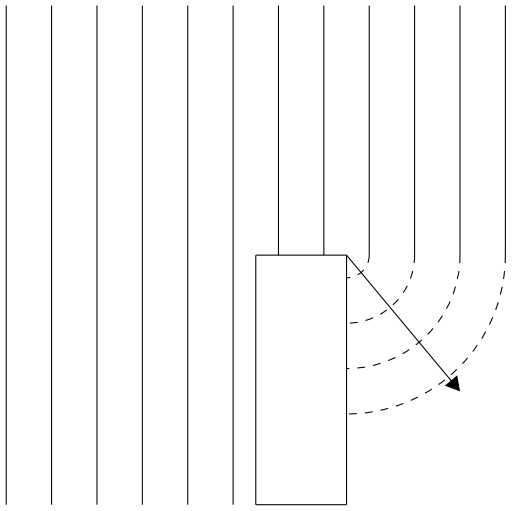

Figure 2.4: Example of sound diffracting around the corner of an object, with the sound radiating into the area cast by the object referred to as the 'shadow zone'.

By considering the properties of sound propagation and reflection, it is evident that from a listeners perspective, the arrival time and amplitude of direct source-to-listener and reflected sound is linked to: the relative positioning of the source and listener; the presence and acoustic properties of any boundaries and objects; and the temperature and humidity of the air within space. A common method of representing these combined acoustic properties for an environment is through the use of a RIR.

2.2.3 The Room Impulse Response

A RIR is the measured or simulated characteristic response of any enclosed space to an input excitation from a known impulse-like broadband test signal. It is a superposition of the direct source-to-receiver sound component, early reflections produced through limited interactions with the most significant boundaries or surfaces in the space, and a densely-distributed and exponentially decaying reverberant field, as shown in Figure 2.5. A RIR is therefore defined by the location, shape and acoustic properties of the reflective boundaries or surfaces, together with the source and receiver positions. A frequency independent representation of RIR $\mathbf{h}(t)$ can be mathematically defined in discrete-time as a superposition of sinc functions($sinc()$) with peaks located at the time-of-arrival (ToA) τ, individual signal amplitude defined as a, and the addition of a residual time-variant ambient noise component $r(t)$

$$\mathbf{h}(t) = \sum_{i=1}^{\infty} a_i \text{sinc}(t - \tau_i) + r(t) \tag{2.16}$$

where i refers to the i_{th} arrival at the receiver.

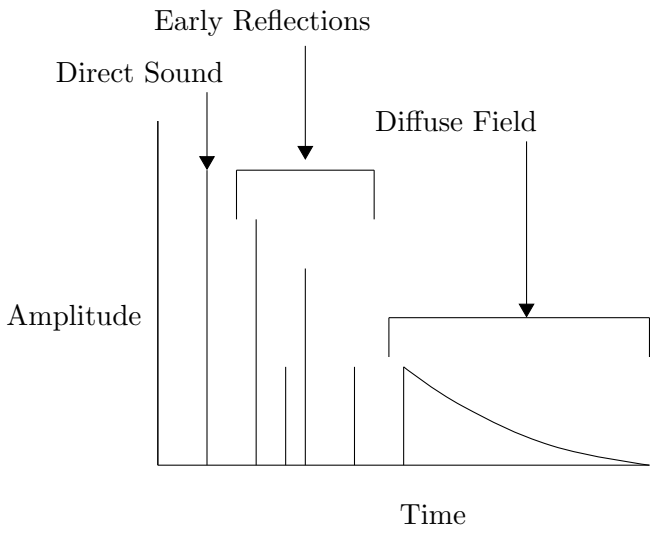

Figure 2.5: Generalised depiction of an impulse response, split into direct sound, early reflections and diffuse field.

The first sound to arrive at the receiver upon excitation of the space is the direct sound, having the shortest propagation distance from the source [21]. The direct sound, in the majority of cases, is the loudest signal arriving at the receiver, with the direct signals amplitude and the direct-to-reverberant sound energy providing auditory cues pertaining to the distance between the sound source and the receiver.

The early reflections are discrete reflections arriving at the receiver from different directions and points in time. As the source and/or receiver locations change in relation to the environment, the time and amplitude of the early reflections will change as a result of differences in the propagation path between reflective boundary and receiver [21]. The time and amplitude of arrival of these early reflections provides the listener with information about the size and shape of the environment [21]. When referring to reflections present within a RIR it is common to refer to them based on the number of interactions they have had with different boundaries or surfaces - *i.e.* their *reflection order*. For example, if a reflection has interacted with two boundaries when it arrives at the receiver, it is referred to as a *second-order* reflection.

The diffuse field, sometimes referred to as the reverberant sound or late reverberation, is made up of densely-distributed reflections from multiple combinations of boundaries and surfaces arriving from multiple directions [21]. Therefore, the diffuse field does not contain arrivals that are individually identifiable as discrete acoustic events. This leads to the generalisation that the diffuse field for a given enclosed space does not vary significantly as either source and/or receiver change position [21].

2.3 Image-Source Method

While numerous geometric acoustic modelling techniques exist, for geometry inference it is common to adopt an approach based on an inverse-model of the image-source method, as a result of the direct relationship that can be drawn between the ToA of a reflection, the location of an image-source, and the location of a reflective boundary. Therefore, in this section the image-source method, as originally proposed by Allen and Berkley in [17] will be introduced.

The image-source model is a geometric acoustic modelling technique, where the solution is a RIR derived as a summation of specular reflections produced by rigid walls. As introduced in Section 2.2.2 a specular reflection is defined such that the angle of reflection relative to the normal of the plane in question is equal to the angle of incidence. Therefore, the reflective conditions of a rigid boundary is equivalent to defining a secondary source, an image-source $\widetilde{\mathbf{s}}$, that is produced by mirroring the source $\mathbf{s}$ (or other image-sources for higher-order reflections) perpendicularly across the reflective boundary (see Figure 2.6). This can be summarised as [6],

$$\widetilde{\mathbf{s}} = \mathbf{s} + 2 < \mathbf{b} - s, \mathbf{n} > \mathbf{n} \tag{2.17}$$

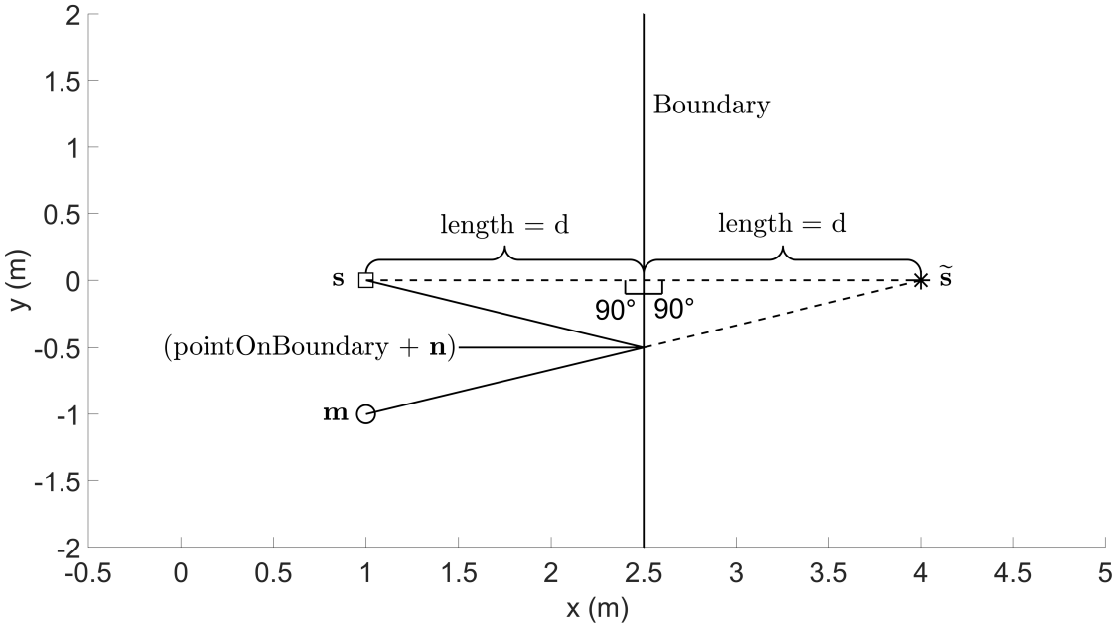

Figure 2.6: Example of an image-source produced by mirroring the source perpendicularly across the boundary. As can be seen the reflection path produced is specular with the angle of reflection relative to the normal of the plane equal to the angle of incidence.

where $\mathbf{b}$ is the $[x, y, z]$ coordinates that define a point on the boundary, $\mathbf{n}$ is the unit normal for the boundary, and $< ., . >$ denotes the dot product. These image-sources $\widetilde{\mathbf{s}}$ are computed for the source $\mathbf{s}$, all previously computed image-sources, and all L boundaries up to a reflection order N, producing L^N image-sources. An example of the image-sources produced for a cuboid-shaped room, with a reflection order of $N = 2$ can be seen in Figure 2.7.

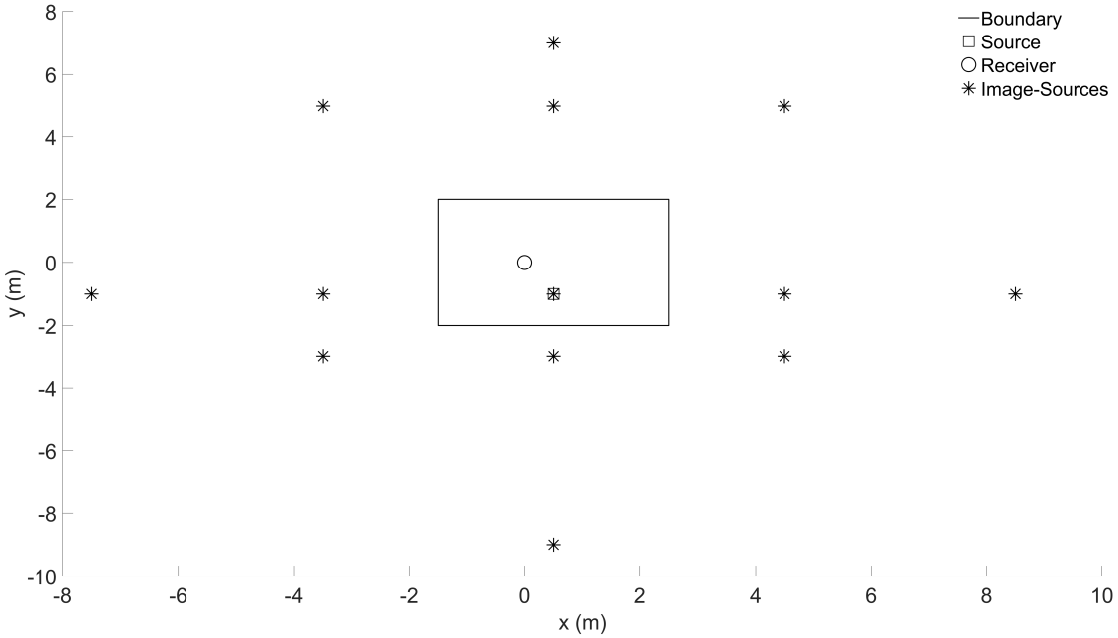

Figure 2.7: Example image-sources computed for a cuboid shaped room. Asterisks denote image-sources, the square denotes the source location, and the circle denotes the receiver location.

The RIR $\mathbf{h}(t)$ representing the arrival of signals up-to a reflection order N, for a convex room with L boundaries, can be, from [17], expressed as,

$$\mathbf{h}(t) = \frac{\text{sinc}\left(t - \frac{||\mathbf{s}-\mathbf{m}||}{c}\right)}{4\pi||\mathbf{s}-\mathbf{m}||} + \sum_{i=1}^{i=L^N} \frac{\text{sinc}\left(t - \frac{||\widetilde{\mathbf{s}}_i-\mathbf{m}||}{c}\right)}{4\pi||\widetilde{\mathbf{s}}_i-\mathbf{m}||} \tag{2.18}$$

where $\mathbf{m}$ is the microphone location and $4\pi||\widehat{\mathbf{s}}_i - \mathbf{m}||$ defines energy loss as defined by the inverse-square law. An example RIR produced from the scenario presented in Figure 2.7 can be seen in Figure 2.8.

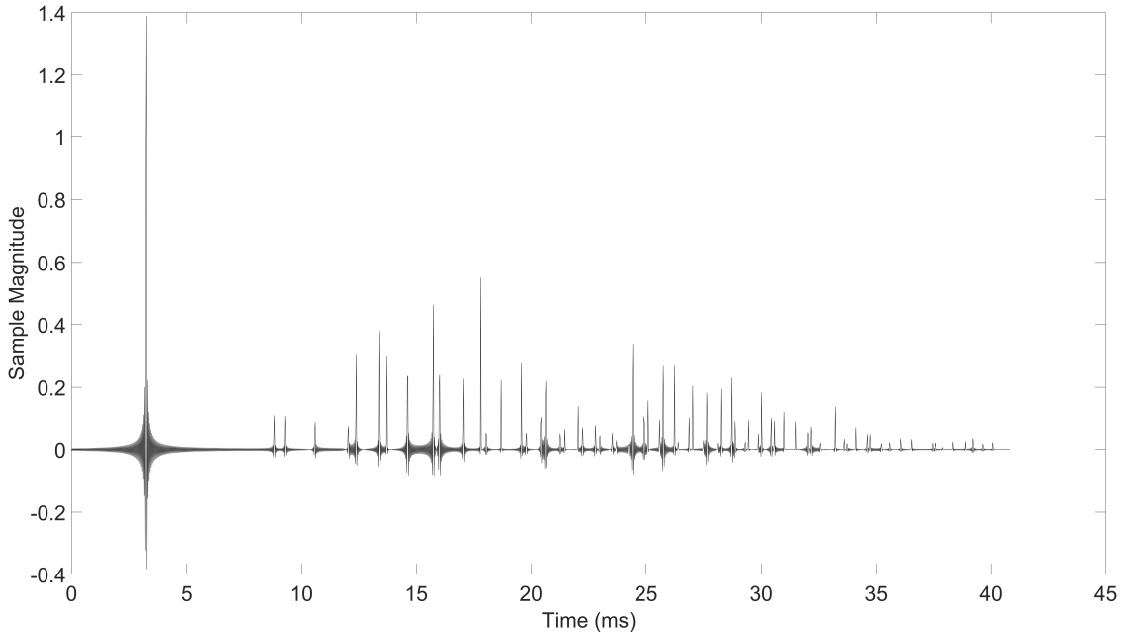

Figure 2.8: Room impulse response computed from the image-sources in Figure 2.7 using (2.18).

When considering the relationship between the image-source and its previous-source (either the source or another image-source) (as seen in Figure 2.6), the distance from previous-source-to-boundary and boundary-to-image-source are equal, and the line between the previous-source and image-source is parallel to the boundary's normal. Therefore, the normal of the boundary, $\widetilde{\mathbf{n}}$, and a point on the boundary, $\widetilde{\mathbf{b}}$, can be inferred from these two points, from [6], as,

$$\widetilde{\mathbf{b}} = \frac{\widetilde{\mathbf{s}} + \mathbf{s}}{2} \tag{2.19}$$

$$\widetilde{\mathbf{n}} = \frac{\widetilde{\mathbf{s}} - \mathbf{s}}{||\widetilde{\mathbf{s}} - \mathbf{s}||} \tag{2.20}$$

This relationship is what makes the image-source model appealing for geometry inference, as the location of the image-sources can be extracted from a set of candidate ToA estimates for reflections in a RIR measurements. Therefore, the aim of image-source-based geometry inference is to find the most likely previous-source in the reflection path, either from the candidate image-sources extracted from the RIR or the source position.

2.4 Binaural Room Impulse Responses

A Binaural Room Impulse Response (BRIR) is a RIR measured with a receiver that is characterised by having the properties of a typical human head, that is two channels of information separated appropriately, and subject to spatially-dependant spectral and temporal variations imparted by the pinnae and head. The spatial information contained within a BRIR is encoded as time-of-arrival and level differences between the signals arriving at each ear, referred to as Interaural Time Difference (ITD) and Interaural Level Difference (ILD) respectively [22]. Both of these cues are a function of frequency and source position relative to the head, and therefore provide the listener with cues that pertain to a sound source's location [22]. The ITD is due to differences in propagation paths from the source to each ear, and diffraction of the propagating sound wave around the head [21]. The ILD is due to the shadowing of the head, which results in the sound arriving at the far ear at a lower amplitude. Furthermore, there will be spectral differences between the signals arriving at each ear due to sound reflecting off the pinnae, which as a result of the short delay-time for these reflections, produces a comb-filtering effect [21]. These cues will vary between people and dummy heads as a result of differences in ear and head morphology. Therefore, when recording binaural audio it is desirable to have a microphone array which represents an average human head with microphones situated inside the ears, referred to as a *binaural dummy head microphone*, ideally producing a recording that will produce adequate spatialisation for different listeners.

These interaural cues form the basis by which computer models attempt to replicate a human's ability to localise sound. However, it should be noted that in the context of human sound localisation, visual cues are often used to relay additional information about sound source's location to the listener [21], which will not be considered further for the rest of this thesis. These interaural cues can be measured for a given head by measuring the response, in each ear, to an input excitation from a known impulse-like broadband test signal produced at a distance from the head, the resulting signal is referred to as the *Head Related Impulse Response (HRIR)*.

2.5 Spatial Room Impulse Responses - Spherical Microphone Arrays

Spatial Room Impulse Response (SRIR) are RIR that have been measured with a microphone array, or simulated, such that they contain spatial information about the acoustic environment. Considering a spherical microphone array, as used in this thesis, the recorded sound field can be represented in the spherical harmonic domain through the use of spherical harmonics [27–29].Therefore, the time-domain RIR from 2.16 can be expressed as the SRIR $\mathbf{H}(t)$ by using the real-valued spherical harmonic vector $\mathbf{y}(\theta, \phi)$ to steer the sinc function towards a given azimuth and elevation DoA [27–29],

$$\mathbf{H}(t) = \sum_{i=1}^{\infty} \mathbf{y}(\theta_i, \phi_i) a_i \text{sinc}(t - \tau_i) + \mathbf{R}(t) \tag{2.21}$$

where $\mathbf{R}(t)$ is the time-variant residual noise component, $\mathbf{y}(\theta, \phi)$ is defined as,

$$\mathbf{y}(\Psi) = [Y_0^0(\theta, \phi), Y_1^{-1}(\theta, \phi), Y_1^0(\theta, \phi), Y_1^1(\theta, \phi), ..., Y_N^M(\theta, \phi)]^T \tag{2.22}$$

where the real-valued spherical harmonics of order n and degree m are from [30, 31] expressed as,

$$Y_n^m = \begin{cases} \sqrt{\frac{2n+1}{4\pi} \frac{(n-m)!}{(n+m)!}} P_n^m(cos(\phi)) \sqrt{2} cos(m\theta), & \text{if } m > 0 \\ \sqrt{\frac{2n+1}{4\pi} \frac{(n-m)!}{(n+m)!}} P_n^m(cos(\phi)), & \text{if } m = 0 \\ \sqrt{\frac{2n+1}{4\pi} \frac{(n-m)!}{(n+m)!}} P_n^m(cos(\phi)) \sqrt{2} sin(m\theta) & \text{if } m < 0 \end{cases} \tag{2.23}$$

where P_n^m is the associated Legendre polynomial of order n and degree m. The spherical harmonics can also be expressed as a complex-valued function for frequency domain processing, from [26, 30], as,

$$Y_n^m = \sqrt{\frac{2n+1}{4\pi} \frac{(n-m)!}{(n+m)!}} P_n^m(\cos(\phi)) e^{im\theta} \tag{2.24}$$

Spherical harmonics will be further used in Chapter 6 to define a spherical harmonic domain spatiotemporal decomposition method for detecting reflections in SRIR measured with a spherical microphone array.

2.6 Summary

In this chapter the fundamental concepts that underpin work presented in this thesis have been presented, including, sound propagation, acoustic reflection, the image-source method, room impulse responses, binaural room impulse responses, spatial room impulse responses, and spherical harmonics. As sound waves propagate through an enclosed space inevitably it will incidence upon and interact with a boundary or surface, at which point the sound will either be specularly reflected, scattered, or, in the case of incidence upon the corner of a boundary of finite-length, diffracted. While scattering and diffraction are important in defining the acoustics of a room, they are hard, if not impossible, to use as a means of geometry inference, as the estimation of possible reflection paths is not possible without *a priori* knowledge of reflection order and room shape to calculate possible points of reflections. Therefore, in geometry inference methods it is useful to ignore these acoustic properties, and assume that all detectable reflections contain a dominant specular component. The Room Impulse Response is the measured or simulated characteristic response of any enclosed space to an input excitation from a known impulse-like broadband test signal and is representative of the reflective boundaries of the room, and the source and receiver location. This principle defines the underlying mechanics by which numerous acoustic signal processing techniques are developed, in particular geometry inference. To this extent, one of the common approaches to geometry inference considers the use of an inverse image-source model, using the relationship between a reflection's time-of-arrival, the consequent location of an image-source, and the reflective boundary's location. This relationship allows boundary locations to be estimated without requiring *a priori* knowledge of any boundary parameters.

Chapter 3

Reflection Detection and Direction-of-Arrival Estimation: Relevant Previous Work

3.1 Introduction

In the previous chapter the fundamentals of sound propagation, acoustic reflection, the image-source model, neural networks, beamformers, and spherical harmonics were presented, which underpins the future work presented in this thesis. The Room Impulse Response (RIR) represents the acoustics of a room for a given source and receiver location, and therefore, contains a superposition of the direct source-to-receiver sound and reflections, produced by interactions with the boundaries and surfaces present in the space, arriving at the receiver. It can be desirable in some cases to be able to identify the location of specific reflections within the RIR, and when using a microphone array, the incident direction-of-arrival (DoA) for each reflection.

A microphone array samples the sound field, or in this case RIRs, at different points in space. The resulting multi-channel signals will, therefore, contain spatial information about the arriving sounds, and so a RIR captured in this way is commonly referred to as a Spatial Room Impulse Response (SRIR). Assuming such an array is capable of representing the arrival of sound across three-dimensions, these SRIRs can be expressed in a general form through the use of the steering vector $\mathbf{\Delta}(\theta, \phi)$, which defines the directionally-dependant temporal and level differences between each channel, for azimuth θ and elevation ϕ DoA. Adapting (2.16) the SRIR can be

expressed as,

$$\mathbf{H}(t) = \sum_{i=1}^{I} \mathbf{\Delta}(\theta_i, \phi_i)\text{sinc}(t - \tau_i)\alpha_i + \mathbf{R}(t) \tag{3.1}$$

where $\mathbf{R}(t)$ is spatially-white and time-varying residual noise component. The array response matrix $\mathbf{\Delta}(\theta, \phi)$ will vary for different microphone arrays. For example, in the case of a BRIR measured with a binaural system, the steering vector is representative of the spectral and temporal differences between signals arriving at each ear as a result of differences in propagation path, pinnae shape, and acoustic occlusion as a result of the head.

This chapter will outline methods presented in the literature for DoA estimation (Section 3.2) and reflection detection (Section 3.3), which are either applicable to the microphone arrays used in this thesis, or are used by relevant geometry inference methods. As some of the reflection detection methods use DoA to detect candidate reflections, DoA estimation will be discussed first.

3.2 Direction-of-Arrival Estimation

DoA refers to the direction from which a propagating sound wave arrives at a microphone array. As such, DoA estimators attempts to determine the direction from which a sound arrived at a microphone array, based on differences between the signals arriving at each channel. The means by which this DoA estimation is performed will vary for different microphone array geometries, as a consequence of how the spatial information for the sound field is sampled. In this section DoA estimation methods relevant to the arrays used in this thesis - spherical microphone arrays and binaural dummy heads - will be explored.

3.2.1 Spherical Microphone Arrays

3.2.1.1 Intensity Vector Analysis

Intensity vector analysis is a DoA estimator which represents the magnitude and direction of acoustic energy using intensity vectors. The intensity vectors $\mathbf{I}$ is computed from the sound pressure p and particle velocity vector $\mathbf{v} = \begin{bmatrix} v_x & v_y & v_z \end{bmatrix}$, where v_x, v_y, and v_z is the particle velocity, with dipole directivity, in the x,y, and z direction respectively, from [26], as,

$$\mathbf{I} = \frac{1}{2}\Re(p^*\mathbf{v}) \tag{3.2}$$

In practice particle velocity is difficult to measure [26], and specialist equipment such as the Microflown is required [33].

An implementation of intensity vector analysis as DoA estimator was applied to first-order ambisonic signals (B-Format microphone) in [34, 35]. This method derives the instantaneous intensity values of the zero- and first-order spherical harmonic domain signals, commonly referred to as the *W* (omnidirectional - Y_0^0 spherical harmonic); *X* (x-axis - Y_1^1 spherical harmonic); *Y* (y-axis - Y_1^{-1} spherical harmonic); and *Z* (z-axis - Y_1^0 spherical harmonic) channels [35], to estimate DoA. These intensity vectors are calculated from the Short Time Fourier Transform (STFT)[1] of the channels and calculated for each frequency bin and time-frame as,

$$\mathbf{I}_X(\omega, t_f) = \frac{\sqrt{2}}{Z_0}\Re(\mathbf{W}^*(\omega, t_f)\mathbf{X}(\omega, t_f)) \tag{3.3a}$$

$$\mathbf{I}_Y(\omega, t_f) = \frac{\sqrt{2}}{Z_0}\Re(\mathbf{W}^*(\omega, t_f)\mathbf{Y}(\omega, t_f)) \tag{3.3b}$$

$$\mathbf{I}_Z(\omega, t_f) = \frac{\sqrt{2}}{Z_0}\Re(\mathbf{W}^*(\omega)\mathbf{Z}(\omega, t_f)) \tag{3.3c}$$

where $\mathbf{I}_X(\omega, t_f)$, $\mathbf{I}_Y(\omega, t_f)$, and $\mathbf{I}_Z(\omega, t_f)$ is the instantaneous intensity at angular frequency ω and time-frame t_f for the $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ channels respectively, Z_0 is the acoustic impedance of air, and $\mathbf{W}^*$ is the complex conjugate of the *W* channel at angular frequency ω and time-frame t_f [34–36].

The DoA of time-frame t_f at each angular frequency is then calculated, from [34], as,

$$\theta(\omega, t_f) = \tan^{-1}\left[\frac{-\mathbf{I}_Y(\omega, t_f)}{-\mathbf{I}_X(\omega, t_f)}\right] \tag{3.4.1}$$

$$\phi(\omega, t_f) = \tan^{-1}\left[\frac{-\mathbf{I}_Z(\omega, t_f)}{\sqrt{\mathbf{I}_X^2(\omega, t_f) + \mathbf{I}_Y^2(\omega, t_f)}}\right] \tag{3.4.2}$$

[1]The STFT is the Fast Fourier Transform (FFT) computed over short time-frames, representing the change in frequency and phase content of a signal over time.

where $\mathbf{I}_X$, $\mathbf{I}_Y$ and $\mathbf{I}_Z$ are the instantaneous intensity vectors for the X, Y and Z channels respectively.

Merimaa and Pulkki's used this B-Format implementation of intensity vector analysis in their research into Spatial Impulse Response Rendering (SIRR) [34], which used this directional information to render measured RIRs over arbitrary speaker arrays. Results presented show that the directional quality of the rendered audio over such loudspeaker arrays was improved when using SIRR over microphone arrays such as coincident pairs and ambisonics [34]. This method was also used in Directional Audio Coding (DIRAC) [37] for analysing spatial audio for reproduction over arbitrary loudspeaker arrays, building on the work presented in [34]. Furthermore, this method was used in [14] to estimate the DoA of six reflections, and the estimated DoA were used to retrace the reflection paths using ray-tracing.

Pseudo-intensity vector

Pseudo-intensity vector analysis is conceptually similar to intensity vector analysis [26], and treats the zero- and first-order eigenbeams as being proportional to sound pressure and particle velocity. Prior to computation of the intensity vectors, the raw microphone output is transformed into the spherical Fourier domain using a weighted spherical harmonic transform $\mathbf{g}_{q,n,m}$ [26] as,

$$\mathbf{X}_n^m(\omega) \approx \sum_{q=1}^{M} \mathbf{g}_{q,n,m} \widehat{\mathbf{X}}(\omega, \theta_q, \phi_q, r_q) \tag{3.5}$$

where $\widehat{\mathbf{X}}(\omega, \theta_q, \phi_q, r_q)$ is the Fourier transformed raw microphone signal at angular frequency ω for the microphone at polar coordinates azimuth (θ_q), elevation (ϕ_q), and radius (r_q), n is the order of the spherical harmonics, m is the degree of the spherical harmonics, and M is the total number of microphones in the array. The weighted spherical harmonic transform $\mathbf{g}_{q,n,m}$ is expressed as,

$$\mathbf{g}_{q,n,m} = \frac{4\pi}{M} Y_n^{m*}(\theta_q, \phi_q) \tag{3.6}$$

where Y_n^{m*} is the complex conjugate of the spherical harmonic of order n and degree m, evaluated in the direction of the microphone at azimuth θ and elevation ϕ, and calculated using using the complex spherical harmonic equation from [26] as (2.24).

Using the zero- and first-order spherical Fourier domain signals the pseudo-intensity vector [26] is expressed as,

$$\mathbf{I}(\omega) = \frac{1}{2}\Re\left\{ \left(\frac{\mathbf{x_0^0}(\omega)}{b_0(\omega)} \right)^* \begin{bmatrix} \mathbf{x}_x(\omega) \\ \mathbf{x}_y(\omega) \\ \mathbf{x}_z(\omega) \end{bmatrix} \right\} \tag{3.7}$$

where $\Re(.)$ denotes the real part, $\mathbf{I}(\omega)$ is the intensity vector at angular frequency ω, $\mathbf{x}_x(\omega)$, $\mathbf{x}_y(\omega)$ are dipoles steered in the opposite direction to the x, y, and z axes ((3.8)), and b_0 are the mode coefficients of order $n = 0$ for a rigid sphere, calculated using (3.10) [26].

$$\mathbf{x}_a(\omega) = \frac{1}{b_1(\omega)} \sum_{m=-1}^{1} \alpha_a^m \mathbf{x}_1^m(\omega), a = x, y, z \tag{3.8}$$

where [26],

$$\alpha_x^m = Y_1^m(\pi/2, \pi) \tag{3.9.1}$$

$$\alpha_y^m = Y_1^m(\pi/2, -\pi/2) \tag{3.9.2}$$

$$\alpha_z^m = Y_1^m(\pi, 0) \tag{3.9.3}$$

$$b_n(\omega r, \omega r_a) = 4\pi i^l \left[\mathfrak{J}_n(\omega r_a) - \frac{\mathfrak{J}_n'(\omega r_a)}{\mathfrak{H}_n^{(2)'}(\omega r_a)} \mathfrak{H}_n^{(2)}(\omega r_a) \right] \tag{3.10}$$

where $\mathfrak{J}_n(\omega, r_a)$ is the spherical Bessel function of order n, $\mathfrak{H}_n^{(2)}(\omega, r_a)$ is a second kind spherical Hankel function of order n, $(.)'$ is the first derivative, r_a is the array radius, and k is the wave number for the frequency band [26].

Once the pseudo-intensity vector has been calculated, an average of the intensity vector is taken across(ω) [26], giving:

$$\mathbf{I} = \sum_{\omega} \mathbf{w}(\omega)\mathbf{I}(\omega) \tag{3.11}$$

where $\mathbf{w}(\omega)$ is a weighting function, allowing certain frequencies to be ignored, such as low frequency noise. The DoA can then be estimated as a unit vector pointing in the direction of the sound source as given in [26]:

$$\hat{\mathbf{u}} = -\frac{\mathbf{I}}{||\mathbf{I}||} \tag{3.12}$$

where $||.||$ is the ℓ_2 norm of the intensity vector. While this is similar in principle to variation of intensity vector analysis presented in [34], the x, y, and z axis signal intensity is computed as a weighted average of the spherical harmonic channels of a higher-order spherical harmonic domain signal steered in the opposite direction to the x, y, and z axes with dipole directivity, as opposed to directly using the zero- and first-order spherical harmonic domain signals.

In [26], DoA estimation based on pseudo-intensity vector analysis was tested using an EigenMike em32 [13], to capture a sound source positioned at azimuth 0° and elevation -90°, in a 2.9 $\times$ 2.7$\times$ 3.3 m room with a reverberation time of approximately 300 ms. The results showed that the method was capable of producing accurate results with, in typical environments, a mean error of less than 0.5°. The results presented were, however, only for a single source position. Furthermore, the results showed that the accuracy with which the DoA was estimated decreased as reverb time increased. Results in [38] showed that across static DoA tests with three active sound sources and one with three actively moving in 5° steps with a SNR of 40 dB, an average angular error between 5.71° and 9.28° was achieved, this varied depending on the length of the test signal and whether the source was moving or static [38]. Additional results presented in [39] show that the accuracy of pseudo-intensity vector decreased with respect to reverb time and signal-to-noise ratio (SNR) for the case of single and multiple sound source localisation, with at most a 2.5° error when localising a single source with a SNR of 10 dB and reverb time of 700 ms.

3.2.1.2 Multiple Signal Classification (MUSIC)

The Multiple Signal Classification (MUSIC) algorithm is a subspace-method for estimating parameters for signals arriving at arbitrarily shaped sensor arrays. This is achieved by decomposing

the covariance matrix of a time-frame into two subspaces relating to the signal and noise component. The method proposed in [40], calculates the MUSIC spectrum for the signal subspace, and uses the largest $\widetilde{d}$ peaks in the spectrum to estimate signal parameters, where $\widetilde{d}$ is the number of predicted signals in a time-frame. The MUSIC algorithm can be used to estimate the DoA, strength and cross-correlation of the signals, polarizations, and strength of the noise component. To compute the MUSIC spectrum for a signal, the $[M \times M]$ covariance matrix, $\boldsymbol{R}_{XX}$, of the $[N \times M]$ (M is number of sensors and N is the length of the signal being analysed) signal, $\boldsymbol{X}$, is first computed as [40, 41],

$$\boldsymbol{R}_{XX} = E\{\boldsymbol{X}^*\boldsymbol{X}\} \tag{3.13a}$$

$$\boldsymbol{R}_{XX} = \frac{1}{N-1}((\boldsymbol{X} - \mu_{\boldsymbol{X}})^T(\boldsymbol{X} - \mu_{\boldsymbol{X}})) \tag{3.13b}$$

where $E\{.\}$ denotes statistical expectation, ^{T} denotes matrix transposition, and $\mu_{\boldsymbol{X}}$ is the mean of the signal matrix $\boldsymbol{X}$. From the covariance matrix the eigensystem can thus be solved for the matrix of generalised eigenvectors, $\boldsymbol{\Omega}_n$, [42] using the relationship,

$$\boldsymbol{R}_{XX}\bar{\boldsymbol{E}} = \boldsymbol{\Omega}_n\bar{\boldsymbol{E}}\boldsymbol{\Lambda} \tag{3.14}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix of the eigenvalues of the covariance matrix and $\bar{\boldsymbol{E}}$ is a matrix containing the eigenvectors ($\mathbf{e}$) corresponding to the eigenvalues ($\bar{\lambda}$) of the covariance matrix [42].

An estimate of the number of sound sources $\widetilde{d}$ present in the signal is then required, which can be calculated using, for example, the Akaike (AIC) approach as defined in [43].

The MUSIC spectrum, $\mathbf{p_{MUSIC}}$, is then calculated from the estimated noise subspace $\boldsymbol{E}_N = \boldsymbol{\Omega}_n[e_{\widetilde{d}+1}|\cdots|e_m]$ [40] as,

$$\mathbf{p_{MUSIC}}(\theta) = \frac{\mathbf{w}^*(\theta)\mathbf{w}(\theta)}{\mathbf{w}^*(\theta)\mathbf{E}_n\mathbf{E}_n^*\mathbf{w}(\theta)} \tag{3.15}$$

where $\mathbf{w}(\theta)$ is the simulated or measured array response for a plane wave arriving at the array

from DoA θ, containing both gain and temporal information [42]. The θ values of the $\widetilde{d}$ largest peaks in the MUSIC spectrum correspond to the estimated DoA for the signals arriving at the microphone array [40].

Results presented in [42] showed that the MUSIC had at most an error of 0.9° as average over 5000 trials with two active sources. However, in 37% of the trials using MUSIC one or both of the sources DoAs were not estimated [42]. Results presented in [45] showed that, when localising two active sources, only small errors [45] were introduced as reverb time increased, with tests performed at 0, 150, and 300 ms reverb times with a SNR of 20 dB. The accuracy of the MUSIC algorithm has made it a popular area of further research [42], with it having applications in a wide array of fields that make use of sensor arrays, particularly with relation to radars.

3.2.1.3 Eigenbeam-Multiple Signal Classification (EB-MUSIC)

Eigenbeam-Multiple Signal Classification (EB-MUSIC) is an implementation of the MUSIC algorithm designed for use with spherical microphone arrays. EB-MUSIC uses spherical beampatterns, or eigenbeams, in place of the steering vector in (3.15) [46]. The steering vector is therefore expressed using the spherical harmonic transform vector $\mathbf{y}$ as defined in (2.22).

As with pseudo-intensity vector analysis (see Section 3.2.1.1), the raw microphone output is transformed into the spherical Fourier domain using (3.5). Then using (3.13a) the noise subspace for the time-frame is computed. The MUSIC spectrum is then computed by replacing the steering vector, $\mathbf{w}(\theta)$, in (3.15) with the spherical harmonic transform vector.

Results presented in [12], which analysed DoA for early reflections measured with a spherical microphone array, presented azimuth error values up to 10° and elevation error values up to 9°, with frequency smoothing[2] applied. Additionally, their results showed that the accuracy with which DoA was estimated decreased with SNR [12]. While it can be beneficial to represent the steering vector analytically, as opposed to through physical measurements, beamforming techniques performance can degrade as a result of differences between the analytical and measured sensor gain, phase, position, and mutual coupling [47–49]. Furthermore, while the results appear less accurate than that of MUSIC, it is important to consider that the types of signals being analysed, in this case reflections, differ from the continuous signals used to test MUSIC.

[2]Frequency smoothing techniques make use of focusing matrices that map all frequency bins into a single reference frequency, effectively focusing the spectral content [12]. Derivation of this process can be found in [12].

Where reflections are short signals that last a fraction of a second, providing less data to estimate DoA from. Additional results presented in [50], looked at the application of frequency smoothing (see [50] for more details) when analysing the DoA of reflections recorded in a 444 seat auditorium (2268 m^3) using a dual sphere scanning microphone with 882 positions per sphere (20^{th} order spherical harmonic signal). When performing the frequency smoothing over the 1.91 kHz–2.73 kHz band, they reported an enhanced spatial spectrum, with smaller regions of higher power exhibited in the spatial spectrum, which should allow for more accurate estimates of DoA. Furthermore, results presented in [51], which used time-domain smoothing (see [51]) to analyse the DoA of the direct sound and first seven reflections in a SRIR measured in a 162 m^3 seminar room using an EigenMike, reported azimuth angular errors between $1°$–$4°$ and elevation between $0°$–$9.5°$.

3.2.1.4 Estimation of Signal Parameters by Rotational Invariance Techniques (ESPRIT)

The Estimation of Signal parameters by Rotational Invariance Techniques (ESPRIT) algorithm was developed for estimating various signal parameters from signals captured using microphone arrays. This method imposes constraints on the sensor array geometry to improve the computational efficiency of the signal parameter estimation [42]. As such, the microphone array is set up in matched pairs, as in Figure 3.1, so as to display a displacement invariance [42].

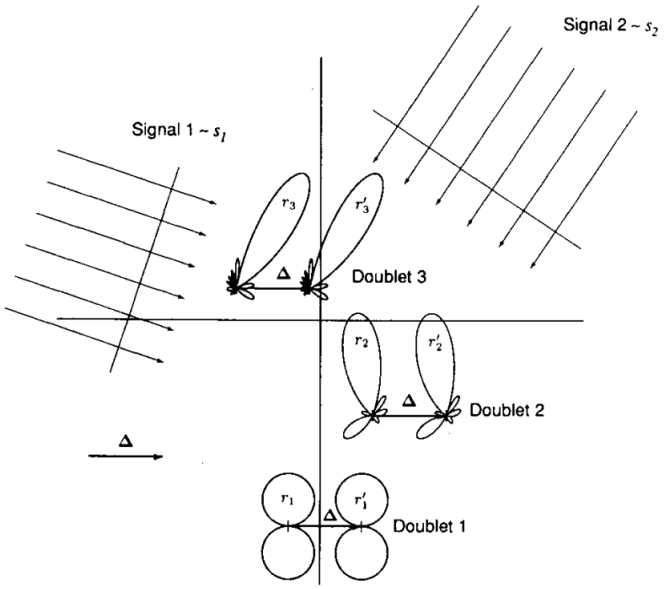

Figure 3.1: An example sensor array geometry for the ESPRIT algorithm, showing three sets of microphone set up in matched pairs. The Δ represents the displacement between the matched pairs. Image from [42]

As with MUSIC, eigendecomposition is performed on the covariance matrix $\mathbf{R}_{XX}$ from (3.13a)

to estimate $\mathbf{\Omega}_n$ as (3.14). The signal subspace $\mathbf{E_s}$ of the covariance matrix is then estimated as the $\widetilde{d}$ eigenvectors corresponding to the $\widetilde{d}$ largest eigenvalues in Ω_n. As with the MUSIC algorithm, $\widetilde{d}$ is defined as the number of sources present, computed using for example the Akaike (AIC) approach as defined in [43].

The $\widetilde{d}$ eigenvectors are partitioned into a $\widetilde{d}$ x $\widetilde{d}$ sub-matrix [42] as,

$$\mathbf{E} \triangleq \begin{bmatrix} \mathbf{E_s}_{1,1} & \mathbf{E_s}_{1,2} \\ \mathbf{E_s}_{2,1} & \mathbf{E_s}_{2,2} \end{bmatrix} \tag{3.16}$$

where $\mathbf{E_s}_{1,1}$ are the first $\widetilde{d}$ x $\widetilde{d}$ entries of the matrix $\mathbf{E_s}$. The final step before estimating the DoA is to calculate the eigenvalues $\hat{\phi}_k$ [42].

$$\hat{\phi}_k = \lambda_k(-\boldsymbol{E}_{12}\boldsymbol{E}_{22}^{-1}) \qquad \forall k = 1, \ ....., \ \widetilde{d} \tag{3.17}$$

where λ_k are the $\widetilde{d}$ eigenvalues of $\mathbf{\Psi} = (-\boldsymbol{E}_{12}\boldsymbol{E}_{22}^{-1})$. The azimuth DoA can then be estimated as,

$$\hat{\theta}_k = \sin^{-1}(c \ arg(\hat{\phi}_k)/(\omega_o\widetilde{\Delta})) \tag{3.18}$$

where ω_o is the centre frequency of the narrow band signal, $\widetilde{\Delta}$ is the array displacement vector and c is the speed of sound [42]. Full derivations of this method can be found in [42].

Comparisons made in [42] showed that the ESPRIT algorithm produced larger variance in angle predictions, $\pm 1.43^\circ$ reported, when compared to the MUSIC algorithm ($\pm 0.9^\circ$), when localising two active source under simulated conditions. The larger variance is a product of the reduced knowledge of the array geometry required for the algorithm to work. This, however, comes with the benefit of improved computational efficiency as only computations of order $\widehat{d^3}$ are required compared to the MUSIC algorithm that performs a search over all steering vectors. This is as a result of the array constraints, which removes the requirement that the whole parameter space be searched.

In [52] an extension to the ESPRIT algorithm was presented for use with uniform linear arrays.

The results presented for this variation of the ESPRIT algorithm produced azimuth estimations within $\pm 0.56^{\circ}$ of the known source positions, offering an improvement over the original implementation, but still not giving results as accurate as those obtained using MUSIC.

3.2.1.5 Eigenbeam - Estimation of Signal Parameters via Rotational Invariance Techniques (EB-ESPRIT)

In [53] an extensions of ESPRIT, referred to as Eigenbeam - Estimation of Signal Parameters via Rotational Invariance Techniques (EB-ESPRIT) was proposed for use with spherical harmonic domain signals (see [53, 54] for derivations). Results presented in [55] showed that the EB-ESPRIT method performs best with higher SNRs, demonstrating that the performance of the method decreases as interfering noise increases. Results are presented as an average root mean squared error across angles from $0^{\circ} \leq \theta \leq 90^{\circ}$ and -10 dB $\leq$ SNR ≤ 30 dB which for the case of uncorrelated signal and noise produced an root mean squared error angular error of 11.6° and for the correlated case 20.9°, and therefore, would be less accurate for analysing reflections. While these errors are generally greater than MUSIC or ESPRIT, for the case of a SNR of 30 dB the angular error was close to zero, although an exact value is not possible to extract from the presented heat maps. Furthermore, findings presented in [12] showed that EB-ESPRIT was only able to localise the direct sound and one reflection within a SRIR, and as such is not suitable for cases when multiple reflections are present.

3.2.1.6 Delay-and-Sum Beamformer

The delay-and-sum beamformer is a classic beamforming technique, which steers the array towards a specific DoA by delaying and summing the received signal [25]. This method uses *a priori* knowledge of the time-difference-of-arrival, $\tau_m(\Psi)$, between each microphone m in the array to a signal from a known DoA $\Psi = [\theta, \phi]$. The microphone array can therefore be steered in the direction of a specific DoA by delaying the recorded signal at each microphone by $\tau_m(\Psi)$, which from [25], is expressed as,

$$\widetilde{\mathbf{x}}(\Psi) = \sum_{m=1}^{M} \mathbf{w}_m \mathbf{X}_m(t - \tau_m(\Psi)) \tag{3.19}$$

where M is the total number of microphones, $\mathbf{w}_m$ is a vector of weights for each microphone, $\mathbf{X}$ is the matrix containing the recorded arrays response at each microphone, and $\widetilde{\mathbf{x}}(\Psi)$ is the resulting beamformer output [25]. For the case of a spherical microphone array the delay-and-

sum beamformer can be, from [56], expressed as,

$$\widetilde{\mathbf{x}}(\Psi, k) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \mathbf{X}_n^m(k) \mathbf{w}_n^m(k) \tag{3.20}$$

where $\mathbf{X}_n^m(k)$ the spherical Fourier domain version of signal $\widehat{\mathbf{X}}$ (3.5) of order n, degree m, and wave number k, and the beamforming weights $\mathbf{w}_n^m(k)$ are computed as [56],

$$\mathbf{w}_n^m(k)^* = b_n^* Y_n^m(\Psi) \tag{3.21}$$

where $(.)^*$ denotes complex conjugate and b_n for a rigid-sphere is, from [27], computed as (3.10).

In [12], the spherical harmonic domain delay-and-sum beamformer was used to estimate the DoA of the first five reflections in a SRIR measured with the EigenMike EM32 [13]. Results show that the delay-and-sum beamformer produced DoA estimation errors up to 5° for azimuth and up to 11° for elevation - when using frequency smoothing, as applied in [12]. While the azimuth DoA estimation accuracy was comparable to the results they presented for the MVDR beamformer, 6°, the delay-and-sum beamformer is less accurate at estimating elevation DoA. This was suggested to be as a result of lower resolution in the acoustic map produced by the delay-and-sum beamforming technique, which resulted in wider regions of higher intensity [12].

3.2.1.7 Plane-Wave Decomposition

A Plane-Wave Decomposition (PWD) beamformer decomposes the sound-field into its component plane-waves based on the measured sound pressure at a microphone array [57]. The beamformer weights $\mathbf{w}_n^m(\Psi) = [M \times 1]$, where M is the number of spherical harmonic channels, can be computed, from [58], as

$$\mathbf{w}_n^m(\Psi) = \frac{Y_n^m(\Psi)}{b_n(k)} \tag{3.22}$$

where $b_n(k)$ for the case of a rigid sphere is, from [27], computed as (3.10). As the array order n approach ∞ the output of the beamformer tends towards a delta function in the DoA [57, 58].

Additional variations on this formulation can be found in: [12, 26, 30, 57, 59, 60]. From these weights, the power in a given steered direction (Ψ) can be estimated, from [30], as,

$$\zeta(\Psi) = \mathbf{w}_n^m(\Psi)^* \mathbf{R}_{\mathbf{H}}(t_f) \mathbf{w}_n^m(\Psi) \tag{3.23}$$

Jarrett *et al.* [26] presented a comparison of results between the plane-wave decomposition technique using 16384 beams (steering directions, the distribution of which was not specified), and the pseudo-intensity vector analysis method. The results presented considered the case of a single source present in a reverberant environment with reverb times between 300-600 ms. Results showed a DoA estimation error of 0.6° and was consistent across reverb times. While the accuracy of pseudo-intensity vector analysis decreased as reverb time increased, it still outperformed the plane-wave decomposition beamformer in the tests presented. Furthermore, in [12] the plane-wave decomposition beamformer was used to estimate the DoA of reflections in a SRIR, only one estimation exactly matched the expected DoA, there was a minimum angular error value of 2° and a maximum of 16° - greater than the results presented for EB-MUSIC and the MVDR beamformer. This implies that the plane-wave decomposition beamformer is not necessarily the best tool for DoA estimation for SRIR.

3.2.1.8 Minimum Variance Distortionless Response (MVDR) Beamformer

The Minimum-Variance Distortionless Response (MVDR) beamformer, sometimes referred to as the *Capon Beamformer*, is a high-resolution beamforming technique that aims to improve the robustness of DoA estimation to noise interference [61]. The MVDR is an adaptive beamformer where the beamformer weights are adjusted based on a signal's covariance matrix, with the aim of minimising the impact of the residual noise component, by minimising the total array output and setting the gain in the steered direction to unity [62]. The beamforming weights are computed as,

$$\widehat{\mathbf{w}}(\Psi) = \frac{\mathbf{R}_{XX}^{-1}\mathbf{w}(\Psi)}{\mathbf{w}^H(\Psi)\mathbf{R}_{XX}^{-1}\mathbf{w}(\Psi)} \tag{3.24}$$

where $\mathbf{R}_{XX}$ is the signal covariance matrix, $\mathbf{w}(\Psi)$ is the steering vector in direction $\Psi = \begin{bmatrix}\theta & \phi\end{bmatrix}$, $(.)^{-1}$ denotes the inverse of the matrix, and $\widehat{\mathbf{w}}$ is the adapted beamforming weights [63]. The directional intensity map, $\zeta(\Psi)$, is computed from the beamformer output, from [63],

across a grid of azimuth and elevation angles as,

$$\zeta(\Psi) = \widehat{\mathbf{w}}^{H}(\Psi)\mathbf{R}_{XX}\widehat{\mathbf{w}}(\Psi) \tag{3.25}$$

where $(.)^{H}$ denotes Hermitian transpose. The steering vector weights can be swapped with the spherical harmonic vector $\mathbf{y}(\Psi_k)$ for application with spherical microphone arrays.

In [12] a spherical array-based MVDR beamformer was compared with EB-MUSIC. The results showed that the MVDR beamformer produced DoA estimation errors up to 6° for azimuth and up to 4° for elevation - when using frequency smoothing, as applied in [12]. These values are slightly improved over the EB-MUSIC algorithm for the same conditions, which had a maximum azimuth error of 10° and maximum elevation error of 9° [12]. Additional results presented in [50], looked at the application of frequency smoothing (see [50] for more details) when analysing the DoA of reflections recorded in a 444 seat auditorium (2268 m^3) using a dual sphere scanning microphone with 882 positions per sphere (20^{th} order spherical harmonic signal). When performing the frequency smoothing over the 1.91 kHz–2.73 kHz band, they reported an enhanced spatial spectrum, with smaller regions of higher power exhibited in the spatial spectrum, which should allow for more accurate estimates of DoA.

3.2.2 Binaural Dummy Heads

3.2.2.1 Interaural Level and Interaural Time Difference Lookup Direction-of-Arrival analysis

In [64] Vesa and Lokki suggested a method for DoA estimation for reflections in BRIRs using measured ILD and ITD from a set of known Head Related Impulse Responses (HRIRs). This method used the continuous wavelet transform to produce a high-resolution frequency domain representations of the BRIR, as,

$$\mathbf{W}_{x_l}(n,s) = \frac{1}{\sqrt{s}}\sum_{n'=0}^{N-1}\mathbf{x}_l(n')\psi_0^{*}\left(\frac{n'-n}{s}\right) \tag{3.26}$$

where $\mathbf{W}_{x_l}(n,s)$ is the transformed signal $\mathbf{x}_l$ (for the left channel) at discrete time index n' and scale s, n is the translation, and $\psi_0(\frac{n'-n}{s})$ is the translated wavelet function at frequency scale s. This equation can be expressed more compactly and efficiently using the FFT as,

$$W_x(s) = \text{IFFT}(\text{FFT}(\mathbf{x}_l)^T . * \psi_0(s)) \tag{3.27}$$

In [64] the Morlet wavelet is used as the complex weighted wavelet function $\psi_0(s)$, which, from [65], in the frequency domain is expressed as,

$$\psi_0(s) = \pi^{-0.25} H(\omega) e^{\frac{-(s\omega-\omega_0)^2}{2}} \tag{3.28}$$

where $H(\omega)$ is the Heaviside step function which equals 1 if $\omega > 0$, ω_0 is the dimensionless oscillating period of the wavelet that determines the frequency resolution of the continuous wavelet transform, and s is the scale factor calculated as [64]:

$$s = s_0 2^{j\delta_j} \qquad j = 0, 1, ..., J \tag{3.29}$$

where s_0 is the smallest resolvable scale, δ_j is the step size of the scale function, and J is the maximum value of the scale calculated as,

$$J = \frac{1}{\delta_j} * log_2\left(\frac{s_{max}}{s_0}\right) \tag{3.30}$$

where s_{max} is the maximum scale value [64]. The scaling of the wavelet functions can be expressed in Hz, from [65], by calculating the relationship between the scale and the Fourier period f_λ as,

$$f_\lambda = \frac{4\pi s}{\omega_0 + \sqrt{2 + \omega_0^2}} \tag{3.31}$$

the frequency in Hz can then be expressed as,

$$f = \frac{fs}{\lambda} \tag{3.32}$$

In [64] the following parameters were used when formulating the wavelets: $\delta_j = \frac{1}{32}$, $J = 288$, $s_0 = 2$ and $s_{max} = 1024$.

Before estimating the DoA, the ILD at each frequency scale (f) is computed as the ratio of total signal energy between the right $\mathbf{W}_{x_r}$ and left $\mathbf{W}_{x_l}$ transformed signals as,

$$\text{ILD}(f) = 20 \times \log_{10}\left(\frac{\sum_{n=1}^{N}|\Re\{\mathbf{W}_{x_r}(f,n)\}|}{\sum_{n=1}^{N}|\Re\{\mathbf{W}_{x_l}(f,n)\}|}\right) \tag{3.33}$$

where $\Re$ denotes the real part of the complex transformed audio vector. The ITD is computed from the cross-correlation function, which is defined as,

$$\mathbf{c}(f,t) = \int_{-1\ ms}^{1\ ms} \mathbf{W}_{x_l}(f,\tau)\mathbf{W}_{x_r}(f,t-\tau)d\tau \tag{3.34}$$

where the maximum peak within the cross-correlation function relates to the ITD between the left and right channel. To produce a more accurate estimate of ITD the method proposed in [66] is used in [64], which upsamples the area around the maximum peak within the cross-correlation function by a factor of ten, which ideally will improve the precision with which the ITD can be estimated.

As a result of the diameter and shape of the human head, these interaural cues are frequency dependent. This results in ITD values being more prominent at lower frequencies, while ILD values are greater at higher frequencies [21, 64]. To this extent, Vesa *et al.* proposed the use of a crossover frequency at $f_c = 1.5$ kHz [64]. The azimuth and elevation angles can then be calculated by comparing the measured ITD and ILD values of the test signal with the ITD and ILD values of the reference HRIRs. The reference HRIRs used in [64] are the MIT KEMAR database [67] and CIPIC database [68] - using an average ILD and ITD across all participants of the CIPIC dataset. The ILD and ITD comparisons are calculated as,

$$(\theta,\phi) = \underset{(\theta,\phi)}{\text{argmax}} \begin{cases} -\sum_{f_{min}}^{f_{max}}(\text{ITD}_{ref}(f,\theta,\phi)-\text{ITD}(f))^2 & ,f\leq f_c \\ -\sum_{f_{min}}^{f_{max}}(\text{ILD}_{ref}(f,\theta,\phi)-\text{ILD}(f))^2 & ,f\geq f_c \end{cases} \tag{3.35}$$

where the reference ILD and ITD values are denoted as ILD_{ref}, and ITD_{ref} respectively and

the ILD and ITD values measured from the signal being tested are denoted as ILD and ITD respectively.

This method was found to be inaccurate for estimating the DoA of reflections in a BRIR, with errors greater than 80° estimated for some reflections, particularly as reflection density, and consequently the number of overlapping reflections, increases [64]. Furthermore, the figures presented in [64] show very few reflections with angular errors less than 10°. While the angular errors are significant this is, to the author's knowledge, the first paper to consider the problem of DoA estimation of reflections within a BRIR. It is possible that through the use of sophisticated pattern recognition algorithms, such as machine learning, these results could be improved on.

3.2.2.2 Machine Learning for Direction-of-Arrival Estimation of Binaural Signals

Machine-learning refers to a category of computer programs that can adapt and learn through trained experience, as opposed to being explicitly programmed to act in a specific way when provided with a given data input [23]. When considering neural networks, as used in this thesis, learning is performed using a training dataset with known solutions. During each iteration of the training process, the neural networks adjustable parameters, referred to as weights and biases, are tuned to minimise the difference between the expected and estimated solutions to the training data [23, 24, 69]. Once trained, the neural network can then be used to find solutions to unknown data of the same type as it was trained with.

While numerous methods for machine-learning exist, the most common approach to binaural DoA estimation involves the use of neural networks [70–78]. Neural Networks (NNs) are powerful machine-learning tools, which can be used to solve complex problems with relative ease, and have been an active area of research in binaural DoA estimation over the past 30 years [70–77].

A NN a is a highly interconnected structure of simple non-linear processing units called *neurons* and in their simplest form have three main layers: the input layer; the hidden layer; and the output layer (see Figure 3.2) [24]. The input layer is a passive layer that takes the input pattern of size I and passes each element of the pattern to each neuron of the first hidden layer, where most of the processing occurs. The hidden layer is formed of neurons that have weighted connections to each feature within the input pattern, typically each neuron will have an additional tunable bias value, which is added to the weighted sum of its connections. The sum of these

weighted connections produce the neuron's *degree of activation* (or a Boolean on/off if a thresholding function is being used), which is formulated using a mathematical function referred to as the *activation function* (for example sigmoid, linear, etc.). Unlike the input and output layer, multiple hidden layers can be defined of different sizes. The output layer defines the solution that the NN has arrived at based on the hidden layer's processing of the input feature vector, and contains neurons equal to the number of possible predefined solutions that exist for a single input pattern. The output is defined by mapping the weighted sum of the hidden layers' neurons to an activation function [24], and therefore defines the probability that an input vector belongs to a specific output value. A basic neuron model can be seen in Figure 3.3.

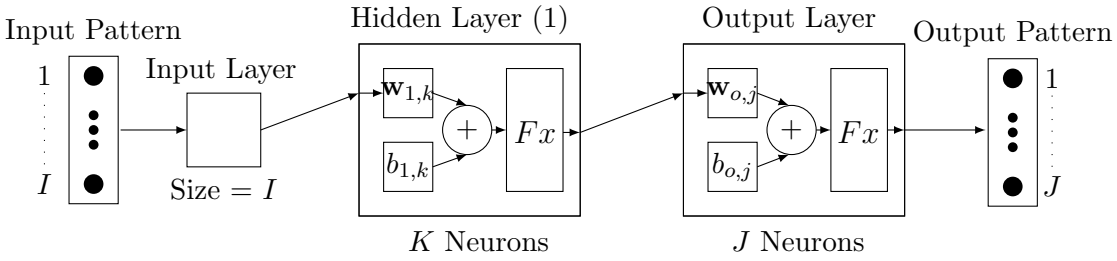

Figure 3.2: Example simple feed-forward neural network topology with an input layer, hidden layer, and output layer. I symbolises the number of elements in a single input pattern, K is the number of neurons in the hidden layer, and J is the number of neurons in the output layer, which corresponds to the number of expected outputs for one input pattern. In this example $\mathbf{w}_{1,k}$ and $\mathbf{w}_{o,k}$ are a vector of weights representing the weighted connections to neuron k in the input and j in output layer respectively, $b_{1,j}$ and $b_{o,j}$ is the bias values for neuron k in the input and j output layer respectively, and Fx is a mathematical function which defines the degree of activation of the neuron. Based on [79]

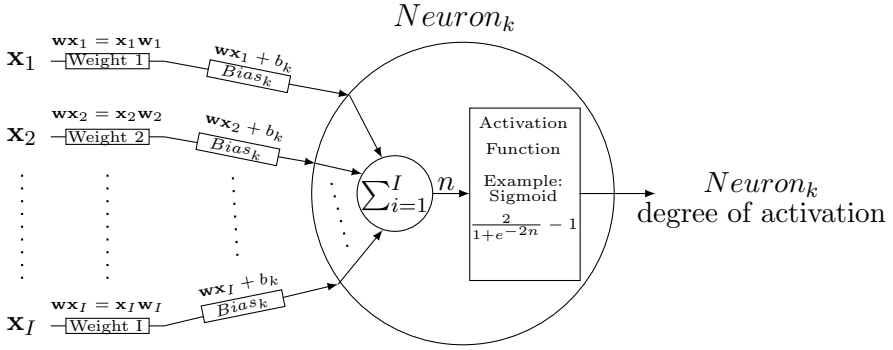

Figure 3.3: Example neuron model for the k^{th} neuron ($Neuron_k$) in the hidden layer of the network. $\mathbf{x}_1 : \mathbf{x}_I$ are the I input elements of the input pattern $\mathbf{x}$, $\mathbf{w}_1 : \mathbf{w}_I$ are the I weights associated with $Neuron_k$, b_k is the bias value associated with $Neuron_k$, and n is the summed processed input. In this case a sigmoid activation function is used to express the degree of activation of the neuron, which will vary from -1 to 1, alternative functions can also be used.

There are two main ways of training a NN: *supervised*, where the NN is given the solution to the

training data; and unsupervised, where the NN is not provided with the desired solution to the training data. In the supervised case, a NN learns by adjusting the weights and biases of each neuron, in an attempt to minimise the error between the predicted and desired output solutions [24]. In the case of unsupervised learning, the NN is looking within the dataset for patterns, and each output layer neuron represents a specific pattern. When a neuron is presented with input data that best matches its prescribed pattern it should have the highest degree of activation in the competitive learning sense. In both cases the training process is iterative until the best solution or maximum number of iterations is reached [24, 69, 80]. The exact procedure of the training process will vary between different network types (feed-forward, feedback, etc.) and different training functions (Bayesian Regularisation [81], Levenberg-Marquardt [82], Scaled Conjugate Gradient [83], etc.). Each network type and training function will lend itself to finding solutions to different problems. Once the NN has been trained, and the weights and biases of the connections fixed, it can be used to predict the probability of an unknown feature vector belonging to one of the predefined solutions.

Direction-of-Arrival Estimation

Considering the sound localisation capabilities of the human auditory system it should be possible to develop a computational approach to binaural sound localisation, by attempting to mimic the neural processing performed by the auditory system. Therefore, machine learning has played a significant role in binaural sound localisation over the past thirty years, as a result of the parallels that can be drawn between machine learning techniques and that of the biological neural system. In the literature for machine learning based binaural DoA estimation, the most common approach is to use a feature space comprised of the ITD and/or ILD as computed through the use of a binaural model [32, 70, 71, 76, 84–86].

When calculating the ILD and ITD, the binaural signals are first filtered into separate frequency bands using either Gammatone filters[3] [32, 76, 84–86], Bark scale filters[4] [89], or logarithmically spaced non-linear filters[5] [70, 71]. A conceptual design for this process can be seen in Figure 3.4.

[3]Gammatone filters are commonly used in computational models of the auditory system, and are designed such that they mimic the frequency separation and resolution of the auditory system [87]. Gammatone filter banks are spaced using equivalent rectangular bandwidths, which distributes the filters across frequency based on their bandwidths [87].

[4]Bark scale filters refer to bandpass filters corresponding to the first 24 critical bands of hearing [88].

[5]Bandpass filters spaced logarithmically along the frequency range, with increasing bandwidths at higher frequencies [71]

One of the key measures of the success of a trained machine-learning algorithm is its ability to produce comparably accurate results for unknown data, which are measured under different conditions (for example: reverb time, noise, source signal, etc.) than that of the training data. This will detail the applicability of these method to real-world situations where the measurement conditions are not controlled. Therefore, when testing these methods, the estimation accuracy can be more realistically defined when the test data is measured under different conditions than that of the training.

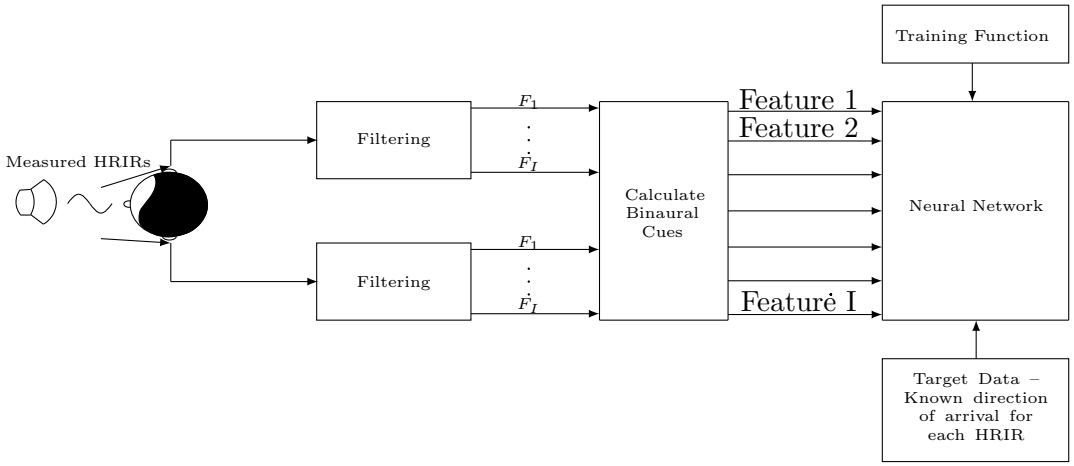

Figure 3.4: Conceptual design of an interaural level and time difference model for neural network based direction-of-arrival estimation, where I is the number of features within the input feature vector.

Neti *et al.* [70] used an ILD based feature space for DoA estimation using 128 logarithmically spaced frequency bands, however, the model was focused on using transfer functions of a cat's external ear. The paper evaluated the importance of different frequency bands for source localisation within the context of NNs trained with a cat's transfer functions. They discovered that the localisation accuracy was better when only using the spectral region 5 kHz to 18 kHz, where prominent notches were found in the transfer functions. The results for a three layer (Input, Hidden, Output) NN, showed an average error of $\pm 6.30^{\circ}$. It is important to note that the test data used here corresponded to measurements from the training data that were not used as part of the training procedure. Therefore, the results do not evidence whether the model is generalisable to different sound source material or measurement conditions than those of the training data. In addition to using ILD Neti et al. [74] explored monoaural DoA estimation using just the spectra from one channel. Their findings showed that the NNs used were still capable of reasonable DoA estimation, with the best average error for monoaural testing being $\pm 9.6^{\circ}$ [74].

Yuhas *et al.* [71] tested two binaural models (the Jeffress model [90] and the Shamma model [91, 92]) to compute the ITD and ILD for azimuthal DoA estimation. Furthermore, they considered

feeding the NN with the raw output of the cochlea model used in the Shamma model [91]. The NN was trained over 1500 epochs for ITD and ILD data, and over 500,000 epochs for the raw cochlea output. The type of NN was not specified, nor the number of hidden layers. The results presented showed high output accuracy within the training data, but only a maximum of 66% of the test data being accurately predicted (see Table 3.1), as with [70] the test data were measurements from the training data set that was not used in the training procedure. However, it is hard to tell to what degree of precision the NN was capable of providing a DoA estimate. The paper suggests an output layer consisting of seven neurons with the neuron with the maximum activation level defining the DoA, which would imply that each neuron represents a possible DoA range of 51° unless further, unknown, processing is considered.

Model	Percent Correct			
	Training Set		Test Set	
	Match	Close	Match	Close
Jeffress Model	79%	84%	46%	64%
Shamma Model				
cross-section	82%	86%	27%	49%
summation	47%	66%	35%	57%
Raw Input				
by sample	29%	71%	18%	54%
12.5ms average	81%	88%	57%	66%
31.3ms average	86%	91%	66%	75%

Table 3.1: Comparative performance of the Jeffres model, Shamma model and Raw input data presented in [71]. 'Match' denotes an exact matching DoA prediction and 'Close' denotes when the DoA was predicted as being on either side of the correct DoA.

Juha *et al.* [89] used downsampled HRIRs (sampled at 22.05 kHz as they were only interested in frequencies $\leq$ 10 kHz) to generate the ITD, represented as the cross-correlation function (limited to ± 1 ms) between the left and right channels, and the ILD computed over 24 Bark filters. The proposed NN had a single hidden layer comprising of either two, four, or eight neurons. The NN output was represented as four neurons with an activation range of ± 1, each representing either the sine or cosine of either the azimuth or elevation DoA. They tested their NN using data excluded from the training data set, and two different acoustic conditions, which was referred to as anechoic and reverberant. It is stated that, based on their own analysis of the results, the NN was unable to generalise to new data. However, angular error values are not reported and cannot be estimated from the NN output graphs provided.

In [76] two different NNs were explored for binaural DoA estimation, the Multilayer Perceptron (MLP) (supervised learning) and the Self-Organising Map (SOM) (unsupervised learning), with the training data generated using the MIT KEMAR HRIR database [67]. The HRIRs are first

filtered using a bank of 32 gammatone filters with Equivalent Rectangular Bandwidth (ERB) spacing [93], using the ITD values up to 2.5 Hz and ILD from 1.9 kHz to 20 kHz to train the NN. However, it was found that ITD and ILD alone were insufficient for accurate estimation using the SOM. To this extent, a second set of features is generated corresponding to a $\pm 15^{\circ}$ rotation of the head, and the SOM is presented with both sets of features. The NNs were tested using pink noise and spoken vowel sounds across different azimuth and elevation positions. In [76] it was found that training the MLP for more than 2000 epochs caused the network to be overtrained and unable to generalise for unknown data. Results presented showed that the relative error of the NNs output for real-world data consisting of spoken Finnish vowels that were not included as part of the training data was 24%, and was said to be generalisable due to a similar relative error presented for the training data (20.7%). The key issue in this work related to elevation prediction in the median plane, where the ILD and ITD values are zero [76]. It is speculated that using the cross-correlation summed across the 32 filter bands, and also the composite loudness level spectra of the 32-filter bands may improve localisation along the median plane [76].

May *et al.* [84] approached the problem from the perspective of using Gaussian Mixture Models (GMMs) to analyse the ITD and ILD data to produce an estimate of DoA. The data is analysed over 32-gammatone filters and half-wave rectified to approximate the inner hair cells within the human ear. The ITD is then computed as the maximum value within the cross-correlation function between the left and right ears for each filter band, and the ILD as the ratio (in dB) between the energy of the signal at the left and right ear summed over a time frame. The parameters are computed over 20 ms time-frames and at a sampling frequency of 5 kHz. A GMM is defined for each gammatone filter band, and the DoA is computed as the maximum value within the summed log-likelihood of each possible DoA across frequency bands. The GMM was trained using a multi-conditional training set, which produced mixtures of training signals with different reverb times and signal-to-noise ratios. This attempts to improve the generalisability of the prediction model to different measurement conditions. As the GMM presented by May *et al.* has a stepped DoA estimation of 5°, a correct prediction is defined as a prediction within $\pm 5^{\circ}$ of the expected DoA. Results are therefore presented in terms of the number of predictions with an angular error greater than 5°, referred to as *anomalies*. The results presented showed that the accuracy of the GMM varied as a function of both reverb time and source-receiver distance, with anechoic conditions having less than 5% anomalies, and with a reverb time of 0.6 s, a maximum of 40-45%.

Woodruff *et al.* [85] also applied GMMs to the binaural DoA problem. The binaural feature

space was computed in 10 ms time-frames over 64 gammatone filters spaced from 80 Hz–5 kHz, with the ITD computed from the maximum peak in the cross-correlation function and ILD as the energy ratio in dB. The input vector is defined using all ITD and ILD values, and fed to a single GMM. The output of the GMM is then integrated over all 10 ms time-frames that define the binaural signal, and the DoA estimated from the resulting probability vector. The results showed that the localisation performance degraded as signal length decreased, signal-to-noise ratio, and source-receiver distance. The accuracy with respect to signal length, when two-talkers were present, was approximately 70% at 100 ms and between 89-95% for 500 ms, 1 s, and 2 s. The accuracy with respect to source distance, again with two-speakers, was approximately 99% at 1 m, 97% at 2 m, and 86% at 4 m. Finally, the accuracy with respect to SNR, for the two-speaker scenario, was approximately 96% at infinite SNR, 95% at 6 dB SNR, and 93% at 0 dBSNR.

In [32] an extension to [84] was presented using a Deep Neural Network (DNN) for each frequency band, as opposed to a GMM. The DNNs consisted of 8 hidden layers, each with 128 neurons, and a sigmoid activation function. The output layer was split into 72 neurons each representing a 5° step in DoA. To resolve front-back confusions in source localisation, a random head rotation was introduced between $\pm 30^\circ$, triggered through use of a motorised dummy head. Results presented for one one–three active sources show an average accuracy of 96% when using a DNN with head rotation and 95% when head rotation was not used. Furthermore, comparing the performance between DNNs and GMM showed that on average the DNN outperformed GMM, which had an average accuracy of 94.2% with head rotation.

May *et al.* [86] presented an extension to their previous work in [84] using head-rotation, as implemented in [86] to further improve the accuracy of the model. Their results showed that the use of head-rotation, multi-conditional training, and integrating the GMM output over frequency produced the most accurate results. The mean accuracy of their model, for one–three active sources, was 91.3% across four measurement environments: anechoic data measured with a KEMAR, and three measurement environments using a Cortex MK.2 head and torso simulator [86]. The accuracy of these results show that their proposed system is not just generalisable to different measurement conditions, but also to alternative head shapes - where differences in binaural cues would be observed. This can prove useful when distributing a trained model designed for general purpose use, where different end users will likely have different measurement equipment.

3.2.3 Discussion

Direction-of-arrival (DoA) estimators refer to the set of methods that aim to determine the direction from which a signal arrived at a microphone array. In this section DoA estimators that are applicable to spherical and binaural microphone arrays have been explored.

Considering the results presented for binaural based localisation first (see Table 3.2), it is clear that NN based DoA estimators are capable of producing comparable accuracy, in some cases, to results presented for larger microphone arrays when considering continuous signals such as speech. Furthermore, while [64] ILD and ITD lookup scheme considered localisation of reflections as opposed to continuous signals, the significant improvement in DoA estimation achieved when using NN presents an area of further research in reflection DoA estimation in BRIR. While these estimators have been studied extensively for estimating the DoA of continuous sound sources, they have not as yet been applied to the analysis of reflections within a BRIR. Work presented in this thesis will, therefore, consider the application of NN based DoA estimators to reflection-based data.

DoA estimation for reflections using spherical microphone arrays (see Table 3.3), however, presents a different set of problems. While results suggest that pseudo-intensity vector analysis has the potential to produce accurate estimates of DoA to within $\pm 0.5^{\circ}$ (as tested with continuous signals), it is not necessarily optimal when considering SRIRs, as in such multiple reflections will arrive at the receiver array, some of which will overlap or arrive at the same time. In these cases it would be expected that the DoA estimation accuracy for the pseudo-intensity vector analysis would decrease similarly to how performance degrades with increasing levels of interfering noise. Furthermore, in the case of simultaneous reflections, pseudo-intensity vector analysis method may not be able to resolve between and so estimate DoA for each of the simultaneously arriving reflections, as previous methods for multiple source localisation have relied on estimating the DoA of multiple signals over time-frames of a continuous signal as opposed to a short time signal such as a reflection, and so in such circumstances, a beamforming based approach is more appealing. Discussion and results presented in [12] showed that, generally, the MVDR beamformer produces comparable accuracy to that of EB-MUSIC when estimating the DoA of reflections measured with a spherical microphone array, without the need to estimate the number of signals that are present in a time-frame. Furthermore, the maximum angular errors, 6° for azimuth and 4° for elevation, are lower than those reported for the other spherical microphone-based multi-signal DoA estimators. Based on these findings for DoA estimation

of reflections [12], the MVDR beamformer will form the basis by which the DoA of reflection based data are estimated when using spherical microphone arrays in this thesis.

Method	Test Condition	Angular Error	Reference
	Binaural		
ILD ITD Lookup	Reflections in BRIR	10° – 90°	[64]
NN Spectral Information	Data excluded from training set	6.30°	[70, 74]
NN Raw Cochlea Output	Data excluded from training set	66% within ±5°	[71]
NN ITD and ILD (downsampled)	Data excluded from training set	Reported as not generalisable	[89]
MLP ITD and ILD	pink-noise and Finnish vowel sounds	Relative error of 24%	[76]
GMM ILD and ITD	Continuous speech, 0–600 ms reverb time, different source-to-receiver distance	95% within 5° at 0 ms 55%–60% within 5° at 600 ms	[84]
GMM ILD and ITD	Two active speakers	minimum of 70% and maximum of 99% within 5°	[85]
NN for each frequency band IACC, ILD, and head rotation	One–three active speakers Anechoic and 320–890 ms reverb time	96% within 5°	[32]
GMM ITD and ILD	One–three active speakers Anechoic and 'strong reverberation'	91.3% within 5°	[86]

Table 3.2: Comparison of direction-of-arrival estimators presented in Section 3.2, presenting method, test conditions, and results.

3.3 Reflection Detection

An RIR, in the perfectly ideal cases, consists of a superposition of the direct source-to-receiver sound followed by reflected copies of the direct sound produced by interactions with boundaries present in an environment, with the density of the reflections arriving increasing with reflection order. However, for real-world measurements interfering noise components can mask desired reflections. Furthermore, reflections are not represented by a single peak within the RIR and as such can temporally overlap, and interactions with the boundaries can temporally warp the reflected sound [94]. These factors can present problems when trying to systematically detect these reflections. Therefore, reflection detection refers to the set of methods aiming to detect individual reflections in a RIR as discrete arrivals, while ideally rejecting interfering noise as a possible reflection. Generally such approaches can be split into one of two categories:

- Microphone Array Based - Making use of particular properties for a specific microphone array.

- System Agnostic - Not requiring any specific setup and usable with any number of microphones.

In this section reflection detection techniques proposed in the literature that are either, relevant to the microphone arrays used in this thesis, or have been used in previous work for geometry inference, will be explored.

3.3.1 Microphone Array Based

3.3.1.1 Circular Variance Local Maxima Technique

The circular variance local maxima technique was developed by Tervo et al. in [14], where they used the first-order components of spherical harmonic domain signals and microphone arrays consisting of two microphones per axes. This method uses a discrete-time implementation of intensity vector analysis (See section: 3.2.1.1), where the frequency-domain representation of a time-frame is computed using the FFT, to calculate the DoA across frequency bins for a windowed time-frame of the RIR. The DoA variation (circular variance) between frequency bins is then used as one of the parameters defining whether a discrete reflection is present in the time-frame. The circular variance, v_{t_f}, is calculated as,

$$v_{t_f} = 1 - (s_{t_f}^2 + c_{t_f}^2)^{\frac{1}{2}} \tag{3.36.1}$$

where

$$s_{t_f} = \frac{1}{(k_2 - k_1)} \sum_{k=k_1}^{k_2} \cos(\theta_{t_f,k}) \tag{3.36.2}$$

$$c_{t_f} = \frac{1}{(k_2 - k_1)} \sum_{k=k_1}^{k_2} \sin(\theta_{t_f,k}) \tag{3.36.3}$$

where s_{t_f} is the average cosine azimuth DoA (θ in radians), from the first frequency bin (k_1) to the last frequency bin (k_2) at the time-frame t_f, and c_{t_f} is the average sine value for θ at time-frame t_f. Ideally if the most prominent signal in a time-frame is a singular discrete reflection, then the variation in DoA should be near zero, and the circular variance will be larger in the presence of noise or multiple arrivals. The use of circular variance alone is insufficient for

accurate reflection detection due to potentially unwanted, quieter, directional signals. To this extent, the circular variance value is used in conjunction with local maxima detection [14]. Local maxima detection compares the average energy across the array contained within the current time-frame, against the previous and next time-frame. A discrete reflection is then defined as being present if the current time-frame is a local maximum and the circular variance is below a defined threshold close to zero, defined as a circular-variance less than 0.1 in [14].

Results presented in [14] show that the circular variance local maxima technique was able to detect more reflections when a RIR (in this cases measured in an auditorium) was measured using a highly-directional loudspeaker (54 potential reflections detected), than when an omnidirectional loudspeaker was used (16 potential reflections detected). In the majority of cases the peaks in the RIR are detected, however, there are additional detections around these peaks, which could be as a result of the same reflection being detected as multiple discrete arrivals due to the windowing process [14]. Furthermore, the results present in [14] do not consider the number of false-positive detections that are made.

3.3.1.2 Cross-Wavelet Transforms

In [64], a method for detecting reflections from the Cross Wavelet Transform (XWT) of a BRIR, measured using a binaural dummy head, was proposed. In order to calculate the XWT, the continuous wavelet transforms of the left and right BRIR channels must first be calculated using (3.27-3.32). The cross-wavelet transform is computed from the continuous wavelet transform as [64],

$$|\mathbf{W}_{x_{l,r}}(n,s)| = |\mathbf{W}_{\mathbf{x}_l}(n,s)\mathbf{W}^*_{\mathbf{x}_r}(n,s)| \tag{3.37}$$

where $W_{\mathbf{x}_l}(n,s)$ is the continuous wavelet transform of the left channel of the BRIR, and $W^*_{\mathbf{x}_r}(n,s)$ is the complex conjugate of the continuous wavelet transform of the right channel (3.26) [64]. The cross-wavelet transform in this case measures the similarity between the signals arriving at the left and right ear, with peaks in the spectrum generally indicating a high correlation between the left and right signals or a significant peak in one of the channels. In the proposed implementation a threshold is applied to the XWT, discarding any parts of the spectrum that are over 14 dB lower than the maximum value, effectively removing any areas of low correlation in the XWT [64]. The discarded parts of the spectrum are set to $-\infty$, and the thresholded transform is normalised so all values above $-\infty$ are scaled between 0 and 1 [64].

Searching for maxima within the thresholded XWT (Figure 3.5 (b)) is not sufficient for detecting reflections, as any temporally overlapping reflections will form a single region of high correlation, and therefore not be detectable as individual arrivals [64]. To this extent Vesa *et al.* proposed that the *watershed* [95] algorithm[6] could be applied to a grey-scale image of the XWT to separate these regions of high-intensity [64]. The segmented regions of high-correlation (Figure 3.5 (c)) are then defined as being the temporal regions in which discrete reflections are present.

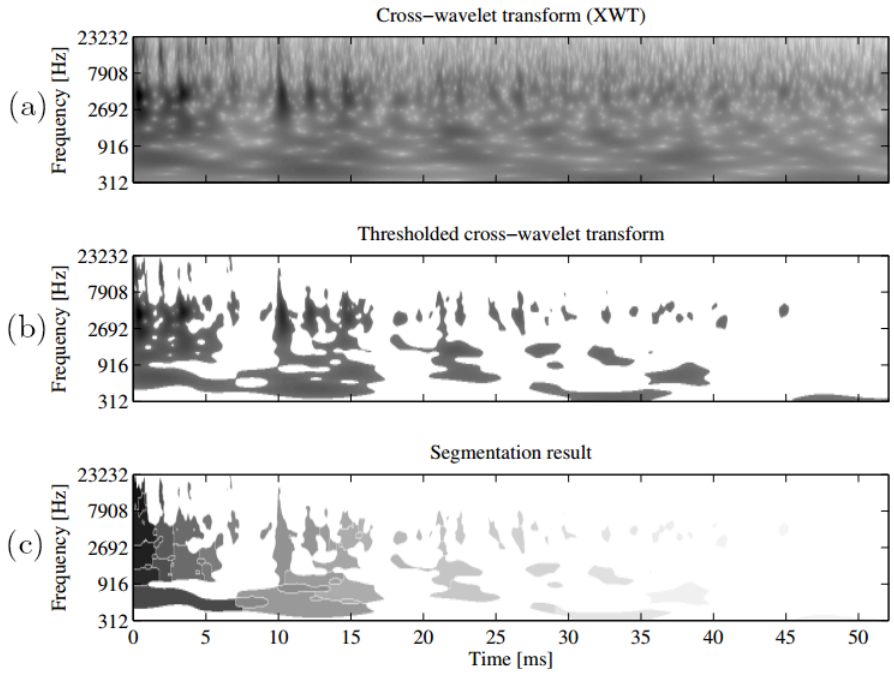

Figure 3.5: Example of the cross-wavelet transform of a binaural impulse response. (a) The cross-wavelet transform between the two channels of the binaural room impulse response. (b) The regions of high correlation present within the binaural room impulse response once the cross-wavelet transform has been thresholded. (c) the segmented regions of high-correlation produced by the watershed algorithm, which are separated by a white outline. Image from [64]

Results presented in [64] show that the reflections detected using this algorithm are all within 2 ms of the expected ToA. The method, however, is unable to disambiguate between reflections that partially overlap, detecting them as a single arrival [64], and therefore, as reflection density increases, the accuracy with which reflections can be detected as individual discrete events decreases. In addition to this drawback, the direct sound is sometimes split into smaller segments by the *watershed* algorithm and may require manual temporal localisation [64]. However, these results do not present the number of reflections that the method fails to detect or the number of

[6]The *watershed* [95] algorithm is an image processing technique that is used to separate overlapping objects within grey-scale images.

false-positive detections.

3.3.1.3 Linear Radon Transform

Baba *et al.* [96] presented a reflection detection method designed for use with RIRs measured with a uniform linear array of loudspeakers positioned along a line with uniform element spacing, and a single receiver position. They exploited the linear temporal displacement property of the array to detect reflections that are common across the RIRs obtained. When considering a uniform linear array of loudspeakers, reflections arriving at the receiver for each loudspeaker will be displaced in time linearly as a product of the distance from the loudspeakers to the walls and receiver. Therefore, the arrival of a reflection from a wall will be linearly displaced across the RIRs, such that the location of the reflection's peaks across the RIRs represent points on a line (see Figure 3.6). Based on this principle the authors proposed the use of the linear Radon transform [97] to detect these lines defined by temporally displaced signals common across the RIRs. These detected lines are then defined as the arrival of a discrete reflection produced by a common wall. The sample index of the point on the line that corresponds to each RIR defines the ToA for a discrete reflection.

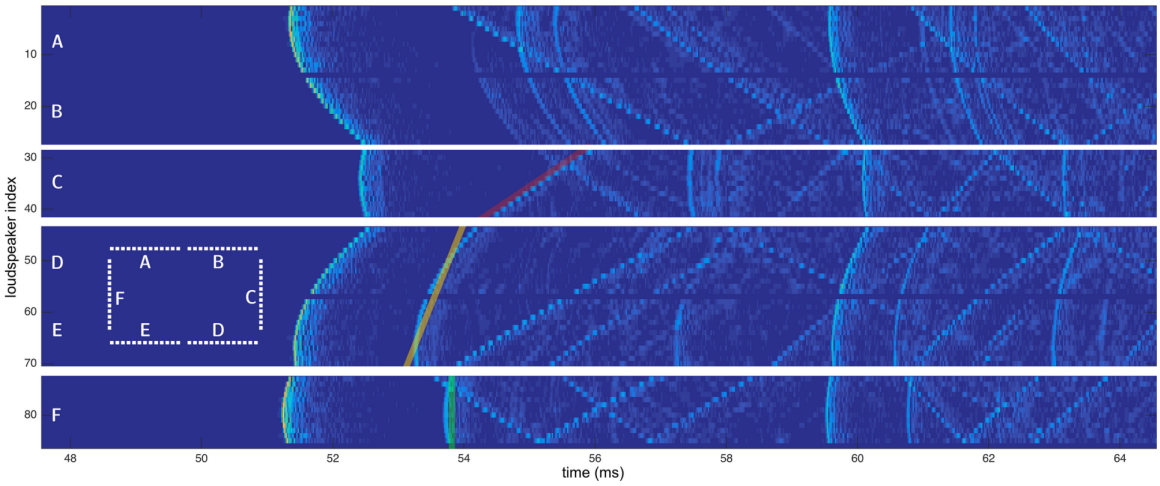

Figure 3.6: Example of a stack of 80 room impulse responses, the white dotted rectangle represents the loudspeaker array used in [9], and the green, yellow, and red lines show the arrival of the reflections linear displaced along the room impulse responses. Image from [9]

To produce the stacked response of the array, the individual RIRs are interpolated such that one pixel on the horizontal axis (time axis) is equal to a time shift of $\frac{1}{Fs}$, and a pixel on the vertical axis (representing microphone spacing) is equal to a distance of $\frac{1}{750}$ m [96]. The Radon transformed image $\mathbf{T}$ is then computed from the interpolated image $\mathbf{R}$ as,

$$\mathbf{T}[j,n] = \sum_{m=1}^{M} \mathbf{R}[m, n + (m - M/2) \times tan(\theta_j)] \tag{3.38}$$

where m is the loudspeaker index, M is the number of loudspeakers, n is the time index, and θ_j is the jth angle - defined as $-15^\circ \leq \theta \leq 15^\circ$ in 0.5° increments. The transformed image is filtered to isolate the highest peaks in the transform, and peak detection is used to detect the time index of the detected reflection in the image.

Results are presented for seven simulated and one real-world case, and consider the detection rates and RMS error of the ToA estimates. Across all eight cases the direct sound is detected with an average error of 104.8 μs for the simulated cases, and 116.1 μs for the real-world case. On average 92.6% of the first-order reflections are detected in the simulated case with an average ToA error of 66.5 μs, and 83.3% of first-order reflections for the real-world case with an average ToA error of 138.7 μs. They report that in the real-world measurements one of the boundaries was near-anechoic, and thus missed first-order reflections could be attributed to these reflections not having been captured. Finally, for the case of second-order reflections there was a large drop in performance, with, on average, only 37.7% of second-order reflections being detected for the simulated case with an average ToA error of 59.9 μs, and 32.0% of second-order reflections detected for the real-world data with an average ToA error of 166.3 μs. From these results it can be seen that as reflection order increases the performance of the algorithm decreases. While this method does not relate to the microphone arrays used in this thesis, this method is used to detect reflections in other work presented by Baba *et al.* on geometry inference in [9], which is discussed in the next chapter.

3.3.1.4 Clustered - Dynamic Phase-Slope Algorithm

The Clustered - Dynamic Phase-Slope Algorithm (C-DYPSA) was proposed in [8] for detecting reflections present across an array of microphones. The proposed C-DYPSA is an extension of DYPSA, which was originally proposed in [98] for the detection of glottal closure instances, in speech research [98]. The phase-slope is computed as the 'centre-of-gravity' $\mathbf{g}(n)$ of the signal energy from a sliding M-length windowed time-frame of the residual signal produced by a linear prediction filter [98] as,

$$\mathbf{g}(n) = \frac{\sum_{m=1}^{M} m\mathbf{x}_n^2(m)}{\sum_{m=1}^{M} \mathbf{x}_n^2(m)} \tag{3.39}$$

where $\mathbf{x}_n$ is the nth windowed time-frame of the residual signal. The vector $\mathbf{g}(n)$ is then centred on sample n as,

$$\mathbf{d}(n) = \mathbf{g}\left(n - \frac{M-1}{2}\right) - \frac{M-1}{2} \tag{3.40}$$

Naylor *et al.* defined the presence of glottal closure instances as indexes where zero-crossings in the phase-slope $\mathbf{d}(n)$ occur [98]. To adapt the algorithm for processing RIRs, any peaks within the phase-slope and RIR that fall below a defined threshold are ignored [8]. Furthermore, the k^{th} reflection in a RIR is defined as a false positive if the median ToA of the k^{th} reflection across all RIRs is closer to the $k + 1^{th}$ than the k^{th} reflection [8]. The remaining zero-crossings within the phase-slope are then defined as the ToA of discrete reflections.

Results presented in [99], however, only present the ToA estimation error for the first reflection that arrives at the microphone array. The results are presented in terms of distance error, representing the error in the estimated distance travelled by the sound wave versus the simulated distance travelled. The average distance error across the four scenarios presented is 109.51 mm, with a maximum error of 192 mm and a minimum of 48 mm. Results considering the detection rates or the number of false-positive detections are not presented, other than stating that it fails for higher-order reflections. This method is not directly relatable to later work presented in this thesis, however, it is used as the reflection detection algorithm in the geometry inference work of Remaggi *et al.* presented in the next chapter of this thesis.

3.3.2 System Agnostic

3.3.2.1 Adaptive Thresholding

Adaptive thresholding is a method originally used for image processing, but was adapted to allow for the detection of individual reflections in a single RIR [100]. Adaptive thresholding compares the average magnitude in a time-frame against neighbouring samples. It is assumed that the magnitude of a specular reflection is a factor of ϵ greater than its neighbouring samples [100]. The mean magnitude value for time (t) can be calculated as,

$$\mu_{local}(t) = \frac{1}{T\mu_{local}} \int_{t-T\mu_{local}}^{t+T\mu+_{local}} |\mathbf{h}(\tau)| d\tau \tag{3.41}$$

where μ_{local} is the average magnitude, $T\mu_{local}$ is the averaging time (2 ms in this study), and $\mathbf{h}(\tau)$ is the RIR at time interval τ [100]. Individual reflections can then be detected through analysis of the mean magnitude values using the following,

$$\mathbf{h}_{peaks}(t) = \begin{cases} 0, & \forall |\mathbf{h}(t)| < \epsilon \mu_{local}(t) \\ 1, & \forall |\mathbf{h}(t)| \geq \epsilon \mu_{local}(t) \end{cases} \tag{3.42}$$

where ϵ is the thresholding parameter which Kuster defined as being two [100]. Therefore, a discrete reflection is defined as being present at points where the sample magnitude is greater than or equal to two times the mean magnitude value of the time-frame.

In [100] the adaptive thresholding method is only able to detect 50% of reflections present in the first 30 ms of a RIR measured in a lecture hall. Furthermore, they present two cases for measurements in a concert hall, both of which had large numbers of false-positive detections in the first 120 ms, although, they do not state how many. Extracting an estimate of false-positive detections from the figures presented in [100], it would seem that approximately 56 detections have been made for the first concert hall measurement, where they suggest there should only be nine reflections. They concluded that this method is not accurate enough for systematic detection of reflections, and at most only one to five reflections can be identifiable with confidence.

3.3.2.2 Matching Pursuit

Defrance *et al.* [101] developed a reflection detection method based on the principal that reflections are filtered occurrences of the direct sound. Therefore, the direct sound will be highly correlated with any reflections present within the RIR, and as such the sample index of the reflection, idx, within the RIR $\mathbf{h}$ can be detected as,

$$idx = \text{argmax}(\langle |\mathbf{r}_h, \mathbf{ds}| \rangle) \tag{3.43}$$

where $\mathbf{r}_h$ is the residual signal, $\mathbf{ds}$ is the direct sound, and $\langle ., . \rangle$ denotes the dot product. This process is then repeated over a finite number of iterations based on a defined stopping criteria, where the residual signal is updated after each iteration by zeroing the time-frame of the previously detected reflection. In [101] it was proposed that this iterative process is stopped once the energy-ratio fell below 20 dB, where the energy ratio is defined as the RIR over the residual $\mathbf{r}_h$ in

dB. The results showed that the choice of value for the stopping criterion is incredibly important when detecting reflections in RIRs, and in this case the higher the allowed energy-ratio the more signals are detected, which can lead to an increased number of false positive detections.

The reliability of this method, however, directly relates to accuracy the with which the direct sound can be windowed out of the RIR by human observation. Furthermore, the process by which reflections are systematically detected, and then removed from the residual RIR, will inevitably result in overlapping reflections being removed, and therefore, not detected as an individual arrival.

3.3.2.3 Dynamic Time Warping Reflection Detection

Kelly and Boland [15] proposed an extension to the work presented by Defrance *et al* [101], by considering the problem of detecting overlapping reflections as individual arrivals. Through the use of DTW they define a likelihood metric which defines the number of reflections present, while also considering the impact of temporal smearing of reflections as a result of sound interacting with the measurement environment. DTW is an operation which computes the minimum warping path required to align the features of one vector with another [102, 103].

To estimate the ToA of the reflections, two cross-correlation functions are computed: one for the cross-correlation between the remaining RIR and the direct sound, and one for the cross-correlation between the remaining RIR and a phase inverted version of the direct sound. The maximum point of correlation is then detected in each vector, with the largest value of correlation between vectors being defined as the most prominent reflection. To detect whether overlapping reflections are present, five new vectors are generated, one for just the direct sound[7], and four with concatenated versions of the direct sound [15] expressed as,

$$\mathbf{ds}_k = [\mathbf{0}_{\tau_k}, \mathbf{ds}, \mathbf{0}_{length(\mathbf{r}_h)-\tau_k-length(\mathbf{ds})}] \tag{3.44.1}$$

$$\widehat{\mathbf{ds}}_k^{+} = [\mathbf{0}_{\tau_k}, \mathbf{ds}, \mathbf{ds}, \mathbf{0}_{length(\mathbf{r}_h)-\tau_k-(2*length(\mathbf{ds}))}] \tag{3.44.2}$$

$$\widehat{\mathbf{ds}}_k^{-} = [\mathbf{0}_{\tau_k}, \mathbf{ds}, -\mathbf{ds}, \mathbf{0}_{length(\mathbf{r}_h)-\tau_k-(2*length(\mathbf{ds}))}] \tag{3.44.3}$$

$$\widetilde{\mathbf{ds}}_k^{+} = [\mathbf{0}_{\tau_k-length(\mathbf{ds})}, ds, \mathbf{ds}, \mathbf{0}_{length(\mathbf{r}_h)-\tau_k-length(\mathbf{ds})}] \tag{3.44.4}$$

$$\widetilde{\mathbf{ds}}_k^{-} = [\mathbf{0}_{\tau_k-length(\mathbf{ds})}, -\mathbf{ds}, -\mathbf{ds}, \mathbf{0}_{length(\mathbf{r}_h)-\tau_k-length(\mathbf{ds})}] \tag{3.44.5}$$

[7] for notation $\mathbf{ds}$ refers to either the phase inverted or original version of the direct sound depending on which produced the largest point of correlation

where $\mathbf{0}_{\tau_k}$ is a zero vector of length τ_k, τ_k is the location of the peak in the cross-correlation vector, $\mathbf{ds}$ denotes the original direct sound, and $\mathbf{r}_k$ is the residual of the RIR being analysed. In the above example $\mathbf{ds}_k$ is just the direct sound, $\widehat{\mathbf{ds}}_k^+$ is with an unchanged direct sound succeeding $\mathbf{ds}$, $\widehat{\mathbf{ds}}_k^-$ is with a phase inverted direct sounding succeeding $\mathbf{ds}$, $\widetilde{\mathbf{ds}}_k^+$ is with an unchanged direct sounding preceding $\mathbf{ds}$, and $\widetilde{\mathbf{ds}}_k^-$ is with a phase inverted direct sounding preceding $\mathbf{ds}$. The main direct sound $\mathbf{ds}$ in these vectors will now be approximately temporally aligned with a possible reflection [15]. DTW is then used to align the features of each variation of $\mathbf{ds}_k$ with the candidate reflection present in residual $\mathbf{r}_h$ of the RIR at τ_k, and the resulting warped versions of $\mathbf{ds}_k$ are scaled to roughly match those of this candidate reflection. The warped and scaled version of $\mathbf{ds}_k$ that best represents the reflections at τ_k is defined using the error value ϵ from [15] as,

$$\upsilon = ||\gamma_{\mathbf{ds}_k} \widehat{\mathbf{w}}_h^\dagger \widehat{\mathbf{w}}_a \mathbf{ds}_k||_{l2} \tag{3.45}$$

where $\widehat{\mathbf{w}}_h$ is the warp vector for the reflection, and $\widehat{\mathbf{w}}_a$ is the warp vector of the direct sound, and $\gamma_{\mathbf{ds}_k}$ is the scaling value. The variation of $\mathbf{ds}$ with the smallest υ is assumed to best represent the reflections present in the time-frame [15].

Results presented by Kelly and Boland in [15] show that the proposed method outperformed the matching pursuit method. Their results show that the proposed method detected 75 reflections, where an estimated 70 reflections should be present based on the image-source method. Unlike the matching pursuit method, the dynamic time warping approach detected individual reflections as a singular arrival as opposed to multiple arrivals [15]. The proposed method also detected far fewer definite false-positive detections, when comparing the number of detections to the number of expected. However, exact numbers of false-positives are not presented so while there are five definite false-positives (based on the number of detections), more could still exist within the 70 other detections.

3.3.3 Discussion

Reflection detection refers to the set of methods designed to find the temporal locations of reflections present in a RIR. In general these methods can be split into one of two categories, microphone array based and system agnostic. Considering the system specific methods that relate to the arrays used in this thesis, which are the circular-variance local maxima (first-order components of a spherical harmonic signal) and cross-wavelet transform (binaural dummy head),

it is evident that neither would be capable of providing accurate enough reflection detection to be considered viable for geometry inference - with large numbers of missed reflections and/or false-positives. Furthermore, even the best system agnostic technique, the DTW based matching pursuit, has drawbacks. While this method can detect overlapping reflections, it would still be unable to detect simultaneous reflections, which can commonly occur in real-world RIRs, as individual reflections. Furthermore, this approach would not fully exploit the spatial information contained within a SRIR measured with a spherical microphone array, which can provide more information about the arrivals of reflections. Therefore, a spatiotemporal decomposition based reflection detection algorithm will be presented in this thesis, and the results compared an implementation of the DTW and circular variance local-maxima approaches.

3.4 Summary

In this chapter, literature relating to reflection analysis, which is a prerequisite step for geometry inference, has been discussed. In the context of this thesis, reflection analysis refers to the temporal and spatial localisation of reflections present in a SRIR, and as such refers to reflection detection and direction-of-arrival estimation.

Reflection detection has been an active area of research in the field of acoustics, with numerous methods proposed. These techniques can be categorised as either being microphone array based or system agnostics - not requiring a specific setup or number of microphones. While some of these techniques have been shown to produce accurate results, the fundamental problem they do not consider is the case where simultaneously arriving, or overlapping, reflections are present. With a view of dealing with this problem, Chapter 6 presents a spatiotemporal decomposition based reflection detection method and will be compared to the DTW based matching pursuit and circular variance local maxima techniques.

As with reflection detection, DoA estimation has been an active area of research, particularly as it is applicable to multiple aspects of audio engineering. From the perspective of reflection analysis, however, very little research has been focused on DoA estimation for reflections in binaural room impulse responses, and none have considered the current state-of-the-art binaural DoA estimators. Hence, a binaural model fronted NN based approach to reflection DoA analysis will be presented in Chapter 5. By way of contrast, DoA estimation of reflections in SRIR captured with spherical microphone arrays has been rigorously studied, and beamforming based techniques have been found to produce at most an angular error 6°. As a consequence, the spherical harmonic domain MVDR beamformer will be used in Chapter 6 for reflection DoA

estimation.

Method	Test Condition	Angular Error	Reference
Spherical Microphone Array			
Pseudo-Intensity Vectors	'Typical Environments'	0.5°	[26]
Pseudo-Intensity Vectors	Three static sound sources with 40 dB SNR	5.71°	[38]
Pseudo-Intensity Vectors	Three moving sound sources with 40 dB SNR	9.28°	[38]
Pseudo-Intensity Vectors	Single and multiple sources 10 dB SNR and 700 ms reverb time	2.5°	[39]
MUSIC	Two active sources (Simulated)	0.9°	[42]
EB-MUSIC	Early Reflections measured with EigenMike EM32	$\theta = 10°$ and $\phi = 9°$	[12]
EB-MUSIC and MVDR	Frequency smoothing, reflections in a SRIR measured with a dual sphere scanning microphone	Reported enhanced spatial spectrum	[50]
EB-MUSIC	Time-domain smoothing, direct sound and reflections in a SRIR measured using an EigenMike	$\theta = 1°$–$4°$ $\phi = 0°$–$9.5°$	[51]
ESPRIT	Two active sources (Simulated)	1.43°	[42]
EB-ESPRIT	-10 dB $\leq$ SNR $\leq$ 30 dB Uncorrelated signal and noise	11.6°	[55]
EB-ESPRIT	-10 dB $\leq$ SNR $\leq$ 30 dB Correlated signal and noise	20.9°	[55]
Eigenbeam-Delay-and-Sum	Early Reflections measured with EigenMike EM32	$\theta = 5°$ and $\phi = 11°$	[12]
Plane-wave decomposition	Single sound source 300-600 ms reverb time	16°	[26]
Plane-wave decomposition	Early Reflections measured with EigenMike EM32	16°	[12]
MVDR	Early Reflections measured with EigenMike EM32	$\theta = 6°$ and $\phi = 4°$	[12]

Table 3.3: Comparison of direction-of-arrival estimators presented in Section 3.2, presenting method, test conditions, and results.

Chapter 4

Geometry Inference: Related Work

4.1 Introduction

In the previous chapter, relevant methods presented in the literature for estimating the time-of-arrival (ToA) and direction-of-arrival (DoA) of reflections in (Spatial) Room Impulse Responses were discussed. These reflection analysis stages are generally a prerequisite for geometry inference, as the information extracted for each reflection is directly relatable to the boundaries present in the measurement environment. This chapter will now consider such geometry inference methods, and discuss the limitations of these techniques.

Geometry inference focusses on the inverse problem of localising reflective boundaries based on temporal or spatiotemporal reflection information from a number of RIRs, exploiting the inherent relationship between reflections arriving at a microphone array, and the location of reflective boundaries present within the environment [6]. Methods for geometry inference generally fall into one of two categories, image-source reversion and direct localisation [8], and both will be discussed in this chapter.

4.2 Image-Source Reversion

Image-source reversion refers to the set of methods that exploit the properties of the image-source model, as discussed in Chapter 2, to estimate the location of dominant reflective bound-

aries in an enclosed space. Typically these methods use the time-of-arrival of first-order reflections extracted from a set of RIRs, measured at different points in an environment, to estimate the location of image-sources, which are then used to estimate boundary locations. Most of these methods exploit the relationship between the source $\mathbf{s}$ and image-source location $\widetilde{\mathbf{s}}$ to define a point on the boundary $\widetilde{\mathbf{b}}$ and the boundary's normal $\widetilde{\mathbf{n}}$ (see Figure 4.1) as,

$$\widetilde{\mathbf{b}} = \frac{\widetilde{\mathbf{s}} + \mathbf{s}}{2} \tag{4.1}$$

$$\widetilde{\mathbf{n}} = \frac{\widetilde{\mathbf{s}} - \mathbf{s}}{||\widetilde{\mathbf{s}} - \mathbf{s}||} \tag{4.2}$$

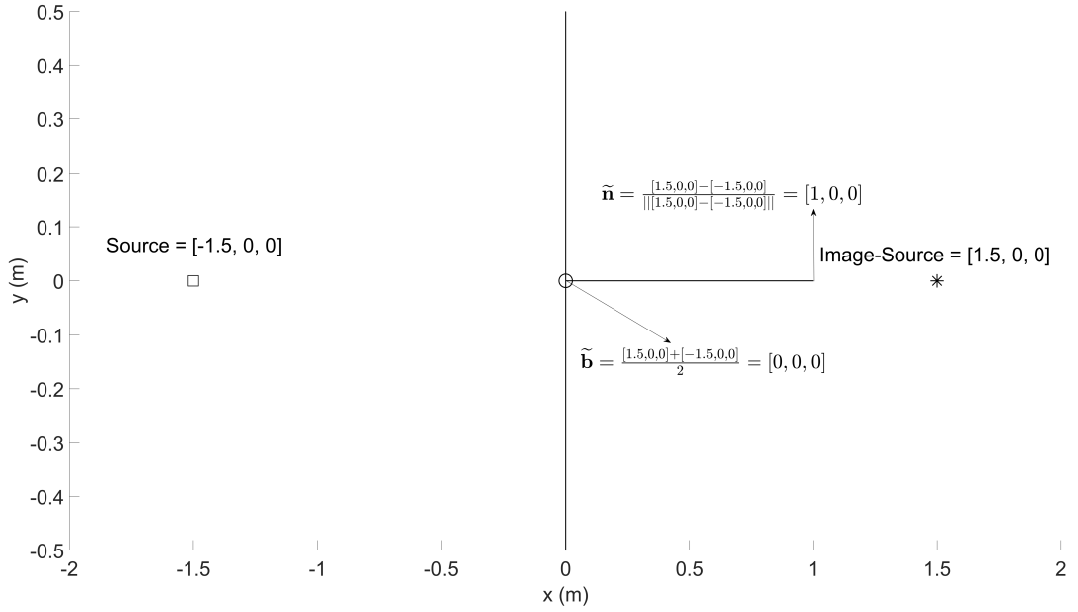

Figure 4.1: Simple example showing the inverse image-source process used to estimate a point on a boundary and the boundary's normal vector.

4.2.1 Euclidean Distance Matrix: Echo Sorting and Geometry Inference

The method proposed by Dokmanic et al. [6] assumes that the geometry of a room is a convex polyhedron, and that five RIRs, each measured at one of five different receiver positions, is sufficient for geometry inference of a three-dimensional space. To estimate the location of image-sources, the ToA of the first-order reflections for each boundary needs to be grouped across the five RIRs. It is proposed that these reflections can be grouped by exploiting the rank property of an Euclidian Distance Matrix (EDM) - the $N \times N$ matrices containing the squared distances between a set of N points in space [104]. That is, if the rank of the EDM $\widetilde{\mathbf{D}}$ is greater

than or equal to the number of microphones M, then the reflections are all produced by the same boundary, and consequently the same image-source. The EDM matrix $\widetilde{\mathbf{D}}$ is, therefore, defined from the squared distances between microphones and the squared distance travelled by a reflection as,

$$\widetilde{\mathbf{D}} = \begin{bmatrix} ||\mathbf{m}_1 - \mathbf{m}_1||^2 & ||\mathbf{m}_1 - \mathbf{m}_2||^2 & \cdots & ||\mathbf{m}_1 - \mathbf{m}_M||^2 & (\tau_{\mathbf{m}_1} * c)^2 \\ ||\mathbf{m}_2 - \mathbf{m}_1||^2 & ||\mathbf{m}_2 - \mathbf{m}_2||^2 & \cdots & ||\mathbf{m}_2 - \mathbf{m}_M||^2 & (\tau_{\mathbf{m}_2} * c)^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ ||\mathbf{m}_5 - \mathbf{m}_1||^2 & ||\mathbf{m}_5 - \mathbf{m}_2||^2 & \cdots & ||\mathbf{m}_M - \mathbf{m}_M||^2 & (\tau_{\mathbf{m}_M} * c)^2 \\ (\tau_{\mathbf{m}_1} * c)^2 & (\tau_{\mathbf{m}_2} * c)^2 & \cdots & (\tau_{\mathbf{m}_M} * c)^2 & 0 \end{bmatrix} \tag{4.3}$$

where $\tau_{\mathbf{m}_1}$ is the ToA of the reflection at microphone $\mathbf{m}_1$ and c is the speed of sound. To account for noise interference on the ToA estimation *multidimensional scaling* [105] is used, this process finds the closest EDM with a rank of M, and therefore, defines the likelihood of a set of reflections belonging to the same image-source.

Once the reflections have been grouped the location of the image-source that produced those reflections are computed by finding the common point of intersection for a set of spheres, centred around each receiver, with the radius defined by the ToA of the reflections in the group. The candidate boundary locations are then defined by estimating each boundary's normal $\widetilde{\mathbf{n}}$ and a point on each boundary $\mathbf{b}$, which, for the i^{th} image-source $\widetilde{\mathbf{s}}_\mathbf{i}$ and the source position $\mathbf{s}$ [6] as (4.1).

Results show that, for the three convex cases presented, first-order reflections can be used to define the geometry of the room uniquely, with a maximum distance error between parallel boundaries of 7 cm. However, it is important to note that the reflections used in this case were manually detected. Therefore, the impact that noise or false-positive detections, as discussed in Chapter 3, has on the accuracy of the geometry inference process cannot be determined. This could have implications when considering real-world applications of this method, where reflections' ToA are systematically extracted from the RIR. Another key drawback to this method, is the limitation imposed by the microphone array used, that is, each microphone must be carefully positioned such that it clearly receives a first-order reflection from each boundary. This requirement limits it applications to that of convex cases where all receivers are inherently in line-of-sight of every boundary.

4.2.2 Room of Best Fit

Arteaga *et al.* [106] approached the problem of geometry inference for a cuboid-shaped room by attempting to find a 'room of best fit'. This method uses the reverb time and source-receiver distance, as estimated from a single RIR, to define a series of possible cuboid shaped rooms [106]. The candidate rooms are constrained such that,

$$\mathbf{l}_z \leq \mathbf{l}_y \leq \mathbf{l}_x \tag{4.4}$$

$$0 \leq \mathbf{s}_{[x,y,z]} \leq \mathbf{l}_{[x,y,z]}/2 \tag{4.5}$$

$$\mathbf{s}_{[x,y,z]} \leq \mathbf{m}_{[x,y,z]} \leq \mathbf{l}_{[x,y,z]} - \mathbf{s}_{[x,y,z]} \tag{4.6}$$

where $\mathbf{s}$ is the possible source location, $\mathbf{m}$ is the possible receiver location, and $\mathbf{l}$ are the dimensions on x, y and z axes. The reverb time, T_{60}, is then used to find combinations of possible room parameters that satisfy Sabine's equation [18, 106], as,

$$T_{60} = \frac{24V \ln 10}{cS\ln(1-\alpha)} \tag{4.7}$$

were V is the volume of the room, α is the absorption coefficient, c is the speed of sound, and S the summed surface area of all walls.

To generate a set of candidate rooms that satisfy the above constraints, stochastic search algorithms such as, *simulated annealing* [107] or *genetic algorithms* [108] are used. These search algorithms are performed over a defined period of time, producing a set of possible candidate rooms [106]. For each candidate cuboid-shaped room, the image-source model is used to generate a test RIR. The candidate room that maximises a utility function, defined using the correlation between the simulated and measured RIR, is assumed to be the one that best matches the geometry of the target room.

Results show that some dimensions were correctly estimated, but other dimensions had relative error values of up to 30%, however, the dimensions of the test rooms are not defined, and as such the error values that the 30% relate to cannot be derived. As this approach does not require any form of reflection detection or associated boundary localisation, additional boundaries are not

produced as a result of false-positive detections or misidentification of higher-order reflections as being first-order, which is important when considering geometry inference. Furthermore, expanding this algorithm to consider more complex environments (convex or not) would likely result in highly inaccurate results, as without constraining the room's shape based on *a priori* knowledge or assumptions about the room's shape, it would be impossible to estimate the shape of a given room using this method.

4.2.3 Synthetic Reflection Fitting

The technique proposed by Ribeiro *et al.* [109] uses SRIRs captured using a conference call device consisting of a uniform circular array of microphones, positioned on-top of a table. Similarly to the work of Arteaga *et al.* [106] the image-source model is used to find the most likely candidate boundary locations, however, here the image-source model is used to generate a set of synthetic reflections with known DoA. The boundary locations are then inferred by fitting these synthetic reflections to the measured SRIR, using least-squares optimisation [106]. The synthetic reflections that best match the reflections present in the measured SRIR then define the distance and angular position of the boundary relative to the receiver. The geometry inference process is constrained such that the resulting room must be cuboid, i.e. all non-parallel boundaries must be at a 90° angle to each other. To further validate candidate boundaries, and remove false-positive detections, each candidate boundary must have at least least one second- or third-order reflection attributable to it.

Results presented for real-world testing showed that out of the five boundaries considered, being the four walls and the ceiling, only three were detectable. Results for boundary location are given in terms of azimuth, elevation, and distance of the centre of the boundary relative to the receiver. For the three detected boundaries, a maximum distance error of ± 2 cm and azimuth position error of $\pm 2^\circ$ was reported. While two boundaries are not detected, and the floor was not considered, the boundaries that are inferred have comparable localisation error to the other techniques presented in this chapter. However, the constraints imposed limit this method's application to cuboid-shaped rooms only. Furthermore, the reflection fitting process used would require a relatively large number of synthetic reflections to account for all possible boundary locations.

4.2.4 Maximum Likelihood Image-Source Estimation

Tervo *et al.* [110] proposed a geometry inference method using maximum-likelihood to estimate image-source locations from RIRs measured at six microphone positions. The RIRs were

measured using a single loudspeaker over multiple orientations rotated in 10° steps from 0° to 360°. The generated RIRs are windowed into short 1.5 ms frames with a 95% overlap, and the loudspeaker direction with the largest absolute pressure on average in the microphone array represents the RIR for each time-frame [14], ideally producing a sparser RIR with less densely distributed reflections. From a set of candidate reflections detected in this new RIR, the location of image-sources $\widetilde{\mathbf{s}}$ are estimated using maximum-likelihood as,

$$\widetilde{\mathbf{s}}(x_k|\hat{\mathbf{t}}, \Sigma_k) = \frac{\exp(-\frac{1}{2}[\hat{\mathbf{t}}_k - \mathbf{t}_k(x)]^T \Sigma_k^{-1} [\hat{\mathbf{t}}_k - \mathbf{t}_k(x)])}{(2\pi)^{N/2}\sqrt{\det(\Sigma_k)}} \tag{4.8}$$

where N is the number of microphones, $\hat{\mathbf{t}}_k$ is a vector of the ToAs for the k^{th} reflection in each measured RIR, $\mathbf{t}_k$ are the true ToAs, and $\Sigma_k = \text{diag}(\sigma_{k,1}^2, \sigma_{k,2}^2, \cdots, \sigma_{k,N}^2)$ is the ToA error covariance matrix[110]. The solution to (4.8) is found through an optimisation process using the *Levenberg-Marquardt-algorithm* [111]. The geometry is then inferred from the estimated image-sources, using (4.1) as in [6]. This method assumed that all reflections are first-order, unless a valid reflection path can be found between existing boundary locations. To validate non-first-order reflection paths the Mahalanobis distance [112] is used. Any reflection path with a large Mahalanobis distance (threshold not defined in [110]) is considered invalid. The Mahalanobis distance between two points, x and y, is, from [110], computed as,

$$D^2 = (x - y)^T (\Sigma_x + \Sigma_y)^{-1} (x - y) \tag{4.9}$$

where Σ_x and Σ_y are the covariances for the boundary at point x and y respectively [110].

The proposed method was tested for four different source-receiver pairs in a cuboid shaped room. Using the expected and estimated boundary parameters provides an estimate of distance error and dihedral angle between the expected and estimated boundaries can be made, showing any errors in boundary position. The dihedral angle represent the angle between two boundaries, and is computed from the boundaries normal vectors $\mathbf{n1}$ and $\mathbf{n2}$ as $\cos(\theta) = (\mathbf{n1}_x\mathbf{n2}_x + \mathbf{n1}_y\mathbf{n2}_y + \mathbf{n1}_z\mathbf{n2}_z)/(\sqrt{\mathbf{n1}_x^2 + \mathbf{n1}_y^2 + \mathbf{n1}_z^2}\sqrt{\mathbf{n2}_x^2 + \mathbf{n2}_y^2 + \mathbf{n2}_z^2})$ [113]. The maximum distance error is 67 cm across the four tests, with an RMS distance error of 20.46 cm, and the maximum dihedral angle is 1°, with a RMS dihedral angle of 0.98°. While the dihedral angle, and therefore inferred shape of the room, is comparable or better than other works

presented in this section, the distance error is generally larger. Furthermore, for one of the test cases presented an additional angled boundary is inferred, as a result of an incorrectly inferred boundary from a ceiling reflection. This method is, however, only applicable to convex-shaped rooms as the process used to define possible reflection paths assumes a convex-shaped room, bounded by a limited number of dominant boundaries, as a starting point.

4.2.5 Image-Source Direction and Ranging-Loudspeaker-Image Bisection

The Image-Source Direction and Ranging-Loudspeaker-Image Bisection (ISDAR-LIB) method was proposed by Remaggi *et al.* [8] as an extension of work from [6] and [110]. A forty-eight bi-circular microphone array is used to obtain the SRIRs measured from multiple loudspeaker positions. For the test cases presented, four, nine, twelve, and twenty-two loudspeaker positions are used. Using ToA computed using the C-DYPSA (Section 3.3.1.4) and the DoA computed using the MUSIC algorithm, the location of image-sources can be defined as,

$$\widetilde{\mathbf{s}}_i = d_i \begin{bmatrix} \cos(\theta_i)\cos(\phi_i) \\ \sin(\theta_i)\cos(\phi_i) \\ \sin(\phi_i) \end{bmatrix}^T \tag{4.10}$$

where d_i is the distance travelled by the i^{th} reflection and θ_i and ϕ_i are the azimuth and elevation DoA respectively. The boundary locations are then inferred from the image-source locations using (4.1), and the boundary normal and point are averaged across these RIR measurements.

Results presented show good localisation of reflectors within an enclosed space with a minimum averaged RMS distance error of 20.8 cm and a maximum of 35.2 cm across four test scenarios. Comparisons between [6, 110] and ISDAR-LIB are presented in [8], which showed that on average ISDAR-LIB outperformed the other two methods, with ISDAR-LIB having an average distance error 24.5 ± 0.4 cm, the method in [110] 33.4 ± 0.6 cm, and the method in [6] 26.7 ± 1 cm. While it is stated that no assumption of room shape is made, the results presented only consider cuboid-shaped rooms, and as such it is not possible to tell if the methodology employed is applicable to complex-shaped rooms.

4.3 Direct Localisation

Unlike image-source reversion techniques, direct localisation techniques use the ToA of reflections to estimate boundary locations without resorting to reflection path calculations using the image-source model [8]. There are three approaches to direct localisation discussed in this section, ellipsoid based [8, 9, 114–117], inverse wave field extrapolation [118], and resonant frequency distance estimation [119]. All of these methods use some direct mathematical expression that relates the features or reflections of a RIR to the room's boundaries.

4.3.1 Elliptical Constraint Method

The elliptical constraint method for geometry inference, proposed by Antonacci et al. in [114, 115], was developed to estimate the locations of boundaries in two-dimensions, requiring *a priori* knowledge of the number of boundaries present, and was the first method to use ellipsoids for geometry inference. A peak detection algorithm is used to estimate the ToA of the $N+1$ most prominent reflections within a RIR where N is the number of boundaries in the measurement environment.

The proposed method uses ToA for individual reflections to create ellipses with the source and receiver positions as their foci. Under the assumption that all reflections are specular, each ellipse represents all the possible reflection paths that could define a given reflection. Therefore, the inferred boundary must be tangential to all ellipses defined by a reflection common across multiple RIRs (a minimum of three is suggested in [115]), as seen in Figure 4.2, [115]. The source and receiver positions used have to be carefully considered to ensure that each boundary has a detectable reflection in each of the measured RIR.

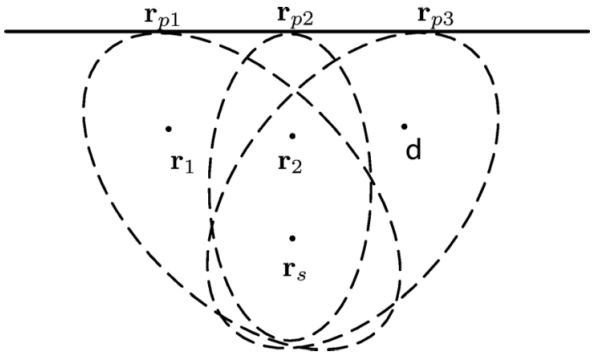

Figure 4.2: Common tangent ($\mathbf{r}_{p1}$, $\mathbf{r}_{p2}$ and $\mathbf{r}_{p3}$) for the ellipses traced for 3 different receiver positions ($\mathbf{r}_1$, $\mathbf{r}_2$ and $\mathbf{r}_3$) with the source at position $\mathbf{r}_s$, from [115]

Results are presented for four scenarios using 4, 8, 12, and 16 different source positions for

the same room. Across the four scenarios, the boundary estimation error values are between 0.5–2.95 cm, with more measurement positions producing smaller boundary estimation errors. However, this method is only applicable to geometry inference in two-dimensional space, and as a result of requiring multiple reflections from the same boundary being uniquely detectable across an array of microphones, it is only applicable to simple convex-shaped rooms. Furthermore, the constraint of using only the $N + 1$ most prominent reflections in the RIR implies *a priori* knowledge of the room's shape, and may not hold when considering larger or more complex rooms..

4.3.2 3D Elliptical Constraint Method

Nastasia *et al.* [116] proposed an extension to the work presented in [114, 115] for 3D room geometry estimation. In the proposed method the ToA of only the first prominent peak after the direct sound is considered. Therefore, a set of RIR measurements is required for each wall with four receiver locations per set (24 RIRs in total for a cuboid room). In this study ellipsoid parameters are defined as a quadric matrix,

$$\mathbf{O} = \begin{bmatrix} a_{n,m} & n_{n,m} & d_{n,m} & g_{n,m} \\ b_{n,m}/2 & c_{n,m} & e_{n,m}/2 & h_{n,m}/2 \\ d_{n,m}/2 & e_{n,m}/2 & f_{n,m} & i_{n,m}/2 \\ g_{n,m}/2 & h_{n,m}/2 & i_{n,m}/2 & l_{n,m} \end{bmatrix} \tag{4.11}$$

where the quadric parameters are computed as,

$$a_{n,m} = 4[(x_m - x_n)^2 - d^2] \tag{4.12a}$$

$$b_{n,m} = 8[(x_m - x_n)(y_m - y_n)] \tag{4.12b}$$

$$c_{n,m} = 4[(y_m - y_n)^2 - d^2] \tag{4.12c}$$

$$d_{n,m} = 8[(x_m - x_n)(z_m - z_n)] \tag{4.12d}$$

$$e_{n,m} = 8[(y_m - y_n)(z_m - z_n)] \tag{4.12e}$$

$$f_{n,m} = 4[(z_m - z_n)^2 - d^2] \tag{4.12f}$$

$$g_{n,m} = 4[T^2(x_n + x_m) - (x_m - x_n)(x_m^2 - x_n^2 + y_m^2 + z_m^2 - y_n^2 - z_n^2)] \tag{4.12g}$$

$$h_{n,m} = 4[T^2(y_n + y_m) - (x_m - x_n)(x_m^2 - x_n^2 + x_n^2 + z_n^2 - x_n^2 - x_n^2)] \tag{4.12h}$$

$$i_{n,m} = 4[(z_n + z_m) - (x_m - x_n)(z_m^2 - z_n^2 + y_m^2 + x_m^2 - x_n^2 - y_n^2)] \tag{4.12i}$$

$$l_{n,m} = [(x_m^2 + y_m^2 + z_m^2) + (x_n^2 + y_n^2 + z_n^2 - d^2)]^2 - 4(x_m^2 + y_m^2 + z_m^2)(x_n^2 + y_n^2 + z_n^2) \tag{4.12j}$$

where d is the distance travelled by the reflection, x_m, y_m, and z_m are the Cartesian coordinates for the source, and x_n, y_n, and z_n are the Cartesian coordinates for the receiver [116]. A boundary is then defined as being tangential to the ellipsoid if,

$$\mathbf{B}^T \mathbf{O}_{n,m} \mathbf{B} = 0 \tag{4.13}$$

where $\mathbf{B} = \begin{bmatrix} B_x & B_y & B_z & 1 \end{bmatrix}$ are the boundary parameters and $\mathbf{O}_{n,m}$ is the ellipsoid quadric matrix for microphones $1 \leq n \leq N$ and sources $1 \leq m \leq M$. The boundary location is inferred as the boundary that is tangential to all ellipsoids defined by the first reflection within the four RIR that were measured specifically for the desired boundary. A cost function, as defined in [116], is used to find the inferred boundary $\widetilde{\mathbf{B}}$ that minimises (4.13) as,

$$\widetilde{\mathbf{B}} = \arg\min_{B} \left(\sum_{m=1}^{M} \sum_{n=1}^{N} ||\mathbf{B}^T \mathbf{O}_{n,m} \mathbf{B}||^2 \right) \tag{4.14}$$

To remove erroneously detected boundaries, the room dimensions are constrained such that there is a minimum and maximum coordinate on the x, y, and z axes, which in this study was a minimum of 0 m and a maximum of 5.5 m [116]. Results presented show good localisation of boundaries with the maximum error between inferred and measured boundaries being 7 cm, and a maximum angle between boundaries of 4.5° [116] - which is larger than the other methods presented in this chapter where the angular error is reported. It is important to note that this method required *a priori* knowledge of the largest dimension of the room, and as with the work presented in [115] the proposed method is only applicable to simple convex-shaped rooms as a result of requiring multiple reflections from the same boundary being uniquely detectable across an array of microphones, which is not necessarily achievable for non-convex rooms.

4.3.3 Ellipsoid based 3D Geometry Inference using a Combination of Linear Estimates

Filos *et al.* [117] proposed an alternative approach to ellipsoid based 3D geometry inference, by splitting the problem into three 2D estimates. A seven-microphone array is used, which can be divided into three sub-arrays consisting of five of the seven microphones. These sub-arrays are located such that they lie on the xy-plane, xz-plane, and yz-plane, and, therefore, are used to estimate the 2D boundary locations on their defined plane. As with [116], only the first reflection in a RIR is used, and so, six loudspeaker positions are required producing a total of 30 RIRs. As with [116] the ellipsoids are defined using a quadric matrix as,

$$\mathbf{O} = \begin{bmatrix} a & b & d \\ b & c & e \\ d & e & f \end{bmatrix} = \widehat{\mathbf{T}}^{-T}\widehat{\mathbf{R}}^{-T}\widehat{\mathbf{S}}^{-T} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \widehat{\mathbf{T}}^{-1}\widehat{\mathbf{R}}^{-1}\widehat{\mathbf{S}}^{-1} \tag{4.15}$$

where $\widehat{\mathbf{T}}$, $\widehat{\mathbf{R}}$, $\widehat{\mathbf{S}}$ are translation, rotation, and scaling matrices as defined in [120], The location of the boundary in two-dimensions, as defined by the microphone sub-array, is inferred as the line that is tangential to the set of N ellipses $\mathbf{O}$ by finding the least-squares solution to the cost function,

$$J(\mathbf{l}, \{\mathbf{O}^*_{n,m}\}) = \sum_{n=1}^{N} ||\mathbf{l}^T \mathbf{O}^*_{n,m} \mathbf{l}||^2 \tag{4.16}$$

where $\mathbf{l}$ defines the line parameters and $\mathbf{O}^*_{n,m} = \det(\mathbf{O}_{n,m})\mathbf{O}^{-1}_{n,m}$ is the adjoint of the conic matrix [117]. This results in six reflector lines, one for each boundary, which theoretically should all intersect. However, in practice ToA estimation errors will likely result in non-intersecting lines, therefore in [117] it is proposed that the most likely boundary be found using a cost function solved in the least-square sense as,

$$\widetilde{\mathbf{B}} = [\widetilde{\mathbf{n}}, \widetilde{d}]^T = \underset{\mathbf{n},d}{\operatorname{argmin}} ||\mathbf{G}[\mathbf{n}, d]^T||^2 \tag{4.17}$$

where $\mathbf{n}$ is the boundary normal, d is the distance from the coordinate origin, $\widehat{\mathbf{B}}$ is the estimated boundary, and $\mathbf{G}[\mathbf{n}, d]$ is defined using two points on two non-parallel and non-intersecting lines

$\mathbf{l1}_1$, $\mathbf{l1}_2$,$\mathbf{l2}_1$, and $\mathbf{l2}_2$ as,

$$\mathbf{G}[\mathbf{n}, d] = \begin{bmatrix} \mathbf{l1}_1 & 1 \\ \mathbf{l1}_2 & 1 \\ \mathbf{l2}_1 & 1 \\ \mathbf{l2}_2 & 1 \end{bmatrix} \tag{4.18}$$

From the estimated set of possible boundaries, the room shape is then inferred from the intersection points between adjacent boundaries [117].

Results are presented for a single real-world test case in a cuboid shaped room. Across the six boundaries estimated a minimum boundary distance error of 0.063 cm and a maximum of 7.95 cm is reported, with a minimum boundary angular error of 0.718° and a maximum of 1.601° [117]. However, as with the other ellipse based methods the constraints imposed as a result of the required measurement positions limits it application to that of convex-shaped rooms. Furthermore, as only one set of tests are presented the accuracy with respect to different room sizes or measurement conditions are not considered.

4.3.4 Ellipsoid Tangent Sample Consensus

The Ellipsoid Tangent Sample Consensus (ETSAC) method was proposed by Remaggi *et al.* [8] and as with [115–117], considered the geometry inference problem through the use of ellipsoids, using the C-DYPSA method to estimate the ToA of reflections. As with [116, 117] ellipsoids are defined using a quadric matrix defined as,

$$\mathbf{O} = \begin{bmatrix} a & d & f & g \\ d & b & e & h \\ f & e & c & i \\ g & h & i & j \end{bmatrix} \tag{4.19}$$

To define the ellipsoid parameters the quadric matrix is initialised to define a unit sphere with $a = b = c = 1$, $j = -1$, and all remaining parameters initialised as 0. The desired ellipsoids $\mathbf{O}_{n,m}$ can then be produced through translation $\mathbf{T}_{n,m}$, rotation $\mathbf{R}$ and scaling $\mathbf{S}$ of the unit sphere matrix $\mathbf{O}$ as

$$\mathbf{O}_{n,m} = \widehat{\mathbf{T}}_{n,m}^{-T}\widehat{\mathbf{R}}_{n,m}^{-T}\widehat{\mathbf{S}}_{n,m}^{-T}\mathbf{O}\widehat{\mathbf{S}}_{n,m}^{-1}\widehat{\mathbf{R}}_{n,m}^{-1}\widehat{\mathbf{T}}_{n,m}^{-1} \tag{4.20}$$

The boundary tangential to all ellipsoids is then searched for by randomly selecting sets of points on the ellipsoid with parameters $i = j = 1$, using the relationship from (4.13).

Results presented show that the proposed method outperforms image-source reversion techniques presented in [6–8], with an average error of 22.0 ± 0.8 cm across the four rooms used. As this method uses information across multiple receiver and source locations to generate each wall, it is likely that the higher accuracy of this method is attributable, in part, to the larger numbers of measurement locations, which in [115] was shown to produce higher boundary estimation accuracy. For the four cuboid-shaped rooms tested a 48-microphone bi-circular array was used with, 4 loudspeaker positions for the 1623 m^3 room, 9 for the 43 m^3 room, 12 for the 189 m^3 room, and 22 for the 23 m^3 room. While the results show good accuracy for boundary localisation, the requirement that a first-order reflection from each boundary be present in every source-receiver pair limits its application to that of simple convex-shaped rooms where fewer boundaries are present, in turn resulting in fewer, more sparsely distributed first-order reflections that are easily detectable.

4.3.5 Image-Microphone Reflector Localisation

In [9], four uniform-linear arrays of loudspeakers, one per wall, with a maximum of 78 loudspeaker (minimum of 64) positions were used to produce stacked plot of RIRs (a vertical concatenation of RIRs in an image). Common reflections within the stacked RIRs for one loudspeaker sub-array are detected and grouped using the linear radon transform technique outlined in Section: 3.3.1.3. From the grouped reflections a set of spheres centred on each source location, with radius defined by the corresponding ToA of the reflection, are defined. From these spheres a set of possible points that define an image-microphone (the mirror of the microphone in the boundary) can be detected as the common points of intersection across each sphere, which fall on a circle. Once the possible image-microphone positions have been defined for all reflections across every sub-array, the location of the image-microphone needs to be refined to a single point. This is achieved by searching for common reflections across every sub-array by finding any image-microphone circles that intersect [9].

After defining the most likely groups of common reflections across each sub-array, spheres are generated centred around every source in the array with radius equal to the ToA for the reflection

detected in the RIR corresponding to that source position. Across all these spheres there will now be a single point of intersection, the location of which defines the image-microphone. The boundary position is then estimated from the point of intersection of a line going from image-microphone to source and an ellipse defined with foci on the source and receiver position, with major and minor axes defined by the ToA of the reflections [121].

Results presented show good localisation of the reflective boundaries, with a maximum boundary location error, averaged over all boundaries in a test case, of 9.05 cm and a maximum average angle between desired and inferred boundaries of 3.5° across seven simulated and one real-world measurement case [9]. Furthermore, the simulated data always results in more accurate inference of the room geometry with a maximum difference of 4.84 cm and a minimum of 0.85 cm. While the results show good accuracy, it is at the expense of requiring a large number of measurement positions, and only considers the case of a cuboid-shaped room. Expanding this to consider non-convex rooms would require the use of multiple microphone positions and more complex shaped loudspeaker arrays, which are not always feasible.

4.3.6 Acoustic Imaging

Kuster *et al.* [118] approached the problem of geometry inference through inverse wave field extrapolation using the Kirchhoff-Helmholtz integral, requiring *a priori* knowledge of the general shape of the room. Inverse wave field extrapolation solves the Kirchhoff-Helmholtz integral by searching for a point I away from the receiver array that satisfies the ray direction and ToA for the reflections as,

$$\langle p_{Im}(\mathbf{r}_i)\rangle = \int\int dr_{R_x} dr_{R_z} [w_{1I}(\mathbf{r}_I, \mathbf{m}, t) v_n(\mathbf{m}, t) + w_{2I}(\mathbf{r}_I, \mathbf{m}, t) p(\mathbf{m}, t)]_{t=\tau(\mathbf{s},\mathbf{r}_I,\mathbf{m})} \tag{4.21}$$

where $p(\mathbf{m}, t)$ and $v_n(\mathbf{m}, t)$ are the measured pressure and normal component of the particle velocity extracted from the measured RIR, $p_{Im}(\mathbf{r}_I)$ is the calculated reflected pressure at point I, τ is the time of arrival given the source $\mathbf{s}$, receiver $\mathbf{m}$, and reflection location $\mathbf{r}_I$, and $w_{1I}(\mathbf{r}_I, \mathbf{r}_R, t)$ and $w_{2I}(\mathbf{r}_I, \mathbf{r}_R, t)$ are computed as,

$$w_{1I}(\mathbf{r}_I, \mathbf{m}, t) = p_0 \frac{1}{4\pi ||\mathbf{r}_I - \mathbf{m}||} \frac{\delta}{\delta_t} \tag{4.22a}$$

$$w_{2I}(\mathbf{r}_I, \mathbf{m}, t) = \frac{\cos(\phi)}{4\pi||\mathbf{r}_I - \mathbf{m}||}\left(\frac{1}{||\mathbf{r}_I - \mathbf{m}||} - \frac{1}{c}\frac{\delta}{\delta_t}\right) \tag{4.22b}$$

where c is the speed of sound and ϕ is the angle between the microphone array normal vector and a ray with end points at the point of reflection and the centre of the receiver array.

The key limitation with this approach is the assumption that all reflections are first-order. To this extent, Kuster *et al.* proposed that through knowledge of the original shape of the room, higher-order reflections that are assumed to be first-order can be manually removed - as the boundary they produce will fall outside of the desired room's geometry.

Results presented for 2D room geometry inference for a simulated shoebox room with 400 receiver locations showed that the four boundaries were located approximately where they should be, with artefacts appearing as a result of reflections from the floor and ceiling, which are still present as the RIRs are measured in a real-world environment. Additional results are presented for the detailed inference of a single wall in 3D. This was achieved through the use of multiple uniform-linear arrays positioned at different heights from the floor. The acoustic image of the wall is produced through concatenation of the 2D wall slices produced by each sub array. While they do not state the number of microphones required, it is likely that a very large number of microphones are used, considering the 400 needed for 2D geometry inference [118]. The results presented for the 3D acoustic image of the wall show that objects such as cabinets within the room were also inferred. However, for both the 2D and 3D test presented, error metrics are not reported. Furthermore, given the large number of measurement needed this method would not be practical for real-world 3D geometry inference, as such large number of measurements require signifficant time-overhead for measurement and analysis.

4.3.7 Reflector Localisation using Room Transfer Functions.

Zamaninezhad *et al.* [119] considered the problem of locating the distance between two reflective boundaries through use of a single room transfer function. To achieve this they initially defined that there is a reflector present at $x = 0$ on the x-axis and that the source is closer to this boundary than the receiver. The distance between two boundaries, $\widetilde{l}$, is defined using the main resonant frequency within the room transfer function, which from [119] are found by optimising

the cost function,

$$f(\lambda) = \int_{\omega} \left| \mathbf{g}_m(\omega) \cdot \left(\frac{\omega}{c}\right) \sin\left(\frac{\omega}{c}\lambda\right) \right| d\omega \tag{4.23a}$$

$$\widetilde{l} = \arg\min_{\lambda} f(\lambda) \tag{4.23b}$$

where $\mathbf{g}_m(\omega)$ is the measured room transfer function at angular frequency ω, λ is the wave length, c is the speed of sound.

The results show the distance between the two boundaries could be estimated with good accuracy, with a distance error of 0.6 cm. While the results are comparable or better that others presented in this chapter, it is at the cost of simplifying the problem to estimating the location of one boundary relative to another parallel one. This approach would become significantly difficult when analysing more complex enclosed spaces, requiring multiple measurement positions, and *a priori* knowledge of the locations of half the boundaries that define the enclosed space.

4.4 Summary

In this chapter geometry inference methods previously presented in the literature have been discussed. These methods make use of temporal/spatiotemporal information from RIRs measured at multiple measurement positions, to locate reflective boundaries within the environment - walls, floor, and ceiling. These methods have been assigned to one of two sub-categories [8], image-source reversion and direct localisation. Image-source reversion uses an inversion of the image-source model, to estimate the boundary locations from image-sources inferred from arriving reflections. Direct localisation techniques locate boundaries without requiring inference of possible reflection paths for each reflection. Generally the methods proposed for each category perform comparably when inferring the geometry of a cuboid shaped room. However, the types of microphone/loudspeaker arrays that these methods are designed for, assumptions on the number of boundaries, assumptions made when retracing reflection paths, and/or the requirement that a first-order reflection from every boundary is attributable to and identifiable at every measurement location, constrains these methods to simple convex-shaped rooms only - especially in the case of ellipsoid based methods.

This thesis will define a geometry inference method for convex and more complex non-convex rooms, using a compact spherical microphone array, with a sufficient number of measurement

positions to ensure each boundary has a first-order reflection attributable to and identifiable in at least one measurement. The proposed method will relax constraints on room shape and reduce the number of measurement positions needed, while relaxing constraints on source and receiver positioning, through the use of a commonly used microphone array. While these relaxed constraints will increase the number of room shapes that geometry inference is applicable too, the proposed method will not accurately infer room geometry for the case of rooms with vertically angled walls or ceilings; such as churches. These constraints have been imposed to improve the method's robustness to false-positive detections, which could lead to inaccurate estimates of the room's shape. From the previous work outlined in this chapter the following constraints, similar to those previously presented, will be implemented as part of the proposed geometry inference algorithm:

- The relative position of all source and receivers are known.
- It is assumed that the source-to-receiver distance is known *a priori* to account for any measurement system latency.
- Knowledge or room temperature to allow estimation of the speed-of-sound.
- It is assumed that the walls are perpendicular to the floor and ceiling, and the floor and ceiling are parallel to each other.
- That all reflections have a dominant specular component allowing their reflection paths to be traced.
- Each boundary has at least one first-order reflection assignable to and detectable in at least one SRIR.
- In this study an empirically defined minimum source/receiver-boundary distance of 50 cm is used (half that of the minimum recommended distance of 1 m in [122] to allow for analysis of smaller/complex rooms). This constraint is imposed to ideally improve the robustness of the method to to false-positive detections, where boundaries inaccurately inferred close to the source or receiver can lead to desired boundaries being invalidated by the proposed boundary validation process.
- The inferred boundaries define a closed geometry.

Part III

Original Research

Chapter 5

Direction of Arrival Analysis for Reflections in Binaural Room Impulse Responses

5.1 Introduction

In the previous chapter, relevant methods presented in the literature for geometry inference were discussed. These previous methods considered the case of convex-shaped rooms only, and were often restricted further to, and only tested with, cuboid-shaped rooms. This is as a consequence of every boundary in the room requiring a first-order reflection attributable to said boundary and identifiable across all, or some subset of, Room Impulse Response (RIR) measurements obtained from different points in the space, and assumptions made about the number of boundaries that define the room. The work in this thesis, therefore, proposes that by using a compact-microphone array capable of representing both time- and direction-of-arrival, a geometry inference method applicable to both convex and non-convex cases might be developed. Therefore, this chapter will explore whether direction-of-arrival (DoA) can be accurately estimated for reflections in a Spatial Room Impulse Response (SRIR) measured with a two-microphone binaural dummy head - the microphone array with the fewest microphones that can encode three-dimensional spatial information - referred to as a Binaural Room Impulse Response (BRIR).

Binaural localisation has been investigated throughout the literature from the perspective of continuous signals (mainly speech), however, only one paper to this author's knowledge [64] has

considered the problem of localisation using reflection information only. The current state-of-the-art approach to binaural DoA estimation is through the use of a binaural model fronted Neural Network (NN). Therefore, this chapter will explore an implementation of this approach for the DoA estimation of reflections in a BRIR. The aim of this chapter is to establish whether a two channel binaural approach can provide accurate enough estimation of DoA for the purpose of geometry inference - where any inaccuracies in the estimated reflection parameters will consequently result in inaccurate estimations of boundary locations.

This chapter is presented as follows: Section 5.2 will discuss the problem domain for binaural DoA estimation, Section 5.3 will describe the binaural model and NN architecture used, Section 5.4 will explain the testing procedure to assess the generalisability of the NN, Section 5.6 will present the results, Section 5.7 will discuss the results in the context of the literature, and Section 5.8 will conclude the chapter.

5.2 Consideration of the Problem Domain

A binaural signal is characterised by the receiver having the properties of a typical human head, that is, two channels of information separated appropriately, and subject to spatially-dependant spectral and temporal variations imparted by the pinnae and head. The spatial information contained within a binaural signal is encoded as the level and time-of-arrival differences (ILD and ITD respectively) between the signals arriving at each ear, which are a function of both frequency and source position relative to the head [22]. Furthermore, both of these cues will vary between different people/dummy heads as a result of differences in ear and head morphology, as such, the binaural DoA estimator ideally needs to be generalisable to different measurement setups. This is to ensure that the resulting trained NN can be implemented outside of the work presented in this study, where different measurement equipment, and conditions, may be in use.

The aim of this research is, therefore, to produce a system inspired by the human auditory system, that can estimate the DoA of sound arriving at the receiver using these cues. The problem of binaural DoA estimation is therefore twofold: first the interaural cues must be extracted from the measured signals, and then from the interaural cues estimate the DoA.

5.3 Method

As with the previous studies discussed in Chapter 3, the method developed here uses a binaural model to produce representations of the frequency-dependent ITD and ILD from the signals

arriving at each ear of a binaural dummy head microphone. These cues alone have been shown to be insufficient [32, 76, 86] to provide accurate localisation of a sound source, due to interaural cue similarities observed at mirrored source positions in the front/rear hemispheres. Therefore, an additional set of binaural cues is generated for the corresponding direct sound and reflected components of a BRIR with the dummy head having been rotated. The use of head rotation has a biological precedence, in that humans use head rotation to focus on the source location. Therefore, the head rotation provides the NN with a second set of features that the NN can use to estimate the DoA of the arriving direct sound or reflection. These sets of interaural cues are then interpreted by a cascade-forward NN, producing a prediction of the DoA for the direct sound and each detected reflection in the BRIR.

To train the NN, a feature matrix is generated using the un-compensated 'raw' SADIE KEMAR dataset [123]. This dataset contains an HRIR grid of 1550 points: 5° increments across the azimuth in steps of 10° elevation, with additional measurement positions based on loudspeaker positions used in ambisonics. To train the NN, only the HRIRs relating to 0° elevation are used, to initially test the accuracy of the approach on a simpler problem domain, providing a dataset of 104 HRIRs. The HRIR dataset alone has been shown to not be sufficient to produce a generalisable NN that produces comparable results across different measurement scenarios [32, 84, 85]. Therefore, a dataset of HRIRs with different simulated measurement conditions, a multi-conditional training (MCT) dataset, is produced by generating additional versions of each HRIRs with simulated uncorrelated diffues noise added to produce SNRs mixtures of 0 dB, 10 dB, and 20 dB as used in [32]. The simulated uncorrelated diffuse noise is generated by convolving Gaussian white noise with all 1550 HRIRs in the SADIE KEMAR dataset and averaging this localised noise across the 1550 positions; producing a simulated uncorrelated diffuse noise matrix [32]. The MCT dataset is then generated from the feature vectors of the original HRIRs and the HRIRs with added diffuse noise. It is important to note that the NNs used are only ever trained with these HRIRs variations, no reflections are incorporated as part of the training data.

5.3.1 Binaural Model

The binaural model used here is inspired by the work presented in [124, 125], representing frequency-dependent temporal information as an Interaural Cross-Correlation (IACC) function, and frequency-dependent level difference as the ratio of the signal energy between the ears across frequency bands. Both the temporal and spectral feature spaces used provide directionally-dependent cues, produced by the path differences between ears and acoustic shadowing formed by the presence of the head, which allow the human auditory system to localise a sound source

in an environment [21, 126]. These directionally-dependent feature spaces are used in this study to produce a feature vector that can be analysed by a NN to estimate DoA. The aim of this binaural model is to process the binaural signals in a manner that is similar to that of the human auditory system, and as such is split into three processing stages, filtering of the audio into frequency bands, processing of the filtered binaural signals to produce a representation of the human auditory system's nerve firing rates, and then computation of the interaural cues. An overview of this process, as will be described in this section, can be seen in Figure 5.1.

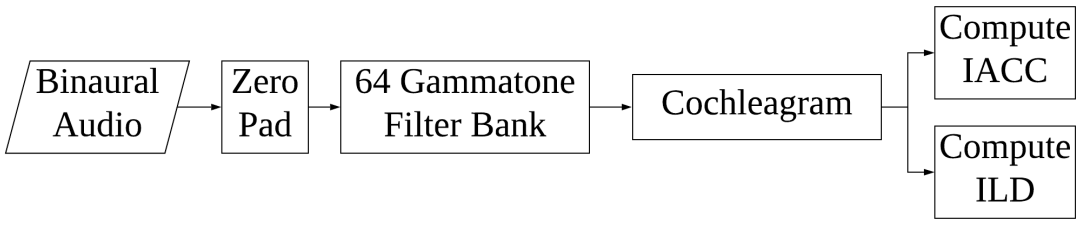

Figure 5.1: Processing diagram for the binaural model starting with the binaural audio, which is zero padded, and filtered using a bank of 64 gammatone filters, the filtered audio is then used to compute the interaural cross correlation and interaural level difference .

Prior to filtering the binaural signals, in this case the direct source or a reflection from a BRIR, the vectors containing the left and right audio channels are zero-padded by 2000 samples to prevent any loss of signal as a result of delay introduced by the gammatone filters used. This ensures that no part of the desired signal is shifted outside of the sample-range represented by the signal vector as a result of this filter delay. The zero-padded signals are then passed through a bank of 64 gammatone filters spaced from 80 Hz to 22 kHz using the equivalent rectangular bandwidth scale. Gammatone filters are chosen as they are designed to mimic the frequency separation and resolution of the human auditory system [87], and are the filters most commonly used in recent work [32, 86]. The gammatone filter implementation in Malcolm Slaney's 'Auditory Toolbox' [127] is used in this study. The output of the cochlea is then approximated using the cochleagram function in [128] with a window size of six samples and an overlap of one sample; this produces an $F \times S$ map of auditory nerve firing rates across time-frequency units (based on findings in [71]), where S is the number of time-frames and F is the number of gammatone filters. The cochleagram is calculated, using [128], as,

$$\widetilde{\mathbf{X}}_l(f, t_f) = \widehat{\mathbf{X}}_l(f, t_f)\widehat{\mathbf{X}}_l(f, t_f)^T \tag{5.1}$$

where $\widetilde{\mathbf{X}}_l(f, t_f)$ is the cochleagram output for the left channel for gammatone filter f at time-

frame t_f, $\widehat{\mathbf{X}}_l(f, t_f)$, which is six samples in length [128]. An example cochleagram output for the left channel of a HRIR, measured with the Knowles' Electronic Manakin for Acoustic Research (KEMAR) dummy head microphone, at azimuth = 90° and elevation = 0°, from the SADIE database [123], can be seen in Figure 5.2, the top image shows the filtered left channel of the HRIR and the bottom image the output of the cochleagram. As can be seen the output of the cochleagram produces a more focussed representation of the HRIR with fewer additional peaks in the signal.

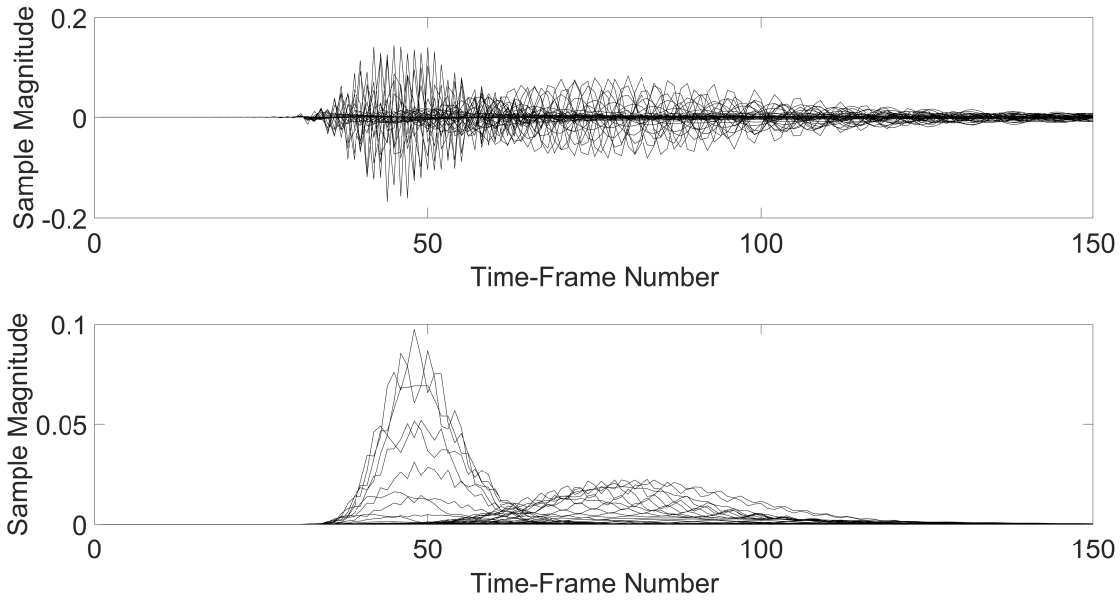

Figure 5.2: Top image shows the left channel of a HRIR after being filtered by the bank of 64 gammatone filters. The bottom image is an example cochleagram output for the left channel of a HRIR measured at azimuth $= 90^\circ$ and elevation $= 0^\circ$. Each of the solid black lines represents a different frequency band.

The IACC function is used to represent the temporal difference between the two channels of audio over time, where the maximum point of correlation between the two channels represents the ITD. It has been shown that the IACC function is directly influenced by the acoustic effect of the head [129], and therefore, features within the IACC function, such as the relationship between the main peak and any side bands, will vary with azimuthal DoA [32]. The IACC function will therefore convey more information about the sound-field arriving at the binaural dummy head than using just the ITD estimate in isolation. The IACC function is computed for each gammatone filter band as the cross-correlation, $\mathbf{C}(f)$, between the approximated cochlea output $\widetilde{\mathbf{X}}_l$ and $\widetilde{\mathbf{X}}_r$ for the left and right channel, respectively, with a maximum lag of ± 1.1 ms as [130],

$$\mathbf{c}(f, t_f) = \int_{\tau=-1.1\ ms}^{\tau=1.1\ ms} \widetilde{\mathbf{X}}_l(f, t_f)\widetilde{\mathbf{X}}_r(f, t_f - \tau)d\tau \tag{5.2}$$

where τ represents the time-delay. The maximum lag of ± 1.1 ms is chosen based on the maximum observed time delays between signals arriving from different DoA as suggested by Pulkki *et al.* in [124]. To produce a more accurate estimate of the IACC function, the cross-correlation function $\mathbf{c}(f, t_f)$ is then normalised, from [124], as,

$$\text{IACC}(f, t_f) = \frac{\mathbf{c}(f, t_f)}{\sqrt{\widetilde{\mathbf{X}}_l(f, t_f)\widetilde{\mathbf{X}}_l(f, t_f)^T\widetilde{\mathbf{X}}_r(f, t_f)\widetilde{\mathbf{X}}_\mathbf{r}(f, t_f)^T}} \tag{5.3}$$

The IACC is then averaged across the 64 gammatone filters, producing the temporal feature space for the analysed signal.

The ILD is calculated from the cochleagram output in decibels as the loudness ratio between the two ears for each gammatone filter f as,

$$\text{ILD}(f) = 10 * log_{10}\left(\frac{\sum_{t_f=1}^{N}\widetilde{\mathbf{X}}_l(f, t_f)}{\sum_{t_f=1}^{N}\widetilde{\mathbf{X}}_r(f, t_f)}\right)\ dB \tag{5.4}$$

where $\widetilde{\mathbf{X}}_l(f, t_f)$ and $\widetilde{\mathbf{X}}_r(f, t_f)$ are the approximated cochlea output of gammatone filter f, for the left (l) and right (r) ear for the time-frame t_f, and N is the total number of time-frames. An example of the IACC and ILD feature vector for a HRIR measured with a KEMAR dummy head at azimuth = $90°$ and elevation = $0°$, from the SADIE database [123], can be seen in Figure 5.3.

When analysing a binaural room impulse response with a sampling rate of 44.1 kHz, the output of this binaural model is a $[1 \times 99]$ IACC function and a $[1 \times 64]$ ILD vector, producing a 163 point feature space for a time-frame. An example MATLAB implementation of the binaural model can be seen in Algorithm 1

5.3.2 Neural Network Data Model

The IACC and ILD computed using the binaural model presented in the previous section defines the feature space for a single HRIR. This defined feature space is, however, still not sufficient for accurately disambiguating between signals arriving from mirrored positions at the front and

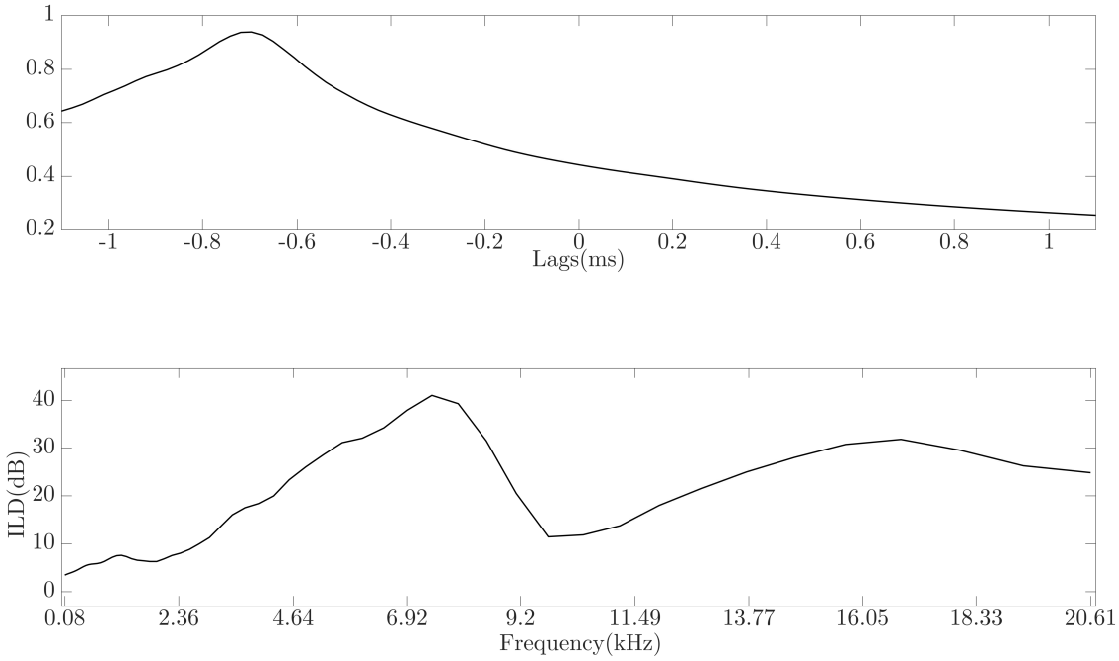

Figure 5.3: Example of the interaural cross-correlation function (top) and interaural level difference (bottom) for a HRIR measured with a KEMAR dummy head microphone with a source positioned at azimuth = 90° and elevation = 0°, from the SADIE database.

back of the head; where the interaural cues will be similar [32]. Therefore, an additional feature space is added to the training data - the interaural cues produced by a HRIR corresponding to either a $+\theta_{\text{rotation}}$ or $-\theta_{\text{rotation}}$ rotation of KEMAR with the same signal-to-noise ratio.

The use of 'head rotation' has a biological precedence, in that humans use head rotation to focus on the location of a sound source, disambiguating front-back confusions that occur due to interaural cue similarities between signals arriving from opposing locations in the front and back hemispheres (the hemisphere regions can be see in Figure 5.4) of the head [21, 126]. In this study, the equivalent effect of implementing a head rotation of θ_{rotation} is realised by taking the BRIR measurements at two additional fixed measurement orientations. The use of fixed rotations reduces the number of additional signals needed to train the NN and reduces the number of additional measurements that need to be recorded.

Two versions of the training matrices are produced, one for the interaural cues of the HRIRs with additional cues at $+\theta_{\text{rotation}}$, and the other for the HRIRs with additional cues at $-\theta_{\text{rotation}}$, producing two 416×326 feature matrices. These training matrices are used to train two NNs, one for each rotation. The NN trained with the $-\theta_{\text{rotation}}$ rotation dataset is used to predict the DoA for signals that originate on the left hemisphere, while the $+\theta_{\text{rotation}}$ NN is used to predict the DoA for signals on the right hemisphere. Each of these NNs are trained with the full azimuth

Algorithm 1: MATLAB implementation of the binaural model. MATLAB functions are indicated in bold, and // indicates a comment.

```
// zero pad the BRIR
BRIR = [BRIR; zeros(2000,2)]
// Apply the F Gammatone Filters the each channel of the BRIR
// Compute the cochleagram output for each channel of the BRIR
maxLag = round(0.0011*Fs);
for f = 1 : F do
    // Compute the normalsiation factor for the IACC
    xAutoCorrelation = sqrt(X̃_l(f) * X̃_l(f)' * X̃_r(f) * X̃_r(f)');
    // Compute the IACC
    [xCorr,lags] = xcorr(X̃_l(f), X̃_r(f), maxLag);
    IACC(f,:) = xCorr ./ xAutoCorrelation;
end
// Average the IACC function over frequency.
IACC = mean(IACC, 1);
// Compute the interaural level difference
ILD = 10 * log10(abs(sum(X̃_l)) ./ abs(sum(X̃_r)));
```

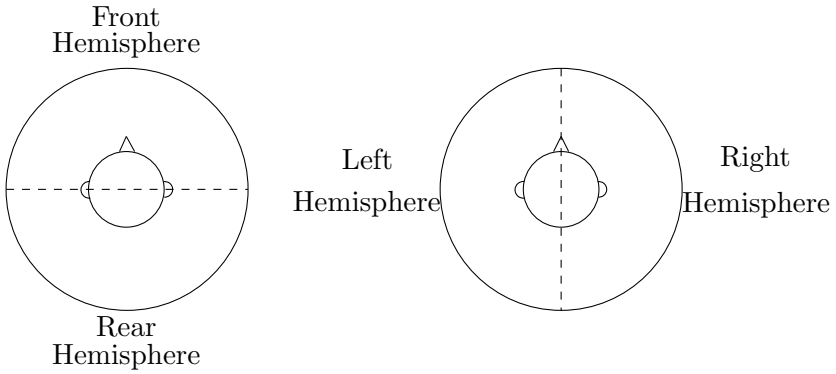

Figure 5.4: Figure showing the regions relating to the front-back hemispheres and the left-right hemispheres.

range to allow the NNs to predict the DoA for signals with ambiguous feature vectors that would otherwise be classified as originating from the opposite hemisphere. As the DoA of the signals is not known *a priori* for test data, the rotation used is chosen based on the location of the maximum peak in the IACC feature vector - if the time index of the peak in the IACC is less than 0 ms (a signal originated in the left hemisphere), a receiver rotation of $-\theta_{\text{rotation}}$ is applied; otherwise, a receiver rotation of $+\theta_{\text{rotation}}$ is used as,

$$\theta_{\text{rotation}} = \begin{cases} +\theta_{\text{rotation}}, & \text{if } \operatorname{argmax}(IACC) > 0\ ms \\ -\theta_{\text{rotation}}, & \text{if } \operatorname{argmax}(IACC) < 0\ ms \end{cases} \tag{5.5}$$

The final step required to generate the training data is to normalise the numeric values of each of the 326 features. This is achieved by *z-normalisation*[131] each of the training data matrices, ensuring each feature has zero mean and unit variance as,

$$\widetilde{\mathbf{x}}_0 = \frac{\left[\text{IACC} \quad \text{ILD} \quad \text{IACC}_{\theta_{\text{rotation}}} \quad \text{ILD}_{\theta_{\text{rotation}}}\right] - \mu}{\sigma} \tag{5.6}$$

where $\widetilde{\mathbf{x}}_0$ is the resulting feature vector, μ is the $[1 \times 326]$ vector of mean values for each feature point from the training dataset, σ is the $[1 \times 326]$ vector of standard-deviation values for each feature point from the training dataset, and $\left[\text{IACC} \quad \text{ILD} \quad \text{IACC}_{\theta_{\text{rotation}}} \quad \text{ILD}_{\theta_{\text{rotation}}}\right]$ is the concatenation of the IACC and ILD feature vectors for the original signal and the signal recorded after head rotation. The mean and standard-deviation used to normalise the training data will then be used to normalise the test data, relating them to each other. An overview of this process can be seen in Figure 5.5.

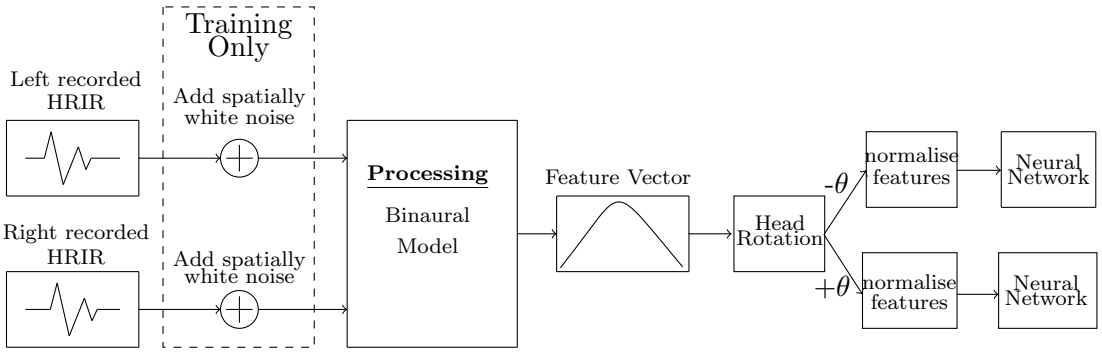

Figure 5.5: Signal processing chain used used to train the neural network. Starting with the HRIRs, in the training phase only simulated diffuse noise is added to create SNR mixtures producing a multi-conditional dataset, the feature vector is then produced for all HRIRs, and the corresponding feature vector for the head rotation is added to the feature vector, these features and then Gaussian normalised and used to train the NN.

5.3.3 Neural Network

To develop, train, and test the NN, the commonly-used, and freely-available, Google python machine learning library TensorFlow [132] is used. The decision to use NNs as opposed to other commonly used machine-learning algorithms for binaural localisation, such as Gaussian Mixture Models (GMMs) or Self-Organising Maps (SOMs), was based on findings in [32, 76] -

which showed that a NN produced more accurate estimates of DoA that GMM [32] and the SOM [76]. Furthermore, instead of the commonly used Multilayer Perceptron (MLP) NN topology, a cascade-forward NN, based on the implementation in [133], approach is taken.

The cascade-forward NN is a highly connected NN architecture that connects both the input feature vector and all previous layers' outputs to the input of each layer [133, 134], an example of which for a one hidden layer cascade-forward NN can be seen in Figure 5.6. As with other NN topologies, the model consists of an input layer which feeds the feature vector into the NN, a number of hidden layers all containing a defined number of neurons that process the input feature vector, and an output layer which defines the output value of the NN based on the output of the hidden layers.

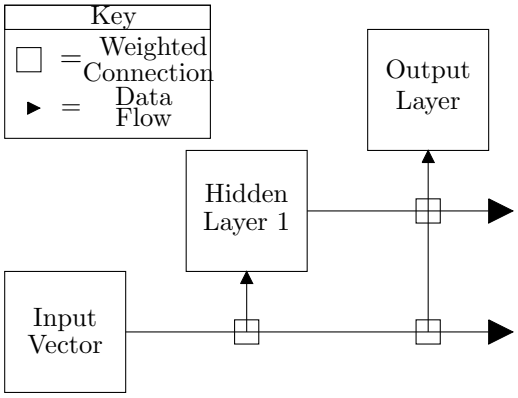

Figure 5.6: Cascade-forward neural network topology used, where squares represent the weighted connections between the hidden layers and the incoming data.

Using the cascade-forward NN topology, each data point, whether it be a feature in the input feature vector or the output of a previous layer, is connected to a neuron via a weighted connection. The summed response of all the weighted connections linked to a neuron defines that neuron's level of activation when presented with a specific data configuration. As with other NNs a bias value is applied to each neuron within the hidden layer. These weights and biases for each layer of the NN are initialised with random values, with the weights distributed such that they are zero mean and have a standard-deviation, σ, defined in [135] as,

$$\sigma_i = q^{-1/2} \tag{5.7}$$

where q is the number of inputs to the i^{th} hidden layer [135]. The output of the I^{th} layer within the cascade-forward NN can therefore be expressed as,

$$\widetilde{\mathbf{x}}_I = \tanh\left(\left(\sum_{i=0}^{I-1} \widetilde{\mathbf{x}}_i \mathbf{W}_{I,i}\right) + \mathbf{b}_I\right) \tag{5.8}$$

where tanh is the hyperbolic tangent function used to define the activation level for each neuron, $\widetilde{\mathbf{x}}_0$ is the [1×Number of features] input feature vector, $\widetilde{\mathbf{x}}_1$ to $\widetilde{\mathbf{x}}_{I-1}$ are the [1×Number of Neurons in layers $i = 1...I-1$] output of the previous $I-1$ layers, $\mathbf{W}_{I,i}$ are the [number of inputs×number of neurons] weights connecting the i^{th} hidden layer to hidden layer I, and $\mathbf{b}_I$ are the neurons' biases for layer I.

When considering DoA estimation in 1° steps, the output layer of the NN will contain 360 neurons, one for each azimuth direction from 0° to 359°. Using 360 output neurons as opposed to 104 defined within the training data will allow the NN to attempt estimation of the DoA for both known and unknown source positions. A softmax activation function is then applied to the output layer of the NN, which turns the activation levels of the neurons into a probability vector that sums to one, defining the likelihood of the analysed signal having arrived from each of the 360 possible DoAs. The DoA is therefore the output neuron with the largest probability value given the feature vector $\widetilde{\mathbf{x}}_0$,

$$\theta_{DoA} = \underset{\theta}{\operatorname{argmax}}\ \mathcal{P}(\theta|\widetilde{\mathbf{x}}_0) \tag{5.9}$$

where $\mathcal{P}(\theta|x)$ represents the probability of azimuth angle θ given the feature vector $\widetilde{\mathbf{x}}_0$, which for a cascade-forward NN is expressed as,

$$\mathcal{P}(\theta|\widetilde{\mathbf{x}}_0) = \text{softmax}\left(\left(\sum_{i=0}^{I-1} \widetilde{\mathbf{x}}_i \mathbf{W}_{I,i}\right) + \mathbf{b}_{out}\right) \tag{5.10}$$

An implementation of the cascade-forward NN using TensorFlow [132] can be seen in Algorithm 2

5.3.4 Training the Neural Network

The training of the cascade-forward NN is performed in stages based on the number of hidden layers being used. Initially only a single hidden layer, and an output layer, is defined, and the NN is trained with the weights and biases of these layers selected using an optimisation algorithm.

Algorithm 2: Pseudocode implementation of the cascade-forward neural network. $\mathbf{x}_0$ is the [1 × number of features] feature vector, $\widetilde{\mathbf{x}}_k$ is the [1 × number of neurons in layer k] output of hidden layer k, $\mathbf{W}_{i,k}$ is the [number of neurons in layer k × number of neurons in i] weigths connecting layer k to layer i, and $\mathbf{b}_i$ are the bias values for hidden layer i.

```
1  i = 1 // Initialise while loop variable
2  while i <= (noLayers) do
3      if i == 1 then
4          x̃_i = tanh((x_0 * W_{i,0}) + b_i)
5      else
6          x̃_i = x_0 * W_{i,0}
7          k = 1 // Initialise while loop variable
8          while k <= i-1 do
               // Add previous layers outputs to the input of this layer.
9              x̃_i += (x̃_k * W_{i,k})
10             k += 1 // Increment while loop variable
11         end
           // add the biases to the weighted sum of the previous layers.
12         x̃_i = tanh(x̃_i + b_i)
13     end
14     i+=1 // increment the while loop variable
15 end
```

The training optimiser used to refine the weights and biases was the Adaptive Moment (ADAM) optimiser [136], which is a computationally efficient first-order gradient optimisation process. The ADAM optimiser computes individual adaptive learning rates for NN parameters based on the estimated gradient of the training cost-function, in this case the cross-entropy cost-function from [137]. The ADAM optimiser parameters are initialised with a learning rate of 0.001, a β_1 value of 0.9, a β_2 value of 0.99 and an ϵ value of 1^{-8}, as used in [136]. The β values define the decay rate of the moving-average of the mean and uncentered variance of the cost-function's gradient and affect the stepsize by which the weights and bias values are adjusted, and ϵ is the numerical stability constant [136]. The ADAM optimiser was chosen over other similar optimisation approaches as it has been shown to be a robust and computational efficient optimisation algorithm that converges on a solution as quick if not quicker than other state-of-the-art optimisation procedures [136].

Once a user defined number of epochs is reached, the accuracy reaches user defined level, or improvement saturates, the weights and biases for this layer are frozen. A new hidden layer is then added and trained, with the aim of minimising the error of the previous layer. This process is repeated until the defined number of hidden layers have been added and trained. In this study a single hidden layer consisting of 128 neurons is used as the defined topology, using more layers resulted in a less accurate NN as a result of over-fitting. The number of neurons is chosen to be the same as the work presented by Ma *et al.* in [32].

The NN is allowed to train for a maximum of 600 epochs, heuristically defined, with the training terminating if the NN reaches 100% accuracy within the training data, or improvement saturates, defined as no improvement over a training period equal to 5% of the total number of epochs. To reduce the likelihood of over-fitting and improve learning efficiency, each epoch is split into four training passes where the NN is given only 25% of the training data in each pass [32]. After each epoch the order of the training data is randomised, so the NN never receives the same batch of data twice.

To define the weight and biases used in testing, fifty NNs, for both the $+\theta_{\text{rotation}}$ and $-\theta_{\text{rotation}}$ NN, are trained until prediction accuracy saturated, which took 122 epochs, achieving generally an accuracy of 95% and a maximum angular error of $\pm 5^{\circ}$ within the training data. The weights and biases for the NN that produced the most predictions within $\pm 5^{\circ}$ of the expected DoA for the KEMAR reflections, for $+\theta_{\text{rotation}}$ and $-\theta_{\text{rotation}}$ NN, are used to define the NN that is used to test the performance of the proposed method.

5.4 Testing

A key measure of the success of a NN is its ability to generalise to new data, where ideally it would produce comparable estimation accuracy for data gathered under different measurement conditions. Therefore, a test dataset of BRIRs are obtained from measurements in an anechoic chamber using both KEMAR 45BC [10] and Neumann KU100 [11] binaural dummy heads. Furthermore, two different loudspeakers were used to measure the BRIRs, the Equator D5 coaxial loudspeaker [138] (the exact loudspeaker used to measure the HRIRs in the SADIE database [123]) and a Genelec 8030 loudspeaker [139]. These provide test cases for the same loudspeaker and dummy head as used in the SADIE KEMAR HRIRs [123], and measurements that use different loudspeakers or dummy heads to the SADIE HRIR used. The source and receiver were positioned 1.5 m off the floor, and the distance between the source and receiver was 1 m. To test the NN's performance at predicting the DoA of reflections, a flat wooden reflective surface mounted on a stand 1.5 m from the receiver at 71° to the front facing dummy head was also placed in the anechoic chamber, such that a reflection with a known DoA would be produced (Figure 5.7). This allows for the accuracy of the NN at predicting the DoA for reflections - without the presence of overlapping reflections that would occur in real-world non-anechoic spaces - to be tested. The speaker stand, reflective boundary stand, and turntable used to rotate the dummy head, were also covered in acoustic foam to minimise any further reflections that might be produced.

To generate the BRIRs the exponential sine sweep method [140] is used with a swept frequency range of 20 Hz to 22 kHz over ten seconds. When performing RIR measurements in real-world environments it is often desirable to have an omnidirectional source [122], to ensure approximately equal acoustic excitation throughout the room. Therefore, to approximate an omnidirectional sound source, the BRIRs are averaged over four speaker rotations (0°, 90°, 180°, and 270°) [141]. The extent to which this averaged loudspeaker response will be omnidirectional will vary across different loudspeakers, particularly at higher frequencies where loudspeakers tend to be more directional. Work presented in [141] showed similar frequency dependent reverb time estimates between a single measurement orientation and multi-orientation, and showed that a greater number of distinct reflections were present in the multi-orientation loudspeaker measurements.

To calculate the required location of the reflective surface such that a known DoA would be produced, a simple MATLAB image-source model based on [17] is used to calculate a point

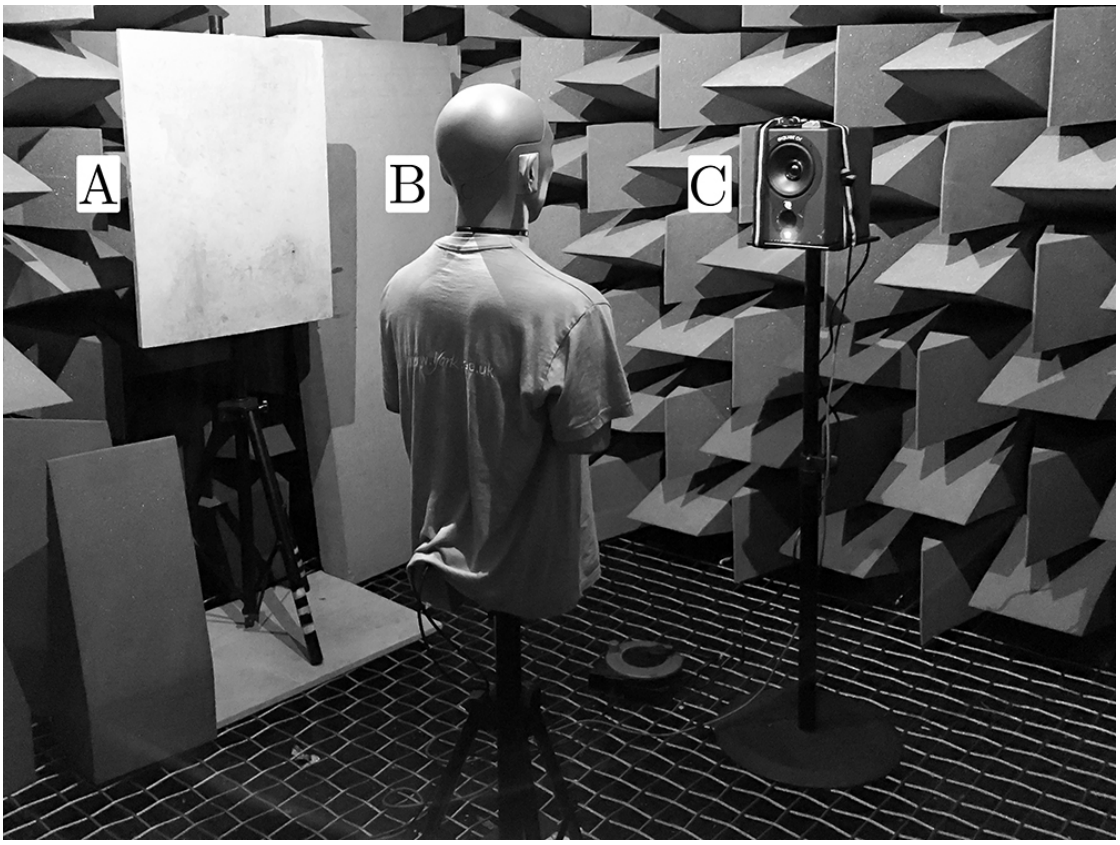

Figure 5.7: Measurement setup showing the reflective surface (**A**), KEMAR 45BC (**B**) and Equator D5 Coaxial Loudspeaker (**C**).

of incidence on a wall that would produce a first order reflection in a 3 m $\times$ 3 m $\times$ 3 m room with the receiver positioned at the centre of the room. The reflective surface is then placed in the anechoic chamber based on the angle of arrival and distance between the receiver and calculated point of incidence. Although care was taken to ensure distances between loudspeaker and receiver, and the position of the boundary relative to the receiver was correct, it is prone to misalignments due to the floating floor in the anechoic chamber, which can lead to possible error in the source, receiver, and boundary placement.

In theory these BRIRs will only have two distinct components, comprising of the direct sound and reflection from the surface, therefore a simple method for separating these signals is employed. Firstly, the maximum absolute peak within binaural signal (whether in the left or right channel) is detected and assumed to belong to the direct sound. A 170 sample (3.9 ms) frame around this peak location, based on observations of the signals, is used to separate the direct sound from both channels of the BRIR. It is also ensured that all segmented audio samples only contained audio pertaining to the direct sound, through observation of the windowed regions of audio. The process is run again to detect the location of the reflected component, and each

segment is again checked to ensure only audio pertaining to the reflected component is present (see Figure 5.8 for an example BRIR with window locations).

Each of the four test scenarios, KEMAR with Equator D5, KEMAR with Genelec 8030, KU100 with Equator D5, and KU100 with Genelec, consisted of 144 BRIRs, with direct sound DoA from $0° \leq \theta \leq 357.5°$ and reflection DoA from $1° \leq \theta \leq 358.5°$ using a turntable to rotate the binaural dummy head in steps of $2.5°$. This provides 288 angles with which to test the NN: 144 direct sound components and 144 reflected components. Therefore, the combined dataset consists of 576 direct sound components and 576 reflected components across all loudspeaker and dummy head microphone combinations.

The separated signals are analysed using the binaural model and the feature matrix is generated by combining the IACC and ILD for the segmented direct or reflected component with the cues for the corresponding component measured at $\pm 90°$ based on the peak location in the IACC function. The positively and negatively rotated test feature vectors are stored in separate matrices, and standardised across each feature in the feature vector, using the mean and standard deviations calculated from the training data. The corresponding test rotation data is fed into the NN trained with the corresponing rotations dataset (as described in Section 5.3.2).

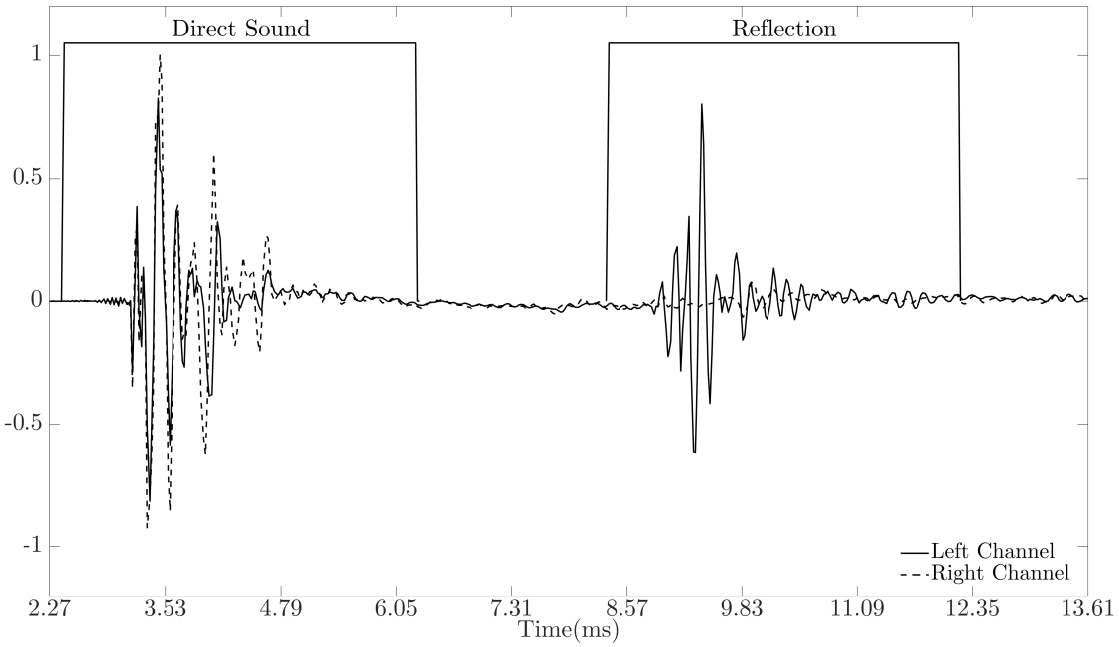

Figure 5.8: Example binaural room impulse response generated with source at azimuth $= 0°$ and reflector at azimuth $= 71°$; the solid line is the left channel of the impulse response; the dotted line is the right channel of the impulse response; and the windowed area denotes the segmented regions using the technique discussed in Section 5.4.

Rotation	Within $\pm5^\circ$	Front-Back Confusions	Max Error
KEMAR Reflections			
no head rotation	17.36%	28.47%	179°
$\pm15^\circ$	29.86%	15.28%	173°
$\pm30^\circ$	34.03%	6.25%	54°
$\pm60^\circ$	29.17%	9.72%	50°
$\pm90^\circ$	32.64%	9.03%	30°

Table 5.1: Direction of arrival accuracy comparison for the reflected component measured with the KEMAR 45BC for different fixed receiver rotation angles.

5.5 Neural Network Parameter Comparisons

To present justification for design choices made in this chapter, this section will present comparisons between: different fixed head rotations, the use of the cochleagram, different feature spaces, the cascade-forward NN and MLP, and number of layers. Fifty versions of the NN are trained using the KEMAR HRIR measurements for elevation = 0° from the SADIE database [123], using the procedures outlined in Section 5.3.3. The NN that produces the most accurate DoA estimations is then used for testing. Test results are presented for the KEMAR reflected components measured with the Equator D5. To define the accuracy of the model the percentage of exact DoA estimations and the percentage of estimates within $\pm5^\circ$ of the expected DoA will be reported. These metrics are based on those previously used in the similar studies presented in Chapter 3.

5.5.1 Head rotation

In Table 5.1 results comparing between fixed receiver rotations of $\pm15^\circ$, $\pm30^\circ$, $\pm60^\circ$, and $\pm90^\circ$ are presented. These results show that, as the angle of rotation increases the maximum error of the DoA estimation decreases. In the context of geometry inference, a lower angular error is desirable, as it would result in angled boundaries being inferred. Furthermore, the results show that the NN produces more accurate results when using head-rotation, with larger maximum errors, larger percentage of front/back confusions, and fewer DoA estimates within $\pm5^\circ$ of the desired DoA when head rotation is not used. Therefore, in this study a θ_{rotation} of $\pm90^\circ$ degrees is used.

5.5.2 Neural Network Comparisons

Comparisons between the cascade-forward NN and the MLP approach used in the previous work are presented in Table 5.2. Both of these NN have the same number of hidden layers, neurons, and are trained using the same procedure. The results show that the cascade-forward

NN converges on a solution 12 s faster than the MLP and has a larger percentage of predictions within $\pm 5^{\circ}$. These findings form the basis by which the decision to use a cascade-forward NN was made.

Neural Network	**Within $\pm 5^{\circ}$**	**Run time**
KEMAR Reflections (Test Data)		
multi-layer perceptron	26.39%	390 Epochs 40 s
cascade-forward	32.64%	244 Epochs 28 s

Table 5.2: Comparison of prediction accuracy for the reflected component measured with the KEMAR 45BC using additional measurements at receiver rotations of $\pm 90^{\circ}$ using a multi-layer perceptron and cascade-forward neural network. Both the multi-layer perceptron and the cascade-forward neural network had one hidden layer with 128 neurons and an output layer with 360 neurons and were trained using the procedure discussed in Section 5.3.3

Further analysis of the cascade-forward NN performance when using one hidden layer with 128 neurons and a two hidden layer with 64 neurons per layer (same number of total neurons), can be seen in Table 5.3. These results show that, when using the training process outlined in this chapter, the use of a single hidden layer is more optimal and results in a more accurate estimation of DoA. The decreased performance for the two hidden layer NN is likely as a result of the NN becoming over fitted to the training data, and as such is less accurate at estimating the DoA for signals that are dissimilar to those used in training.

5.5.3 Binaural Model Comparisons

From the comparisons between different feature spaces presented in Table 5.4, it is clear that using a combination of the IACC and ILD produces the best results, with lower angular error, and a larger number of exact and within $\pm 5^{\circ}$ estimations of DoA. Feature spaces containing the ITD, which was extracted from the maximum peaks in the IACC function across frequency, tend to produce less accurate estimates of DoA. Furthermore, the ILD feature space produces the

Topology	**Exact**	**Within $\pm 5^{\circ}$**	**Max Error**	**Training Accuracy**
KEMAR Reflections				
One Layer 128 Neurons	11.11%	32.64%	30°	98.1%
Two Layer 64 Neurons	4.86%	18.05%	53°	99%
Two Layer 128 Neurons	5.55%	22.22%	63°	99.5%

Table 5.3: Direction of arrival accuracy comparison for the reflected component measured with the KEMAR 45BC for one hidden layer with 128 neurons and two hidden layers with 64 neurons in each. These tests are performed using the $\pm 90^{\circ}$ head rotations with the IACC and ILD feature spaces. The results are presented as the number of exact estimates of DoA, the number of predictions within $\pm 5^{\circ}$, and the maximum error.

Feature Space	Exact	Within $\pm5^\circ$	Max Angular Error
KEMAR Reflections			
ITD	0.69%	6.944%	162°
ILD	0%	29.86%	47°
IACC	0%	13.89%	103°
ITD and ILD	0%	9.03%	73°
IACC and ILD	2.08%	32.64%	30°
IACC and ITD	1.39%	11.81%	151°
IACC, ITD and ILD	0.69%	15.97%	50°

Table 5.4: Comparison of NN performance when using different feature spaces: ITD, ILD, IACC, ITD and ILD, IACC and ILD, IACC and ITD, and IACC, ITD, and ILD. The number of exact, within $\pm5^\circ$, and maximum angular error are presented for the measured test reflection captured using the KEMAR 45BC

Pre-processing	Exact	Within $\pm5^\circ$
KEMAR Reflections		
Raw Signal	0.69%	25%
Cochleagram	2.08%	32.64%

Table 5.5: Comparison of NN performance between the raw gammatone signals and cochleagram based interaural cues. The number of exact and within $\pm5^\circ$ predictions are presented for the measured test reflection captured using the KEMAR 45BC. These tests are performed using the $\pm90^\circ$ head rotations with the IACC and ILD feature spaces.

largest impact on estimation accuracy, with the more within $\pm5^\circ$ and lower maximum angular error when compared to the ITD and IACC. These findings suggest that the combined IACC and ILD produces the best representation of the recorded signal for binaural DoA estimation.

From the results in Table 5.5, it is clear that using the cochleagram as a preprocessing step when generating the interaural cues results in more accurate DoA estimates. These results show that the trained NN produces more exact estimates of DoA when using the cochlea pre-processing. These findings form the basis by which the decision to use the cochleagram pre-processing stage was made.

5.6 Results

To test the accuracy and generalisability of the NN the angular error of the NN's predictions is computed as the angular difference between NN estimated and expected DoA. Five error metrics are used to asses the performance of the NNs: the percentage of the data where the NN exactly predicts DoA; the percentage of the data where the NN predicted the DoA within $\pm1^\circ$ of the expected value; the percentage of the data where the NN predicts the DoA within $\pm5^\circ$ of the expected value; the percentage of the data where front-back confusions occurred

Head	Loudspeaker	Exact	±1°	±5°	Front-Back Confusions	RMS Error
Direct Component						
KEMAR	Equator	19.44%	21.53%	64.58%	1.39%	5.18°
KEMAR	Genelec	12.50%	13.19%	79.86%	0.69%	4.63°
KU100	Equator	13.19%	17.36%	68.05%	0%	6.86°
KU100	Genelec	22.22%	27.08%	81.25%	0%	5.56°
Reflected Component						
KEMAR	Equator	2.08%	11.11%	32.64%	9.03%	13.59°
KEMAR	Genelec	0%	9.03%	27.78%	2.78%	9.74°
KU100	Equator	0%	9.03%	37.5%	2.78%	8.85°
KU100	Genelec	1.39%	14.58%	40.97%	1.39%	10.30°

Table 5.6: Direction of arrival accuracy comparison showing the, percentage of exact estimates of DoA; percentage of estimates within ±1° of the expected DoA; percentage of estimates within ±5° of the expected DoA; percentage of front-back hemisphere confusions; and RMS error in degrees, for the direct sound and reflected components measured with the KEMAR and KU100 binaural dummy heads, for the cascade-forward neural network

defined as DoA being estimated in the opposite front/back hemisphere; and the root mean square (RMS) error of the angular error. As the DoA estimation errors have non-parametric distribution, statistical analysis of this data is performed using the non-parametric Kruskal-Wallis test which in MATLAB is the function *kruskalwallis* [142], and reported as (χ^2 = , p = , degrees of freedom =).

In Table 5.6, the neural network accuracy across the test data is presented. The results show that the NN performs best when it is presented with the direct component, and there is a substantial reduction in performance when used to estimate the DoA of the reflected component. It is interesting to note that, with the exception of the KEMAR reflected data, the number of predictions within ±5° is greater when using the Genelec 8030 loudspeaker than the Equator D5, potentially as a result of differences in system alignment. Furthermore, for both the direct sound and reflected components using the KU100, the NN has a larger percentage of predictions within ±5° of the target value. It is possible that some of these differences are as a result of differences in the morpho-acoustic properties of each head and their ears, which could lead to differences in the observed interaural cues, particularly those dependent on spectral information. However, as the NN was trained with data for KEMAR, it would be expected to give better results, with data obtained from a KEMAR, as opposed to the KU100. It is therefore more likely that these differences are as a result of measurement system misalignment, or as the NN was trained with azimuth data in stepped in 5° intervals, it could be as a result prediction quantisation error.

In Figure 5.9, comparisons between the direct sound and reflected components for BRIRs captured with the KEMAR 45BC are presented. The boxplots show that for the direct sound, a maximum error of 12° and median error of 5° (mean error of 4.20°) were observed when using the Equator, and a median error of 4.5° (mean error of 3.87°) when using the Genelec. The reflected component on the other hand has a maximum error of 30° and median of 8.5° (mean error of 10.87°) for the Equator, and a max error of 25° and a median of 8° (mean error of 8.27°) when using the Genelec. There is a statistically significant difference between the DoA estimation errors for the direct sound and reflected components when using both the Equator D5 and Genelec 8030 loudspeakers, (χ^2 = 50.34, p = <0.0001, degrees of freedom = 287) and (χ^2 = 63, p = <0.0001, degrees of freedom = 287) respectively. The observed difference between direct sound component and reflected component could be due to differences in signal path distance, which was found to reduce prediction accuracy in [84, 85]. Additional sources of error could be attributable to small system misalignments at point of measurement, or lower SNR occurring due to signal absorption at the reflector or longer propagation path (source-reflector-receiver); an average SNR of approximately 22.40 dB and 13.14 dB was observed across the direct and reflected component respectively when comparing the amplitude of the residual signal after BRIR measurement to the desired signal.. Furthermore, the difference in performance between direct sound and reflections could also be as a result of multiple points of reflection and edge diffraction from the finite-length boundary producing additional signal paths from boundary to each ear, which could confuse interaural cues. An example of two reflection paths from the boundary to the left and right ear can be seen in Figure 5.10

In Figure 5.11, the comparison between direct sound and reflected components for BRIRs captured using the KU100 are presented. The boxplots show that for the direct sound, a maximum error of 23° is observed and a median error of 5° (mean error of 5.15°) when using the Equator, and a max error of 23° and a median error of 3° (mean error of 3.79°) when using the Genelec. The reflected component has a maximum error of 19° and median of 7° (mean error of 7.51°) for the Equator, and a max error of 35° and a median error of 6° (mean error of 7.87°) for the Genelec. As with the KEMAR measurements there is a significant difference in performance between the direct and reflected components for both loudspeakers, (χ^2 = 20.84, p = <0.0001, degrees of freedom = 287) and (χ^2 = 40.18, p = <0.0001, degrees of freedom = 287) respectively, with the reflected components producing significantly worse estimates of DoA.

In Figure 5.12, the comparison between the two binaural dummy heads is presented for both the direct sound and reflected components of the BRIRs. The boxplots show that the interquartile

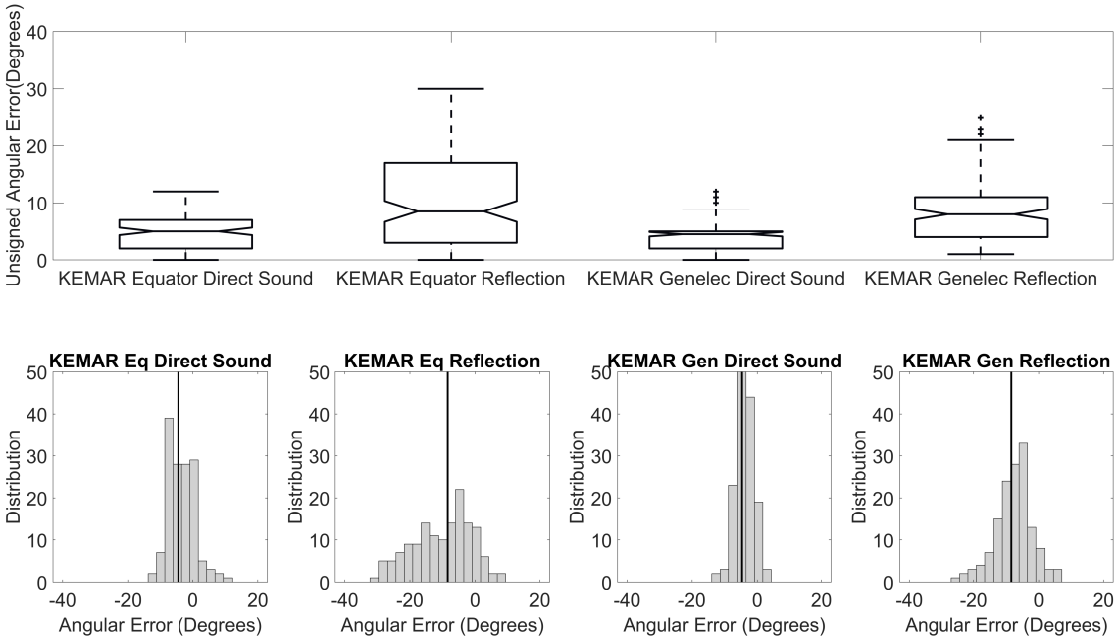

Figure 5.9: Comparison of angular errors for the neural network direction-of-arrival predictions for measurements with the KEMAR 45BC. The top image is a boxplot comparison of the angular error in the neural network predictions for the direct sound and reflected components. The bottom left two are the histograms showing the error distribution for the direction-of-arrival predictions of the direct sound and reflected component using the Equator D5 (denoted as Eq on figure), and the bottom right two are the error distribution for the direction-of-arrival predictions of the direct sound reflected component using the Genelec 8030 (denoted as Gen on the figure). The black line on the histograms depicts the median angular error.

ranges (region between the high- and low-notch for a box on the boxplot) for the direct sound measurements, with the exception of the KU100 Genelec direct sound measurements, overlap. Furthermore, there are not significant differences in the estimation errors, when considering absolute angular errors, between binaural dummy heads for three out of the four scenarios, (χ^2 = 1.08, p = 0.29, degrees of freedom = 287) (Direct Sound Component Equator D5), (χ^2 = 3.4, p = 0.07, degrees of freedom = 287) (Direct Sound Component Genelec 8030), and (χ^2 = 2.5, p = 0.11, degrees of freedom = 287) (Reflected Components Genelec 8030). This would suggest that generally, while RMS errors vary, the NN's performs comparably between the two dummy heads and loudspeakers set ups, therefore, the NN can be considered to be generalisable to different measurement scenarios. Comparing the angular errors observed in the output of the NN for the reflected component when using the Equator D5 shows that the KU100 has a significantly lower median angular error and produces more accurate estimates of DoA. However, this is not the case when using the Genelec loudspeaker where the boxplots show very little difference between the interquartile ranges. Given that the NN was trained with HRIRs captured using a KEMAR, the NN should perform best when analysing test binaural signals measured with a

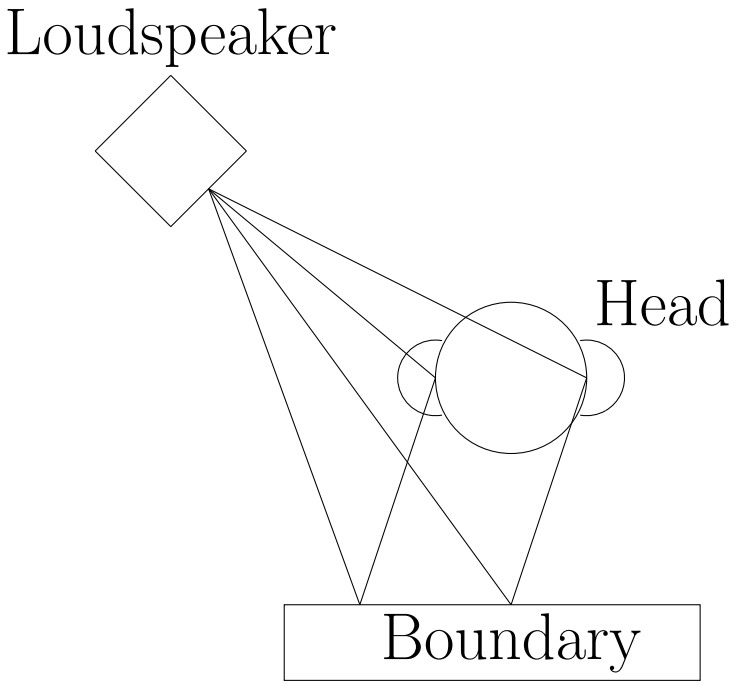

Figure 5.10: Figure showing the different in signal paths between a direct sound and a reflection, where two reflection paths exist from boundary to receiver, which could confuse interaural cues.

similar KEMAR. This suggests that, for the case of the Equator measurements, there could be some external influence, such as misalignment of the reflector or additional noise.

By investigating the neural networks' signed angular error over DoA, insight can be gained into any patterns occurring in the NN output predictions. Additionally, it will show how capable the NN is at predicting the DoA for signals with a DoA not represented within the training data. In Figure 5.13, the predicted DoA by the neural network (red and blue line) is compared against the expected DoA (black line), and the plot shows the comparison for the KEMAR direct sound measurement predictions (top left), KEMAR reflection measurement predictions (bottom left), KU100 direct sound measurement predictions (top right) and KU100 reflection measurement predictions (bottom right). Generally, the direct sound measurement predictions are mapped to the closest matching DoA represented in the training database, suggesting that the NN is incapable of making predictions for untrained directions of arrival. In the case of the reflections, the NN predictions tend to plateau over a larger range of expected azimuth DoA. This observation further shows the impact of the blurring of the interaural cues (Figures 5.16 and 5.17) producing regions of ambiguous cues in the reflection measurements, causing the NN to produces regions of the same DoA prediction.

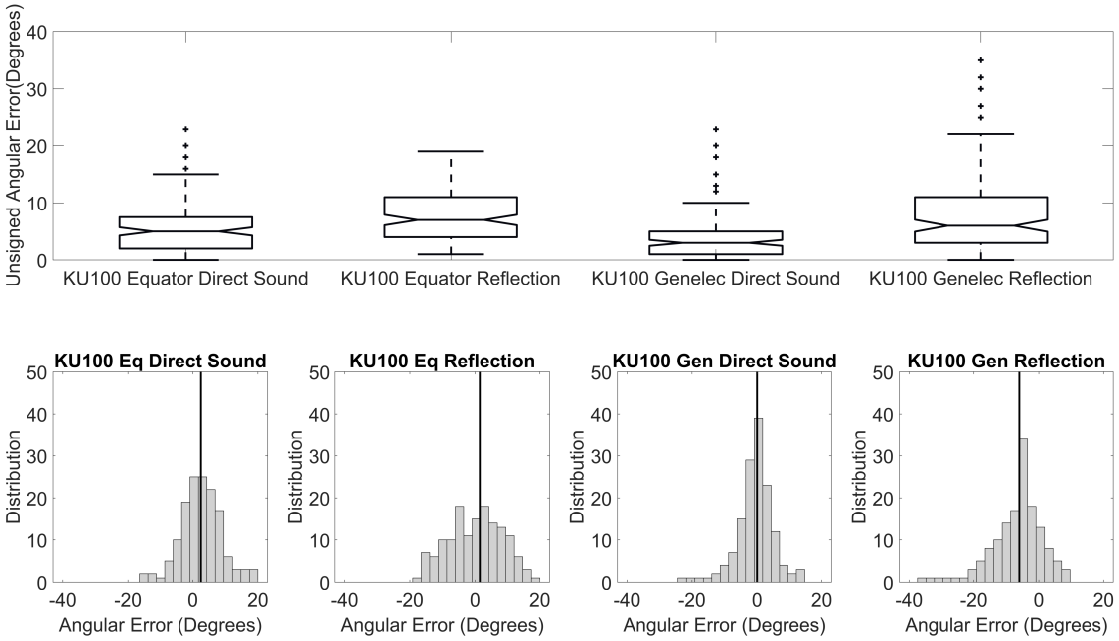

Figure 5.11: Comparison of angular errors for the neural network direction-of-arrival predictions for measurements with the KU100. The top image is a boxplot comparison of the angular error in the neural network predictions for the direct sound and reflected components;The bottom left two are the histograms showing the error distribution for the direction-of-arrival predictions of the direct sound and reflected component using the Equator D5 (denoted as Eq on figure), and the bottom right two are the error distribution for the direction-of-arrival predictions of the direct sound reflected component using the Genelec 8030 (denoted as Gen on the figure). The black line on the histograms depicts the median angular error.

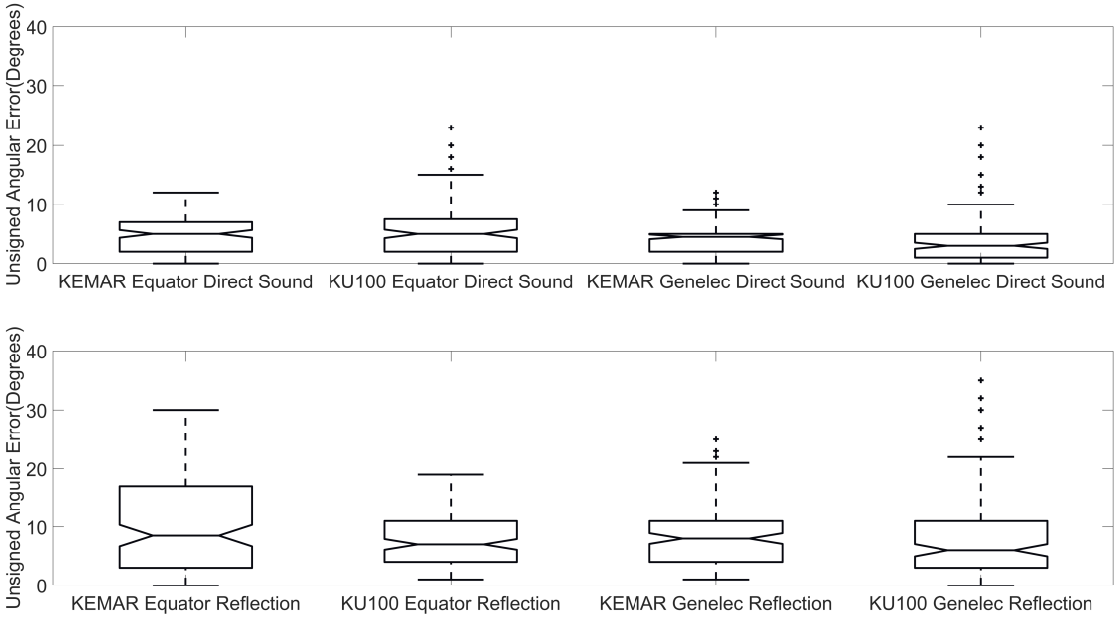

Figure 5.12: Boxplot comparison of angular errors for the neural network direction-of-arrival predictions between the KEMAR and KU100 dummy heads for direct sound (top) and reflected (bottom) components.

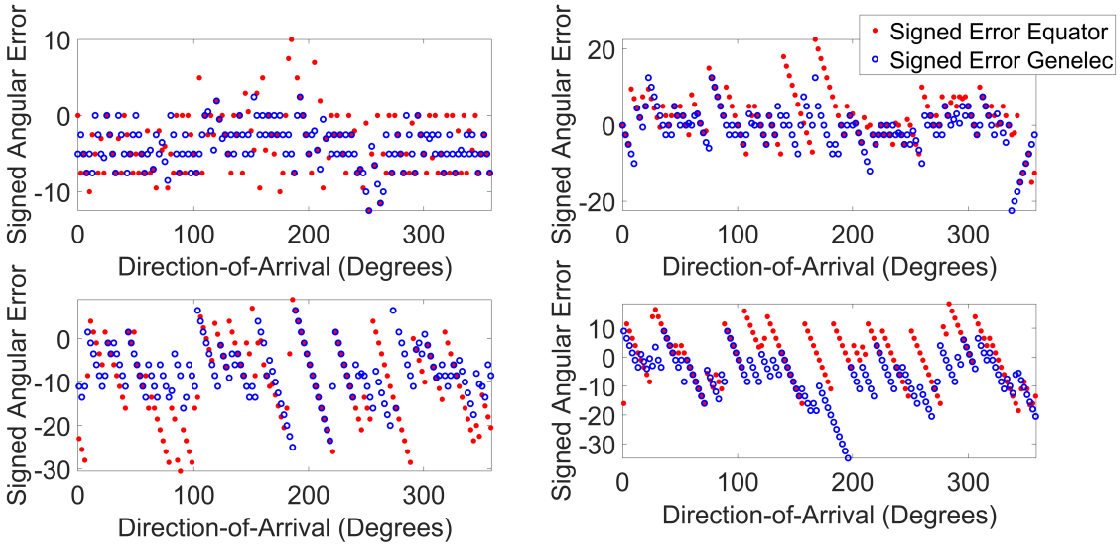

Figure 5.13: Plots of signed angular error over direction-of-arrival. The red line is the estimated DoA using the Equator D5 loudspeaker, the blue line is the estimated DoA using the Genelec 8030 and expected direction-of-arrival is the black line. The top left plot is for the KEMAR direct sound; the top right plot is for the KU100 direct sound; the bottom left is for the KEMAR reflection; and the bottom right is for the KU100 reflections.

5.6.1 Bias Correction

From the distribution of the angular errors in Figures 5.9, 5.11 and 5.13, it can be seen that, when comparing between test cases, the neural network is biased towards underpredicting the DoA when analysing the reflection data. This bias may be a result of a misalignment in the measurement system, specifically the reflector location, and therefore, can be accounted for by adjusting the angular error data such that it is zero-mean. As can be seen in Table 5.7, the number of predictions within 1° and 5° of the expected DoA has increased compared to the results in Table 5.6, with at most 83.33% of the data within 1° and 97.91% within 5°, both of which are for the KEMAR binaural dummy head and Genelec loudspeaker when analysing the direct sound. Furthermore, from the signed errors in Figures 5.14 and 5.15, it can be seen that, for the same comparisons previously presented, there are no statistically significant differences between the different test cases, $p > 0.39$, suggesting that the distribution of the errors are comparable across the test cases. However, when considering the absolute angular errors there are statistically significant differences across all test cases, $p < 0.01$, suggesting that the variance in the magnitude of these angular errors is significantly different. Furthermore, the results still show that the reflection data is less accurately estimated compared to the direct sound, with maximum angular errors between 17° and 28°, suggesting that the larger errors observed are a product of more than just system misalignments.

Head	Loudspeaker	Exact	±1°	±5°	RMS Error	Bias
Direct Component						
KEMAR	Equator	0.69%	66.67%	92.36%	3.98°	−3.70°
KEMAR	Genelec	1.39%	83.33%	97.91%	2.72°	−3.99°
KU100	Equator	15.28%	61.80%	86.11%	6.32°	2.32°
KU100	Genelec	0.69%	68.75%	86.11%	5.45°	0.85°
Reflected Component						
KEMAR	Equator	1.38%	49.31%	66.67%	9.32°	−10.15°
KEMAR	Genelec	3.47%	58.33%	81.94%	6.07°	7.85°
KU100	Equator	0%	54.17%	66.67%	8.80°	0.47°
KU100	Genelec	1.39%	54.17%	70.83%	8.14°	−6.56°

Table 5.7: Direction of arrival accuracy comparison for the bias corrected data showing the percentage of exact estimates of DoA; percentage of estimates within ±1° of the expected DoA; percentage of estimates within ±5° of the expected DoA; and RMS error in degrees, for the direct sound and reflected components measured with the KEMAR and KU100 binaural dummy heads, for the cascade-forward neural network.

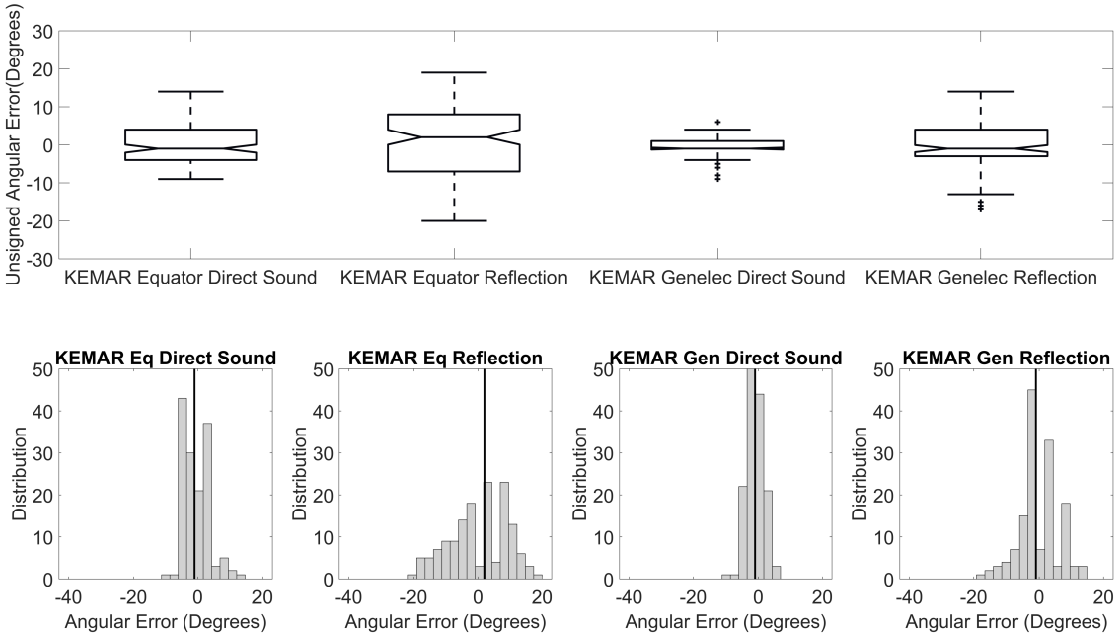

Figure 5.14: Comparison of angular errors after bias correction for the neural network direction-of-arrival predictions for measurements with the KEMAR 45BC. The top image is a boxplot comparison of the angular error in the neural network predictions for the direct sound and reflected components. The bottom left two histograms show the error distribution for the direction-of-arrival predictions of the direct sound and reflected component using the Equator D5 (denoted as Eq on figure). The bottom right two are the error distribution for the direction-of-arrival predictions of the direct sound reflected component using the Genelec 8030 (denoted as Gen on the figure). The black line on the histograms depicts the median angular error.

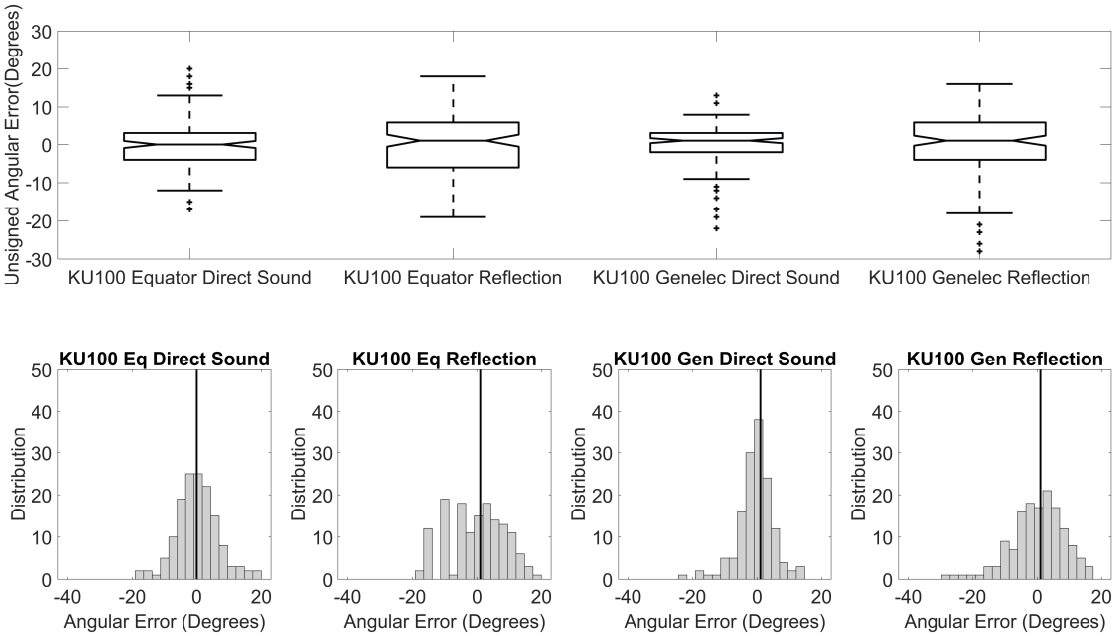

Figure 5.15: Comparison of bias corrected angular errors for the neural network direction-of-arrival predictions for measurements with the KU100. The top image is a boxplot comparison of the angular error in the neural network predictions for the direct sound and reflected components. The bottom left two histograms showing the error distribution for the direction-of-arrival predictions of the direct sound and reflected component using the Equator D5 (denoted as Eq on figure). The bottom right two are the error distribution for the direction-of-arrival predictions of the direct sound reflected component using the Genelec 8030 (denoted as Gen on the figure). The black line on the histograms depicts the median angular error.

5.6.2 Data Comparison

Comparing the IACC and ILD (Figures 5.16 and 5.17) between the direct sound and reflected components of the BRIR for the KEMAR and KU100 measurements shows a more distinct blurring for the reflected components measured with the KEMAR when compared to those measured with the KU100. This is particularly pronounced when comparing the ILD where the high frequency ILD values are smudged, and ripples in the ILD values across DoA start to appear at low frequencies. This again could suggest that a source of interference is present in the KEMAR measurements that is producing ambiguity in the measured signals' interaural cues. This could be due to noise present within the system and environment, or additional reflection paths from the reflective boundary being captured in the case of the reflected component.

5.7 Discussion

The results presented in Section 5.6 show that there is minimal difference in the accuracy of the NN when analysing the direct sound of BRIRs captured with both the KEMAR 45BC and

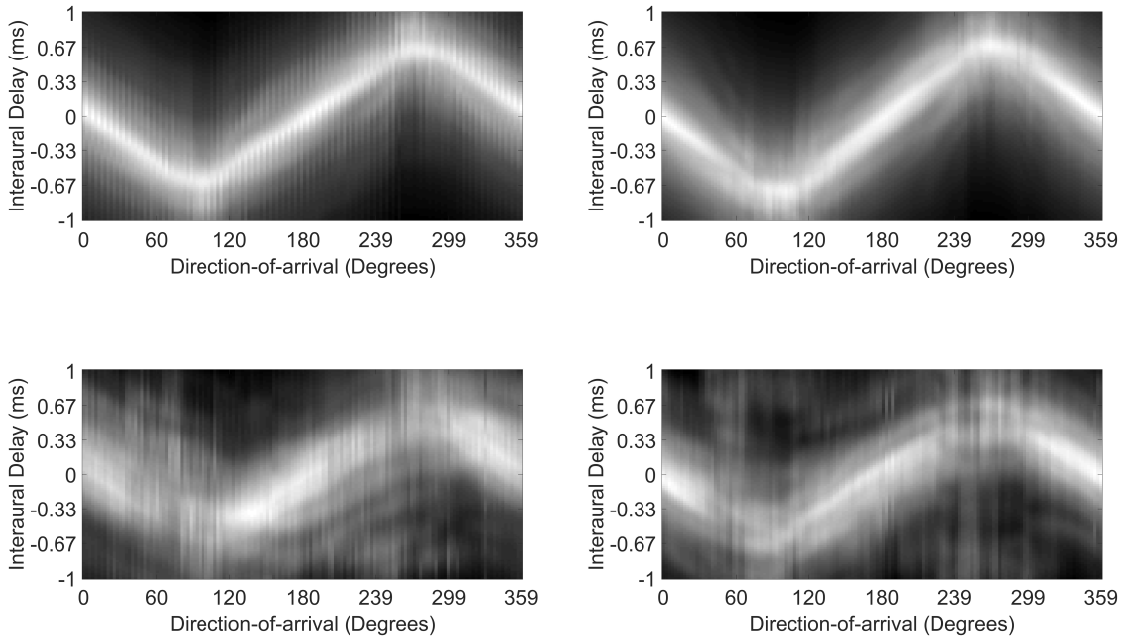

Figure 5.16: Comparison of interaural cross correlation across the direction-of-arrival for the KEMAR measured direct sound (top left), KEMAR measured reflection (bottom left), KU100 measured direct sound (top right) and KU100 measured reflection (bottom right) measured with the Equator D5.

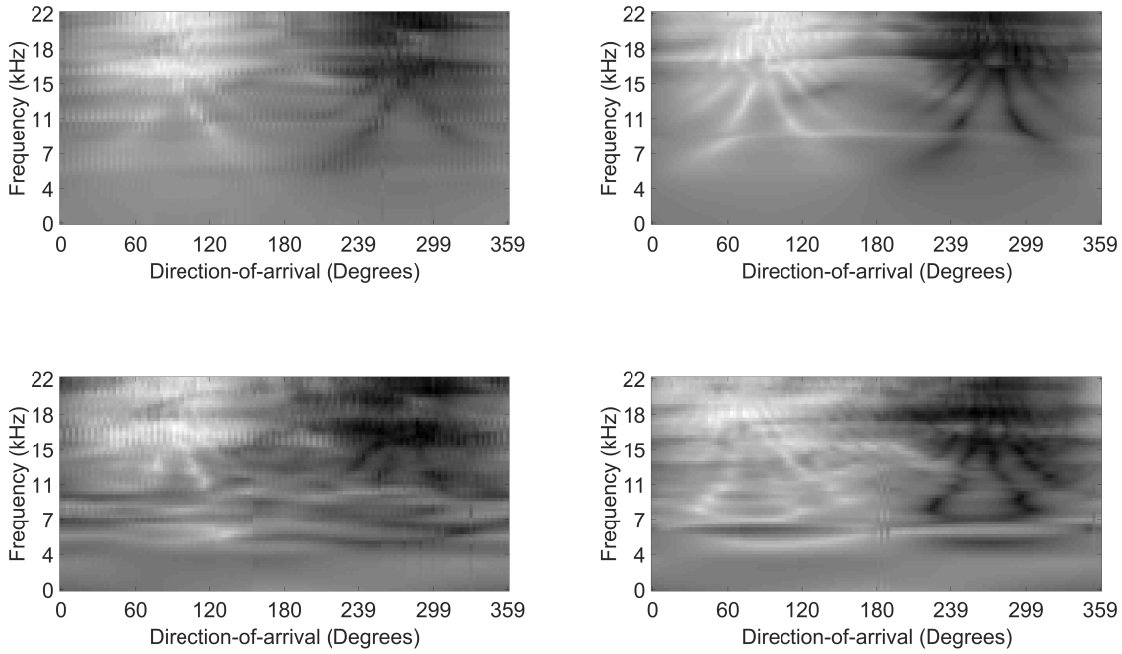

Figure 5.17: Comparison of interaural level difference across the direction-of-arrival for the KEMAR measured direct sound (top left), KEMAR measured reflection (bottom left), KU100 measured direct sound (top right) and KU100 measured reflection (bottom right) measured with the Equator D5.

the KU100, with all combinations not displaying a significant difference in estimation accuracy. However, the accuracy of the NN is significantly reduced when analysing the reflected component of the BRIRs, with a maximum RMS error of $13.59°$ and a minimum of $8.85°$, compared

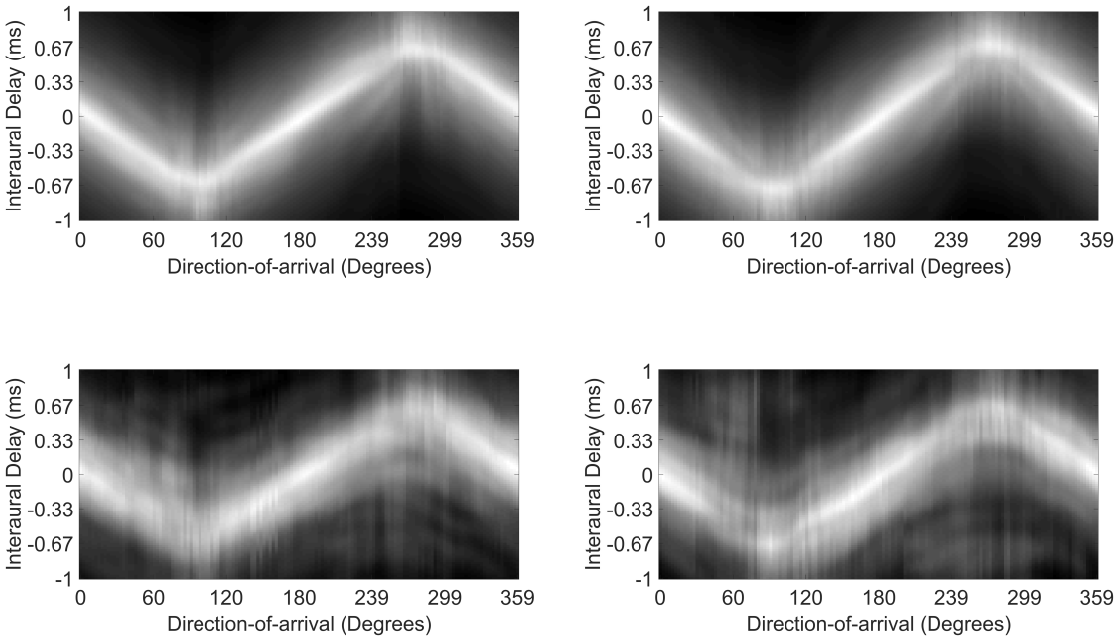

Figure 5.18: Comparison of interaural cross correlation across the direction-of-arrival for the KEMAR measured direct sound (top left), KEMAR measured reflection (bottom left), KU100 measured direct sound (top right) and KU100 measured reflection (bottom right) measured with the Genelec 8030.

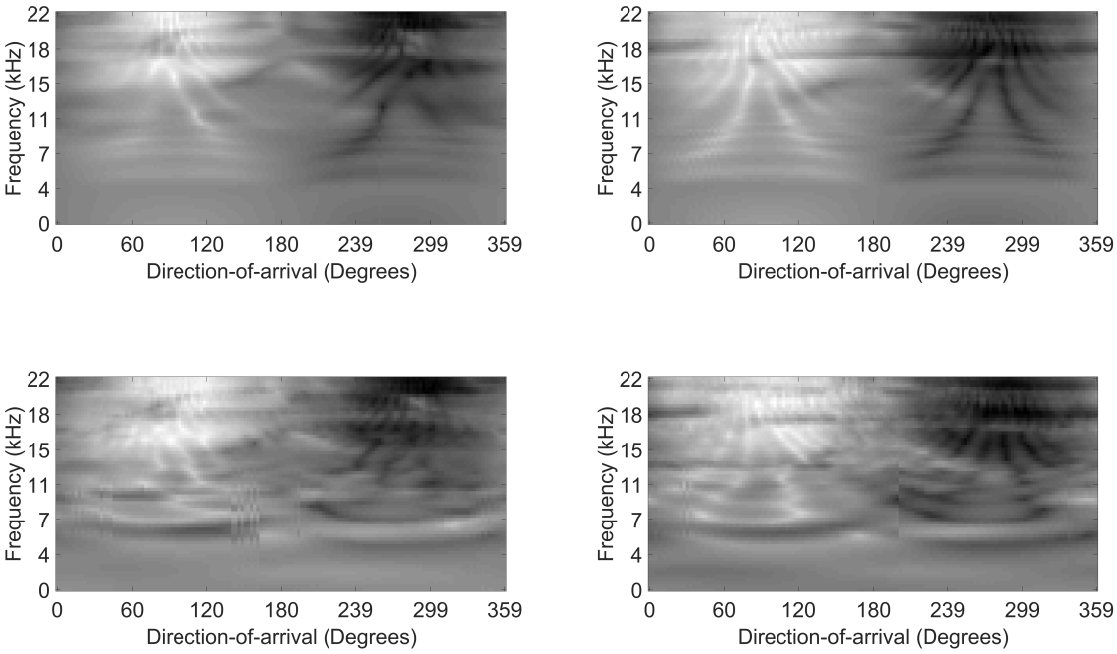

Figure 5.19: Comparison of interaural level difference across the direction-of-arrival for the KEMAR measured direct sound (top left), KEMAR measured reflection (bottom left), KU100 measured direct sound (top right) and KU100 measured reflection (bottom right) measured with the Genelec 8030.

to a maximum of 6.86° and a minimum of 4.63° for the direct sound. As with the direct sound the interquartile ranges overlap between measurement configurations, suggesting that there is minimal difference in the median values for these tests. This reduction in performance would

be expected between the direct and reflected component, due to the lower SNR ratio that would be observed for the reflected component, differences in the signal pathways between direct and reflected components, as a result of multiple points of reflection on the boundary , which could produce multiple closely-arriving reflections at the receiver. It is of interest that, while the maximum errors are greater, a larger number of reflected components measured with the KU100 are estimated within $\pm 5^{\circ}$ of the expected DoA. This difference could be as a result in misalignment of the source, receiver or boundary as a result of the floating floor in the anechoic chamber. Slight differences in the position/orientation of these parts of the measurement system can result in small differences in the direction that the signals arrive at the receiver, given that the NN has been shown to only estimate directions-of-arrival that existed in the training dataset these misalignments can result in larger deviations from the expected DoA. To account for these errors, a bias-corrected version of the angular errors was produced, which showed an improvement in the prediction accuracy of the NN, with a statistically similar distribution of signed angular error across all test cases. However, there were statistically significant differences in the magnitude of the angular error across all test cases. This would suggest that, while some of the DoA estimation errors can be attributed to system misalignment, it is not the sole cause for the increased inaccuracy of the DoA estimation error observed for reflections.

Analysis over different rotations (Table 5.1) shows that while the number of predictions within $\pm 5^{\circ}$ varies little between extent of rotation, the maximum error in the neural network's prediction decreases as the angle of rotation increases. The use of additional measurement orientations decreases the number of front-back confusions, with generally larger degrees of receiver rotations producing fewer front-back hemisphere errors, except when using $\pm 30^{\circ}$. Using larger rotations has the additional benefit of reducing the maximum prediction errors made by the neural network. This could be due to the larger angle of rotation resulting in a source to the rear of the listener being focused towards in the frontal hemisphere, producing a more accurate DoA prediction. It is interesting that there is a greater percentage of front-back confusions for the KEMAR 45BC compared to the KU100, given the low maximum error, this is likely caused by signals arriving near 90° and 270° (source facing the left or right ear) being estimated in the opposite hemisphere, which again could be as a result of small misalignment in the measurement system altering the direction that the signal arrived at the receiver.

The lack of significant differences between the direct sounds measured with the two binaural dummy heads agrees with the findings of May et al. [86], who found that a GMM trained with an MCT dataset was able to localise sounds captured with two different binaural dummy heads.

Notable differences between the KEMAR 45BC and KU100 include: morphological differences of the head and ears between binaural dummy head microphones; the KEMAR 45BC has a torso; the KU100's microphones have a flat diffuse-field frequency response; and the material used for the dummy head microphones.

The overall accuracy of the method presented in this chapter is, however, lower than that found in [86]. This could be a result of the type of signals being analysed, which, in this study, are 3.8 ms-long impulsive signals as opposed to longer speech samples. Comparing the work presented in this chapter to the NN-based implementations in [32], the method proposed in this chapter under-performs compared to reported findings of 83.8–100% accuracy across different test scenarios, with a 2.55% difference between the best case in this study compared to the worst in [32]. However, the work in [32] only considers signals in the frontal hemisphere around the head and again considered longer audio samples for the localisation problem.

The method here out-performs that presented in [76] with lower error values across all measurement sets, when compared to the relative-errors of 24.0% reported in [76] for real-world continuous pink noise using a multi-layered perceptron NN.

The average errors reported in this chapter are lower than that presented in [64], which reported average errors in the range of 28.7° and 54.4° when analysing the components of measured BRIRs. However, the results presented in [64] considered higher reflection orders, and therefore, further analyses of the performance of the NN with full BRIRs is required for a more direct comparison to be made.

More importantly in the context of this thesis, even when considering the bias-corrected results, the results highlight that the accuracy with which DoA is estimated in the case of reflected components is not accurate enough to be considered as a viable method for spatiotemporal based geometry inference. The larger angular errors for the reflections would lead to inaccurate estimation of corresponding image-source locations, which in turn would result in an estimate of the boundary's location with an angular error equal to that of the DoA estimate. Furthermore, considering that BRIRs consist of multiple overlapping reflections it is likely that the DoA estimation would be even less accurate. Therefore, alternative measurement methods that result in more accurate DoA estimates need to be used when considering geometry inference based on limited receiver location data points.

5.8 Conclusions

The aim of the study presented was to investigate the application of neural networks in the spatial analysis of binaural room impulse responses, and whether it is possible to obtain sufficiently accurate DoA for reflected components to allow for accurate geometry inference. The neural network was tested using binaural room impulse responses captured using two different binaural dummy heads with two different loudspeaker sources. The neural network was shown to demonstrate minimal difference in accuracy when analysing the direct sound of the binaural room impulse response across the two binaural dummy heads, with 64.58% (KEMAR Equator), 79.86% (KEMAR Genelec), 68.06% (KU100 Equator), and 81.25% (KU100 Genelec) of the predictions being within $\pm 5^{\circ}$ of expected values. However, upon presenting the NN with reflected components for analysis, the accuracy of the predictions was significantly reduced. The NN also generally produced lower average errors for reflected components of the binaural room impulse response captured with the KU100 than those with the KEMAR. Comparisons of the interaural cues for the direct sound and reflected components show a distinct blurring in the cues for the reflected components measured with KEMAR, which is present to a lesser extent for the KU100. This blurring could be a product of lower signal-to-noise ratios or multiple reflection paths from the boundary arriving at the receiver, leading to greater ambiguity in the measurements. Furthermore, difference in performance between direct and reflected components in all cases could be as a result of difference in signal pathways arriving at the ear as a result of multiple reflection points of origin on the boundary. This would suggest that training the NN with additional data relating to large numbers of reflections could lead to an improvement in DoA estimation accuracy for reflected components. However, in its current state the accuracy of the DoA estimation for all components of a binaural room impulse response is not sufficient to make it viable for geometry inference, where ideally all DoA estimation errors need to be as small as possible to allow for accurate boundary estimation. Further development of this approach, with the intent of use in geometry inference, would need to work on improving the accuracy of DoA estimation while including estimation of elevation DoA. Geometry inference within this thesis will therefore focus on alternative compact microphone arrays receivers with larger channel count than the two that are used in a binaural dummy head, and where more accurate estimates of DoA are therefore easily obtainable.

Chapter 6

Spatiotemporal Decomposition Based Reflection Detection

6.1 Introduction

In the previous chapter a binaural model fronted Neural Network (NN) was used to analyse the direction-of-arrival (DoA) of reflections in a Binaural Room Impulse Response (BRIR). The results showed that, while the direct sound was in the majority of cases predicted within $\pm 5^{\circ}$ of the expected DoA, there was a significant reduction in performance when estimating the DoA of subsequent reflections, with a maximum angular error of $\pm 35^{\circ}$. While this method produces more accurate results than the work presented in [64], it is not as accurate as the spherical microphone based approaches in [12]. Furthermore, the inaccuracies observed for the binaural model fronted NN estimates of DoA for reflections are too large for a BRIR based geometry inference method to be viable. Hence, this chapter will propose a method for estimating time-of-arrival (ToA) and DoA of reflections in Spatial Room Impulse Response (SRIR) measured with a spherical microphone array. This specific microphone array has been chosen based on the accuracy that was achieved for reflection DoA estimation in [12], which also used the EigenMike EM32.

Previous work investigating the analysis of reflection information measured with a spherical microphone array [12, 50, 51], considered the problem as detecting the DoA of the main reflections through spatial decomposition of the SRIR using beamformers. It is proposed that by expanding on these methods to analyse the SRIR over short time-frames, performing spatial and temporal

decomposition, both ToA and DoA can be estimated. The method works as follows. Firstly, beamforming is performed on a time-frame to measure the signal power incident from each direction. Image-processing techniques are then used to extract peak locations in the resulting spatial map representing the DoA of arriving reflections. Finally, the ToA of the reflection is then estimated by steering a beam in the direction of the DoA for that time-frame. By solving the reflection detection problem through such spatiotemporal decomposition, overlapping and simultaneously arriving reflections can be detected as individual arrivals, therefore, addressing a key issue with existing reflection detection techniques as discussed in Chapter 3. The proposed method, processing, and subsequent analysis of results form the main contributions of this chapter.

This chapter will be organised as follows: Section 6.2 will define the problem formulation, Section 6.3 will present the proposed method, Section 6.4 will present the testing methodology, Section 6.5 will present the results, Section 6.6 will discuss the results, and Section 6.7 will conclude the chapter.

6.2 Problem Formulation

A spherical microphone array measures the sound pressure on the surface of a sphere, spatially sampled at the microphone positions distributed on the surface. Therefore, as discussed in Chapter 2, the array's response to a plane wave arriving from a specific DoA can be expressed using spherical harmonics. This property makes spherical microphone arrays ideal for beamforming, as the steering vector can be expressed mathematically and not through physical measurements. As described in Chapter 2 the SRIR $\mathbf{H}(t)$ can be described in the spherical harmonic domain as,

$$\mathbf{H}(t) = \sum_{i=1}^{\infty} \mathbf{y}(\theta_i, \phi_i) a_i \text{sinc}(t - \tau_i) + \mathbf{R}(t) \tag{6.1}$$

where a_i, is the amplitude of the arriving signal, $\text{sinc}(t-\tau_i)$ is a sinc function with time of arrival τ_i, and $\mathbf{R}(t)$ is the time-variant residual noise component, $\mathbf{y}(\theta, \phi)$ is the real-valued spherical harmonic vector as described in Chapter 2.

From this representation of the SRIR it is evident that each reflection arrives at the spherical array at a specific time- and direction-of-arrival. The problem of reflection detection is therefore, to detect all of these reflections as individual arrivals.

6.3 Method

The problem of reflection detection, in this case, can be implemented through spatial decomposition of short time-frames, where the aim is to detect the arrival of directional signals at the microphone array. Contrary to the circular variance local maxima method discussed in Chapter 3, which also uses spatial information (the DoA variance in a time-frame), the proposed approach can disambiguate between multiple arrivals in a single time-frame as a result of the beamforming process used. The problem of reflection detection can, therefore, be split into a four-stage sequential process; (i) windowing of the RIR at the current time-frame; (ii) spatial decomposition of the time-frame using a beamformer; (iii) detection of directional signals in the time-frame; and (iv) estimation of the time- and direction-of-arrival for each reflection. A flowchart of the proposed method can be seen in Figure 6.1.

The approach proposed here, named as the Eigenbeam Detection and Evaluation of Simultaneously Arriving Reflections (EDESAR) method, performs spatiotemporal decomposition on a normalised SRIR, to detect directional impulses arriving at a spherical microphone array. The analysis is performed sequentially over 0.45 ms windowed time-frames (the length of the direct sound in the real-world measurements used in this chapter) with a 50% frame overlap, chosen empirically to reduce the number of false-positive detections. The use of short-time frames allow for reflections arriving from close DoAs but different ToA to be detected separately.

One of the main problems with reflection detection algorithms is false-positive or inaccurate detections which, in the case of geometry inference problems, produces errors in boundary position estimations. To try and reduce the likelihood of inaccuracies and false-positive detections occurring, two assumptions are made. The first assumption is that there is a maximum sample magnitude threshold ϵ_α, below which the time-frame is more likely to be noise or part of the diffuse field. Secondly, it is assumed that there is a diffuseness threshold ϵ_d, above which the number of arriving signals or residual noise present in the time-frame will negatively impact the estimation of both ToA and DoA for any reflections present in that time-frame.

To improve the accuracy with which diffuseness is estimated for a time-frame, filtering is used to remove diffuse spectral components of the SRIR. The cut-off frequency at which the number of microphones positioned on the sphere is inadequate to accurately capture spatial information [143], is referred to as the spatial Nyquist frequency, which for the EigenMike EM32 is at 8 kHz [144]. To account for this the audio is low-pass filtered at 5 kHz, and also high-pass filtered

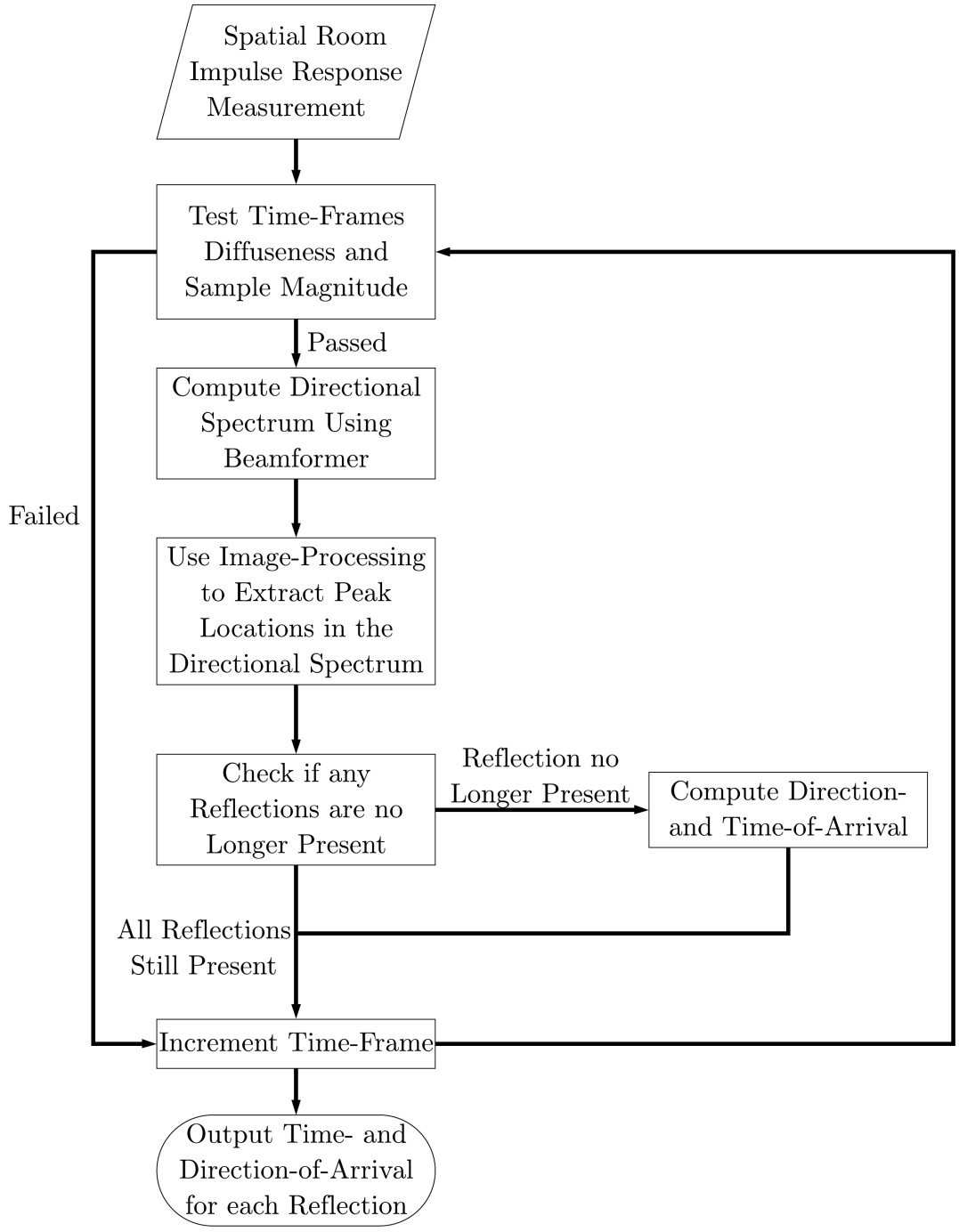

Figure 6.1: Flowchart detailing the sequential processing stages used to compute the time- and direction-of-arrival for reflections detected within a spatial room impulse response.

at 100 Hz to remove any low-frequency rumble. In order to maintain the temporal information contained within the time-frame, filter phase is accounted for by circular-shifting the resulting time-frame to align with the original time-frame.

To compute the diffuseness profile for each time-frame the Covariance Matrix Eigenvalue Dif-

fuseness Estimation (COMEDIE) algorithm [145] is used. The COMEDIE algorithm has been shown to produce a more robust estimate of diffuseness, compared to DirAC[37] and Thiele–Gover Diffuseness Measure [146] in [145], as a result of being able to disambiguate between multiple uncorrelated sound sources and spatially diffuse noise [145]. Furthermore, this diffuseness estimator is shown to produce comparable measurements of diffuseness when presented with multiple correlated sound sources as well [145], making it ideal for analysing SRIRs. The COMEDIE algorithm is based on the observation that in the presence of predominantly diffuse noise the eigenvalues computed from the covariance matrix of a signal will be similar. Therefore, the eigenvalues of the covariance matrix will have the largest variation when only a single plane wave is present. This diffuseness estimation can be computed for each spherical harmonic order as the ratio between the deviation of the measured eigenvalues γ to that of an ideal, completely non-diffuse case γ_0, which from [145], is,

$$d = 1 - \frac{\gamma}{\gamma_0} \tag{6.2}$$

where γ is computed as,

$$\gamma = \frac{1}{\langle u \rangle} \sum_{n=1}^{(N+1)^2} |u_n - \langle u \rangle| \tag{6.3a}$$

$$\langle u \rangle = \frac{1}{(N+1)^2} \sum_{n=1}^{(N+1)^2} u_n \tag{6.3b}$$

where u_n is the nth eigenvalue computed for the covariance matrix and N is the spherical harmonic order [145]. The ideal single plane wave eigenvalue deviation is defined in [145] as,

$$\gamma_0 = 2[(N+1)^2 - 1] \tag{6.4}$$

If a time-frame meets both the sample magnitude and diffuseness conditions, a spherical beamformer is used to perform spatial decompositions of the time-frame to extract any directional signals. In this case the MVDR beamformer, originally proposed by Capon in [61] and adapted for spherical microphone arrays in [30, 63], is used. As discussed in Chapter 3 the key benefit of the MVDR beamformer is its ability to adapt the steering vector weighting based on the variance of

the recorded signal, and an implementation of the MVDR used in [12] was shown to outperform Eigenbeam-Multiple Signal Classification (EB-MUSIC), Eigenbeam - Estimation of Signal Parameters via Rotational Invariance Techniques (EB-ESPRIT), the delay-and-sum beamformer, and the plane wave decomposition beamformer. These adaptive weights improve the robustness of the algorithm to the residual noise component, which in turn should improve the accuracy with which DoA can be estimated. As with the diffuseness estimator, the time-frame is first filtered to remove frequencies above the spatial Nyquist frequency. Before computing the beamformer output, the spherical harmonics vector is weighted to ideally minimise residual noise, by minimising the total array output,based on the signals' covariance matrix, while setting the gain in the desired direction to unity, these MVDR beamformer weights are computed as [62, 63],

$$\widehat{\mathbf{w}}(\Psi) = \frac{{\mathbf{R_{HH}}}^{-1}(t_f)\mathbf{y}(\Psi)}{\mathbf{y}^T(\Psi){\mathbf{R_{HH}}}^{-1}(t_f)\mathbf{y}(\Psi)} \tag{6.5}$$

where $(.)^{-1}$ denotes matrix inversion, $\mathbf{y}(\Psi)$ is the $[16 \times 1]$ spherical harmonic vector computed using the *getSH* function in the *Spherical Harmonic Transform Library* [147], and $\mathbf{R_{HH}}(t_f)$ is the $[15 \times 15]$ covariance matrix of time-frame t_f in RIR $\mathbf{H}(t)$ with dimensions $[2000 \times 16]$ after filtering is computed using the formulation in [30] as,

$$\mathbf{R_{HH}}(t_f) = \mathbf{H}(t_f)^T\mathbf{H}(t_f) + \frac{\mathbf{I}_{(N+1)^2}}{4\pi} \tag{6.6}$$

where $\mathbf{I}_{(N+1)^2}$ is the $[(N+1)^2 \times (N+1)^2]$ identity matrix, and $\frac{\mathbf{I}_{(N+1)^2}}{4\pi}$ represents the covariance matrix of a diffuse sound [30, 145], which is used to improve robustness to rank deficient covariance matrices as a result of transient signals (such as reflections) as described in [61]. The MVDR beamformer output is then computed as,

$$\zeta(\Psi) = \widehat{\mathbf{w}}(\Psi)^T\mathbf{R_{HH}}(t_f)\widehat{\mathbf{w}}(\Psi) \tag{6.7}$$

where $\zeta(\Psi)$ is the power of the signal in the direction Ψ using the $[15 \times 1]$ real-valued weighted spherical harmonic vector $\widehat{\mathbf{w}}(\Psi)$. To improve the accuracy of the DoA estimation, it is proposed that the omnidirectional-channel (first-channel) should be removed, as it will be equally weighted across all DoA, adding a residual bias value to $\zeta(\Psi)$.

Using beamforming, a heat map of the signal power across DoA, which will referred to as the *directional spectrum*, can be extracted for each time-frame. This is computed as the directional power of the signal steered across a grid of azimuth and elevation positions from $0° \leq \theta \leq 359°$ and $0° \leq \phi \leq 180°$ in one degree increments, and is expressed as,

$$\mathbf{\Lambda} = \begin{bmatrix} \zeta(\Psi = [0,0]) & \zeta(\Psi = [1,0]) & \cdots & \zeta(\Psi = [359,0]) \\ \zeta(\Psi = [0,1]) & \zeta(\Psi = [1,1]) & \cdots & \zeta(\Psi = [359,1]) \\ \vdots & \vdots & \vdots & \vdots \\ \zeta(\Psi = [0,180]) & \zeta(\Psi = [1,180]) & \cdots & \zeta(\Psi = [359,180]) \end{bmatrix} \tag{6.8}$$

An example of a directional spectrum, derived from a typical time-frame of a third-order spherical harmonic signal representation of a SRIR containing two distinct simultaneously arriving reflection, can be seen in Figure 6.2.

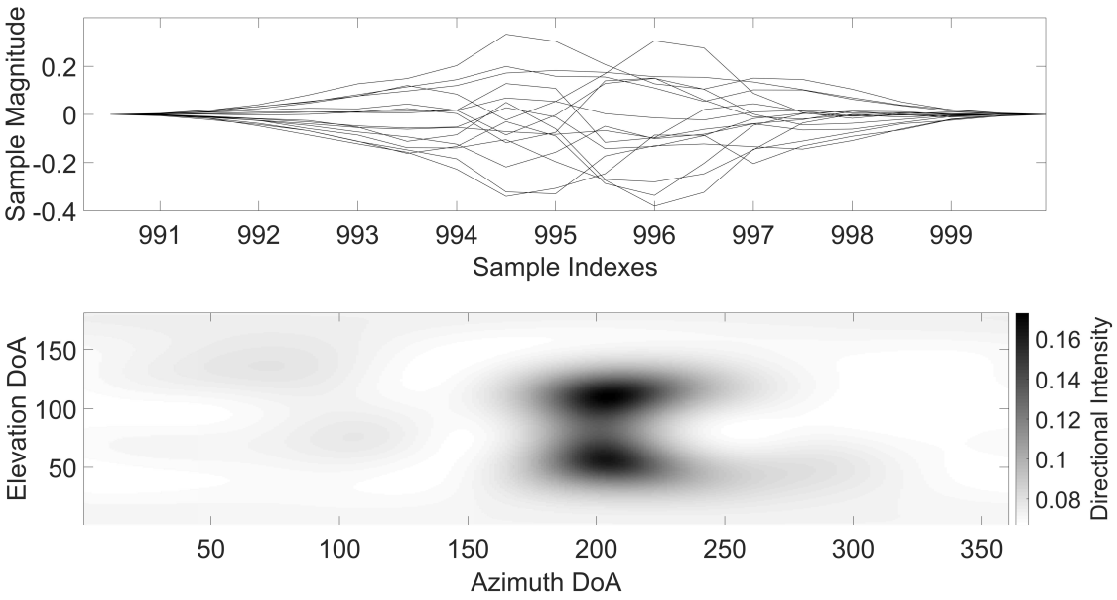

Figure 6.2: Top: A typical time-frame of a third-order spherical harmonic signal representation of a spatial room impulse response containing two distinct simultaneously arriving reflections, where each line represents a different channel in the third-order spherical harmonic signal. Bottom: The directional spectrum computed for the time-frame, where the darker regions indicate the arrival of strong directional components in the signal.

As discussed in the problem formulation, a reflection can be described as a directional signal arriving at the microphone array, and therefore should appear as, and be detected by searching for, regions of higher power in the directional spectrum. When using region detection methods on the directional spectrum, further processing is required to ensure reflections with overlapping spatial regions (Figure 6.2) are detected as individual arrivals. Given that the directional spec-

trum can be represented as a heat map, the segmentation and detection of these regions can be approached as an image-processing problem, as follows.

The directional spectrum is first converted to a greyscale image, by remapping the power matrix to values between 0 (black) and 1 (white), where in this case black represents regions of higher-power. To try and reduce the likelihood of false-positive detections, the dynamic range of the power matrix is compressed, such that $\min(-\mathbf{\Lambda}) \div 2 = 0$ is black, and $\max(-\mathbf{\Lambda}) = 1$ is white. This process in turn increases the dynamic range of the greyscale image, as a smaller range of values are mapped between black and white, which ideally will reduce the likelihood of unwanted detections. The greyscale image $\widehat{\mathbf{\Lambda}}$, which from [148] is computed as,

$$\widehat{\mathbf{\Lambda}} = -\mathbf{\Lambda}\frac{1}{\max(-\mathbf{\Lambda}) - \frac{\min(-\mathbf{\Lambda})}{2}} + \left(\frac{\min(-\mathbf{\Lambda})}{2}\frac{1}{\max(-\mathbf{\Lambda}) - \frac{\min(-\mathbf{\Lambda})}{2}}\right) \tag{6.9a}$$

$$\widehat{\mathbf{\Lambda}} = \begin{cases} 1, & \forall \widehat{\mathbf{\Lambda}} > 1 \\ \widehat{\mathbf{\Lambda}} & \forall \widehat{\mathbf{\Lambda}} \leq 1 \end{cases} \tag{6.9b}$$

To segment any overlapping regions within the greyscale image, which can be accomplished using the *watershed* algorithm as implemented in MATLAB [149]. To improve the segmentation accuracy, additional processing of the directional spectrum is required [150]. The greyscale image is first converted into a binary mask using the threshold $\widetilde{\mathbf{\Lambda}} = \widehat{\mathbf{\Lambda}} \leq \epsilon_{msk}$. This extracts only the regions with high directional power (darkest regions), and ideally removes any unwanted background noise from the directional spectrum, as seen in Figure 6.3. To produce more distinct regions within the remaining directional spectrum an extended-minima mask is applied to the binary mask. The extended-minima transform is a masking technique that uses the distance transform computed for a binary image to focus the regional minima on a central point. The distance transform is a matrix where each index $[i, j]$ represents the Euclidean distance between the pixel at $[i, j]$ in the binary image and the the nearest zero valued index. Using the distance transform the extended-minima mask is computed and then imposed onto the binary image, and an example of this can be seen in Figure 6.4. The *watershed*, as used in [64], algorithm is then applied to the transformed binary mask, producing a label matrix where positive valued integers are assigned to each separate region, with the regions separated by zero valued indices. This label matrix is applied to the binary image, by setting the indices in the binary image where the watershed algorithm outputs a 0 to 0, as shown in Figure 6.5, a MATLAB implementation of this whole process is presented in Algorithm 3.

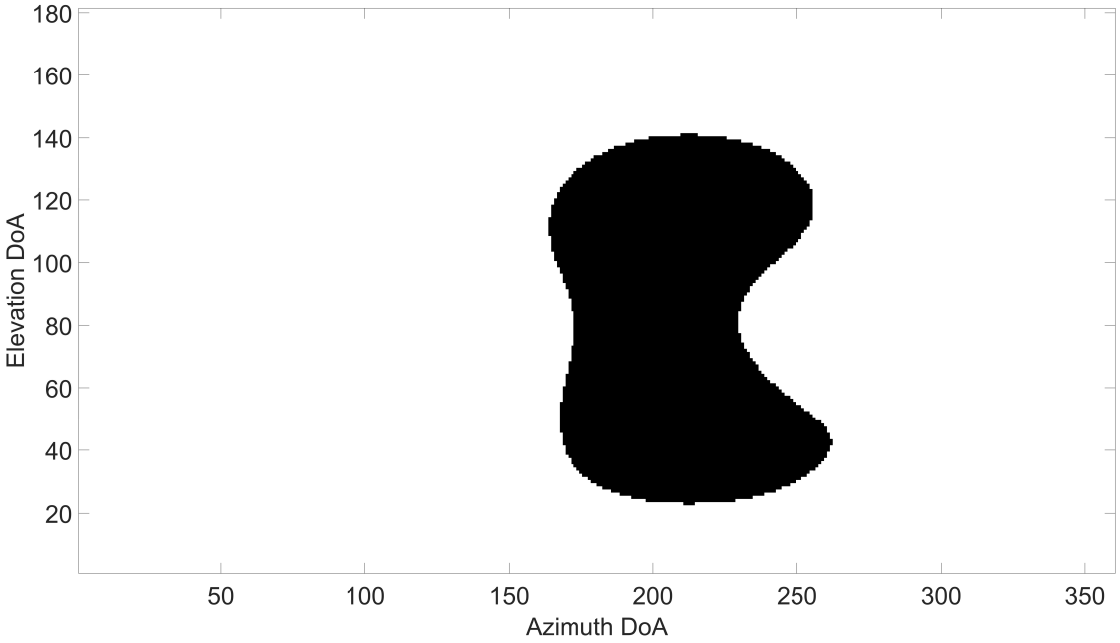

Figure 6.3: Example of the binary mask produced for the directional spectrum of the signal presented in Figure 6.2.

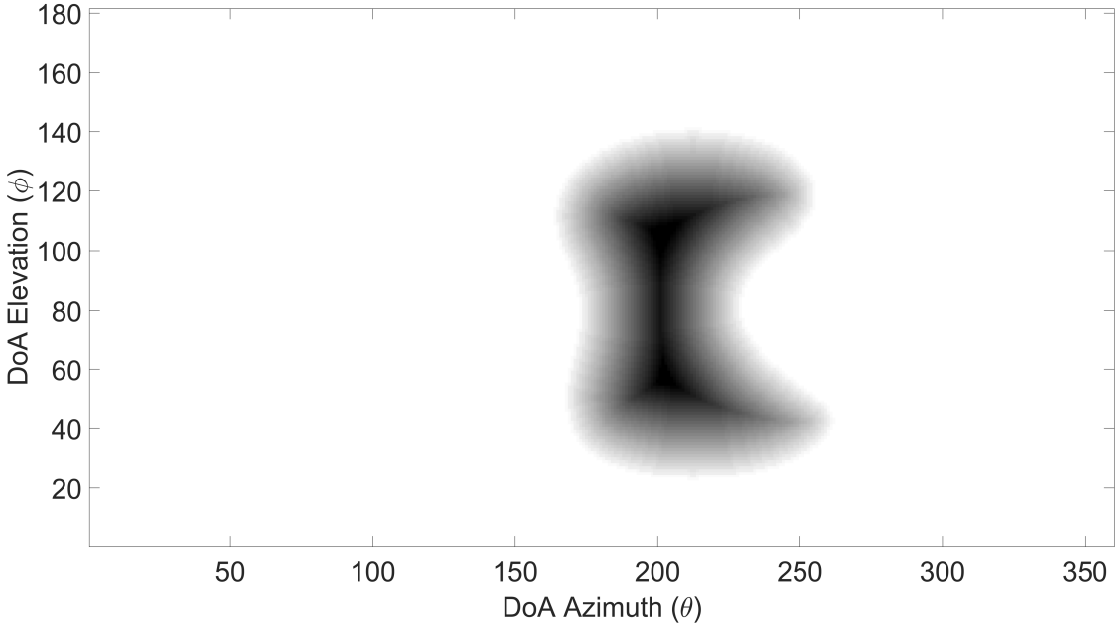

Figure 6.4: An example of the mask of the directional spectrum (as presented in Figure 6.2) after having an extended minima mask applied.

From the segmented image, reflections present in a time-frame can be detected by searching for regions of connected 1s within the masked binary image - which represent the arrival of a signal in the directional spectrum. In image-processing, this region detection process is commonly achieved using a nested loop, iterating over both rows and columns in the binary image, searching for indices where a 1 is present. A grid search is then performed to find all connected 1s, and a unique numeric label is assigned to the detected region [151]. A MATLAB example of

Algorithm 3: Computation and processing of directional spectrum to segment out overlapping regions. MATLAB functions are indicated in bold, and developed functions are denoted in bold and italics.

```
// Step 1: Compute the directional spectrum using the MVDR
   beamformer.
1 directionalSpectrum = MVDR(filteredTimeFrame);
// Step 2: Convert directional spectrum into a grayscale image
2 beta = 1 / (max(-directionalSpectrum(:)) - (min(-directionalSpectrum(:))*0.5));
3 greyScaleImage =
   (-directionalSpectrum*beta)+((min(directionalSpectrum(:))*0.5)*beta);
4 greyScaleImage = max(grayScaleImage, min(grayScaleImage,1));
// Step 3: Compute the binary mask of the grayscale image
5 binaryMask = greyScaleImage <= 0.1;
// Step 4: Compute the watershed transform for the binary mask
6 D = -bwdist(~binaryImage) ;     // Compute the distance transform of the
   binary image
7 mask = imextendedmin(D,2) ;   // Compute the extended minima transform
8 D = imimposemin(D,mask) ;      // Apply the extended minima transform
9 maskedImage = watershed(D) ;          // Compute the watershed transform
10 binaryMask(maskedImage ==0) = 0 ;          // Apply watershed transform
```

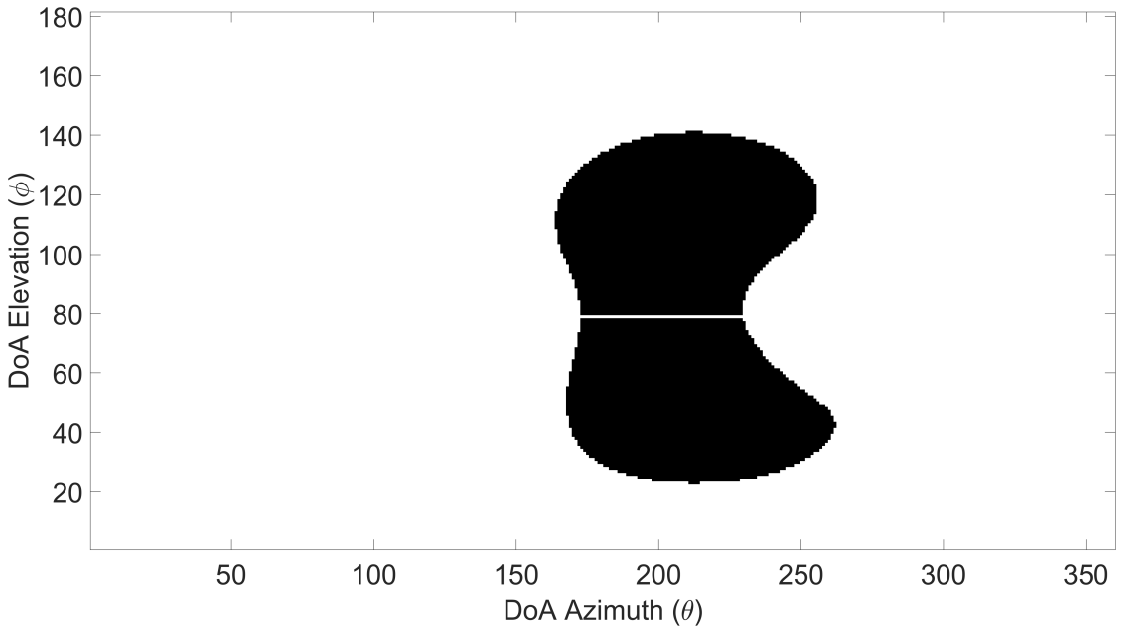

Figure 6.5: An example of the resulting directional regions after the watershed mask has been applied. The overlapping regions as presented in Figure 6.2 are now separated by a white line.

this process, based on [151] can be seen in Algorithm 4. Each unique label (except zero) within this labelled matrix in this case represents the arrival of a different reflection in this time-frame. The spatial region occupied by each reflection can be simply represented as the convex hull of the labelled area. An example of the detected spatial regions within the directional spectrum

can be seen in Figure 6.6. It is important to note that this two-dimensional representation of the directional spectrum represents an unwrapped sphere. The implication of this being that if a reflection arrives at the microphone close to $\theta = 0°$, the spatial region it occupies will exist around both $\theta \approx 0°$ and $\theta \approx 359°$, as seen in Figure 6.7. This will result in the reflection being detected as two arrivals, one for each spatial region. Therefore, if a region that falls close to $\theta \approx 0°$ or $\theta \approx 359°$ is detected, the corresponding region on the opposing side of the image is searched for within the detected regions, and the regions combined if found.

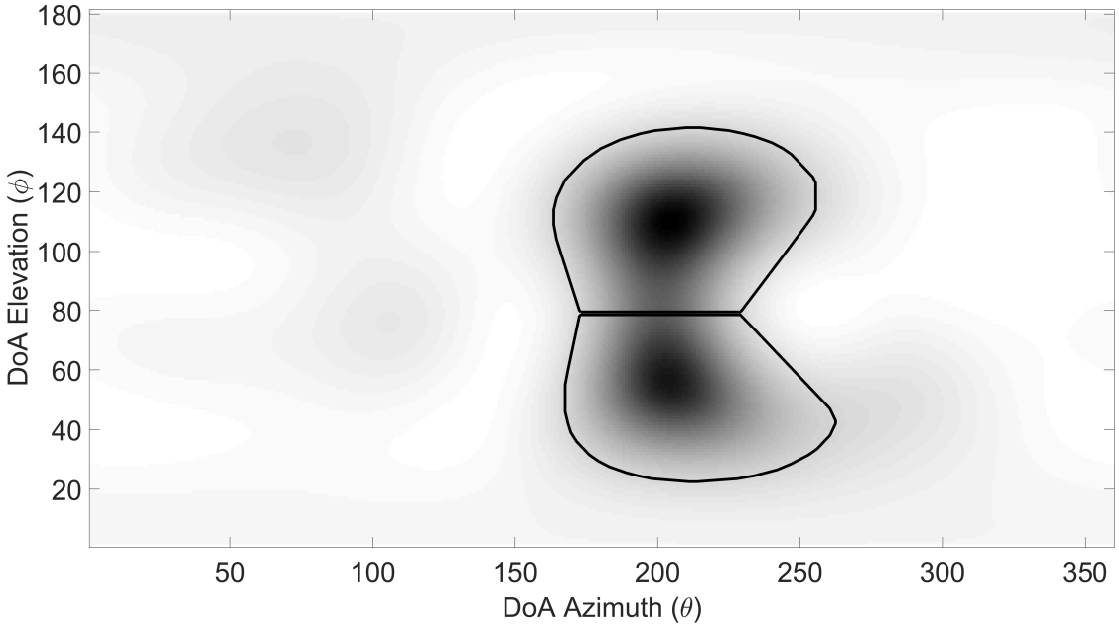

Figure 6.6: An example of the detected regions (black contours) within the directional spectrum as originally presented in Figure 6.2.

As this is an overlapping sequential process, and each reflection occupies a range of samples in the SRIR, the same reflection can be present across multiple subsequent time-frames. Therefore, each detection is either a reflection that was detected in the previous time-frame, or a new reflection. To resolve this ambiguity, the spatial region, as defined using the matrix indices representing the detected region in $\widehat{\mathbf{\Lambda}}$, for each detection within the current time-frame is compared to any detections in the previous time-frame, and if any spatial region in the current time-frame had an overlap of at least 80% with any in the previous time-frame, they are considered to have been produced by the same reflection. The value of 80% is empirically chosen to try and prevent individual reflections, arriving from close DoA, being detected as the same reflection. All reflections are therefore considered unresolved until their spatial region is no longer present in a subsequent time frame, a time-frame is skipped, or the sequential process ends. Once a detected reflection has been resolved using this process, the spatial and temporal region for the reflection is known and can be used to estimate the ToA and DoA.

Algorithm 4: Image processing object labelling for detecting connected regions of ones within a binary image. MATLAB functions are indicated in bold.

```
1  label = 1 ;                                              // Initialise the label
   // Loop over the number of rows
2  for row = 1 : noRows do
       // Loop over the number of columns
3      for col = 1 : noCols do
4          if binaryImage(row,col) == 0 then
5              continue
6          else if already checked binaryImage(row,col) then
7              continue
8          else
               // Find the connected neightbours for binaryImage(row, col)
9              testIndices = [row, col] ;          // Store the row and column index
10             while ∼isempty(testIndices) do
11                 testLocation = testIndices(1,:) ;       // Store test indices for
                    analysing connected regions
12                 if already checked binaryImage(testLocation(1),testLocation(2)) then
13                     continue
14                 end
15                 labelledImage(testLocation(1),testLocation(2)) = label; ;    // Store
                    the label identifier in the labelled image
16                 [gridIndicesY, gridIndicesX] =
                    meshgrid(testLocation(2)-1:testLocation(2)+1,
                    testLocation(1)-1:testLocation(1)+1) ;     // Define a 3 × 3 grid
17                 Remove locations in the grid that are out of bounds of the
                    binaryImage
18                 Remove locations in the grid that are equal to zero in the
                    binaryImage
19                 testIndices = [testIndices; [gridIndicesX gridIndicesY]] ;     // Store
                    indices connected to testLocation
20             end
21         end
22         label = label + 1 ;                              // Increment the label.
23     end
24 end
```

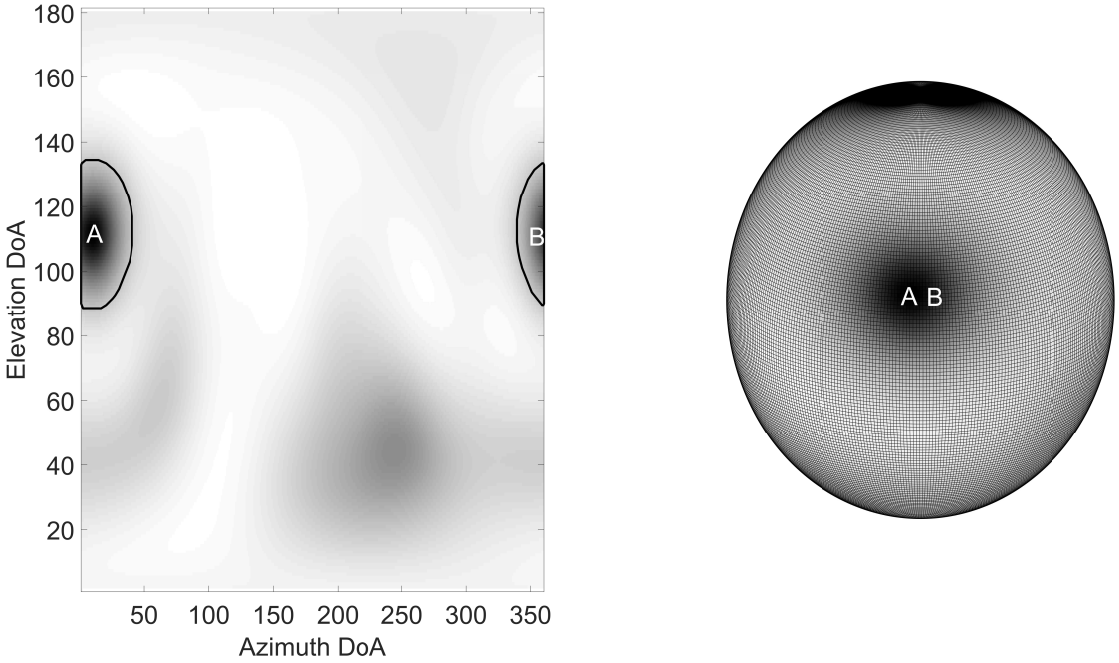

Figure 6.7: Example directional spectrum where two spatial regions (A and B) exist which belong to the same reflection. Image on the left shows the unwrapped directional spectra, and image on the right shows the directional spectra mapped to a sphere showing the overlapping region.

The DoA can be estimated from the spatial region within each time-frame for which a reflection is present. The DoA is computed by adding the directional spectrum across the reflection's time-frames, and taking the steered direction, within the reflection's spatial region, with the largest power as corresponding to the DoA of the reflection:

$$\Psi_{DoA} = \underset{\Psi}{\operatorname{argmax}} \left(\sum_{i=1}^{i=I} \mathbf{\Lambda}_i(\Psi_r) \right) \tag{6.10}$$

where $\mathbf{\Lambda_i}$ is the directional spectrum matrix for the i^{th} time-frame that the reflection is present, r defines the sub-array indices in Ψ that define the spatial region, I is the total number of time-frames over which the reflection is present, and $\operatorname{argmax}(...)$ outputs the steered direction with highest power value.

6.3.1 Time-of-Arrival Estimation

As is the case with previous work [8, 14, 15, 64, 96, 100, 101] it is assumed that the arrival of a reflection at the receiver array is represented by a peak within the SRIR. However, searching for the maximum peak within the temporal region of the reflection does not account for the case when multiple reflections are present, as the maximum peak does not necessarily relate

to the desired reflection. Therefore, the ToA of the desired reflection is estimated by steering the response of the microphone array to the direction of the DoA of the arriving reflection, and searching for the maximum peak index in the resulting signal. This is expressed as:

$$\tau = \underset{\tau \in \{\tau_{st}, \cdots, \tau_{ed}\}}{\operatorname{argmax}} \left(\left| \sum_{n=1}^{(N+1)^2} \mathbf{H}_n(\tau_{st} : \tau_{ed}) \mathbf{y}_n(\Psi_{DoA}) \right| \right) \tag{6.11}$$

where τ_{ref} is the sample range occupied by the reflection, τ is the ToA for the given reflection, Ψ_{DoA} is the azimuth and elevation steering direction defined by the DoA of the reflection, and $\operatorname{argmax}(...)$ returns the peak index $\tau \in \{\tau_{st}, \cdots, \tau_{ed}\}$. Where τ_{st} defines the start of the time-frame where the reflection is first detected, and τ_{ed} defines the first sample of the time-frame where the reflection is no longer present. An overview of the EDESAR method for reflection detection is presented in Algorithm 5.

While this method is theoretically capable of detecting multiple reflections within a single time-frame, there is a special case when this is not true. This special case is when the method is being used to analyse SRIRs generated using geometric modelling. When considering geometric acoustic modelling, a reflection is added to the SRIR using a filtered Dirac delta, and the consequence of this is that the reflections are all highly-correlated. This results in a covariance matrix for each time-frame that is rank-deficient [50, 51], the implication of this being that the MVDR beamformer will be less able to disambiguate between multiple arrivals [61]. An example of this can be seen in Figure 6.8 where multiple regions of higher power exist that are close to the residual signal power spread throughout the directional spectrum. The thresholding applied to the directional spectrum in this case would result in no reflections being detected, and removing the thresholding would result in multiple false positive detections.

To remove the rank requirement of the covariance matrix, an alternative beamforming technique can be used for simulated data. DoA estimation analysis for reflections in SRIR measured with a spherical microphone array in [12], shows that EB-MUSIC was the next best beamformer when compared with a delay-and-sum, plane wave decomposition, and MVDR beamformer. However, EB-MUSIC also requires a full-rank matrix to work effectively [50, 51], and therefore the next best beamformer, the plane wave decomposition beamformer (as implemented in [30]), will be used instead, expressed as,

Algorithm 5: Overview of the EDESAR algorithm. MATLAB functions are indicated in bold.

Input: *SRIR* The spatial room impulse response.

Output: *reflections* structure containing information about each detected reflection.

```
// Initialisation
frameLength = 20 ;                  // Define the length of the analysis window
stepSize = frameLength / 2 ;        // Define the step size between sucessive time frame
noFrames = floor(length(SRIR) / stepSize) - (floor(frameLength/stepSize)-1) ;
// Define the the total number of time-frames
for ii = 1 : noFrames do
    Window out the iith time-frame in the SRIR using a Hann window.
    Filter the windowed time-frame and adjust peak location to account for filter latency.
    if all(max(abs(unnormalisedAnalysisFrame)) < 0.01) —— any(diffusenessProfile > 0.3) then
        Store all unresolved reflections with τ_ed = frame start and compute the ToA
        Increment frame start and end indices and continue to next iteration of the for loop.
    else
        Compute and process the directional spectrum. Algorithm 3.
        Detect any spatial regions in the directional spectrum. Algorithm 4.
        If there is more than one spatial region detected, check if they are artefacts of analysing the unwrapped sphere. If so combine their convex hulls.
        Compute DoA for each region Equation 6.10.
        if Any detected spatial regions existed in the last time-frame then
            Store region information with corresponding previous time-frame
        else
            Define new reflection.
        end
        if Any regions in previous time-frame were not in the current time-frame then
            Store all unresolved reflections and compute the ToA
        end
    end
end
Store all unresolved reflections with τ_ed = frame end and compute the ToA
```

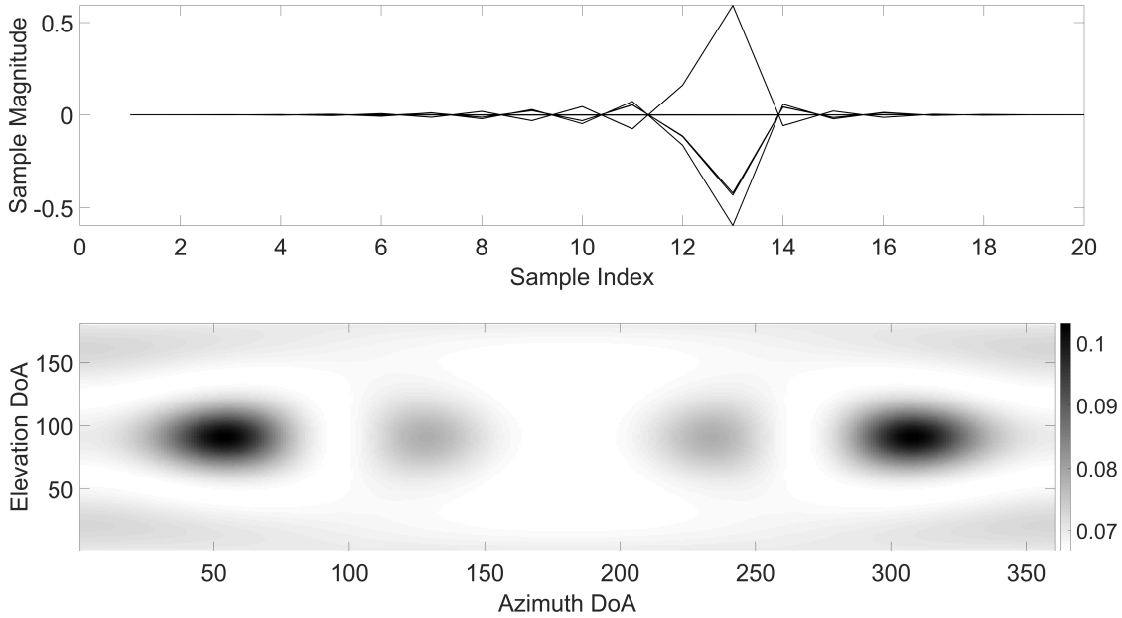

Figure 6.8: Example of the MVDR beamformer's output when given a rank-deficient covariance matrix for a time-frame of a spatial room impulse response, simulated using CATT-Acoustic, where two reflections are present, with DoA at $\theta = 69°$ $\phi = 90°$ and $\theta = 310°$ $\phi = 90°$. As can be seen there is minimal difference between the residual directional power and the desired reflections. Furthermore, it can be seen that there are at least four distinct regions.

$$\zeta(\Psi) = \left(\frac{4\pi}{M}\mathbf{y}(\Psi)\right)\mathbf{R}_{\mathbf{H}}(t_f)\left(\frac{4\pi}{M}\mathbf{y}(\Psi)\right) \tag{6.12}$$

where $\mathbf{R}_{\mathbf{H}}(t_f)$ is the covariance matrix of the SRIR at time-frame t_f and M is the number of microphones in the array. The resulting directional spectrum for the analysis frame in Figure 6.8 when using the steered-response power map can be seen in Figure 6.9.

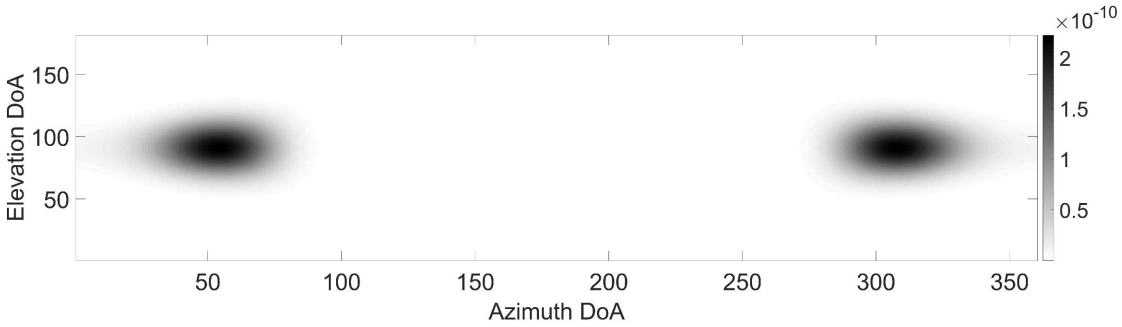

Figure 6.9: Example of the steered-response power map output when given a rank-deficient covariance matrix for a time-frame of a spatial room impulse response, simulated using CATT-Acoustic, where two signals are present. As can be seen there is a larger difference between the residual directional power and the desired reflection compared to Figure 6.8. Furthermore, it can be seen that there are are now only two distinct regions.

6.4 Testing

The threshold values used have been defined empirically by considering the method as applied to different signal types, by observing the resulting detected reflections and adjusting the thresholds to minimise the number of false-positive detections and inaccurately estimated reflection parameters. The sample magnitude threshold $\epsilon_\alpha = 0.01$, is defined as the sample magnitude of the SRIR where no distinct reflections could be visually observed (within the diffuse field). The diffuseness threshold $\epsilon_d = 30\%$ was defined to be at a point where the residual noise component starts to be detected as a reflection producing false-positives, while still ensuring the main discrete reflections are detected. Finally the mask threshold $\epsilon_{msk} = 0.1$ was chosen through observation of the directional spectrum's darkest regions and their power value relative to the residual power value present in the remainder of the spectrum.

To test the accuracy of the proposed method three scenarios using simulated and real-world SRIRs are considered. Scenario 1 will use a simple impulse train where the ToA of every reflection is exactly known. Scenario 2 will use a SRIR simulated using CATT-Acoustic [16] to test the performance of the method for an example where no noise is present. Finally Scenario 3 will use real-world SRIR measurements in a cuboid-shaped room. As the DoA of the reflections is only known for Scenario 1, and results in [12] have already shown that the DoA of reflection can be accurately estimated using these beamformers, the results presented will focus on the use of this method for reflection detection, and will be compared to implementations of the Dynamic Time Warping (DTW) based matching pursuit method [15] and the Circular-Variance Local-Maxima (CVLM) method [14]. As the original implementation for these baseline methods could not be obtained, the author has developed implementations based on the original papers. As with [96] a candidate detection is considered valid if it is within 0.5 ms of the expected ToA, and for each ToA only one detection is validated, that being the one closest to the expected ToA.

Scenario One: Randomly Generated Train of Pulses

Generally initial testing of a reflection detection algorithm will use a train of impulses with known ToA [15], providing a highly controlled test scenario. Each impulse represents the arrival of a reflection with a randomly generated ToA, DoA and amplitude.

In this particular case the direct sound component from an omnidirectional (0th order spherical harmonic) channel of a real-world SRIR measurement, captured using an EigenMike [13],

Genelec 8030 [139], and the exponential sine sweep method [140], is used to generate the 5.89 ms pulse used , as shown in Figure 6.10. As discussed in Chapter 3 one of the main issues with existing techniques is whether overlapping reflections can be individually detected and to test this, the ToA of each impulse is defined, using the peak location of each impulse within the train, such that there is a minimum time-difference-of-arrival beyween subsequent reflections of 3.17 ms and a maximum of 8.61 ms – with at most a 4.53 ms overlap between subsequent pulses. The azimuth DoA of each reflection is randomly generated, with the constraint that adjacent reflections can not have a DoA within 10° of each other.

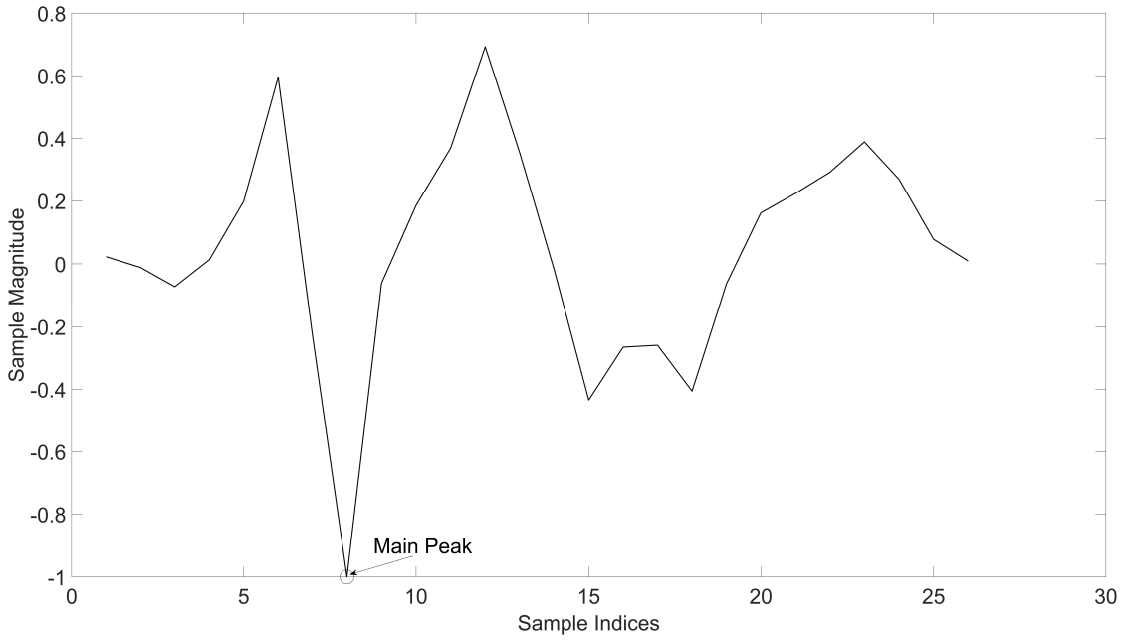

Figure 6.10: The direct sound extracted the zeroth-order component of a real-world spatial room impulse response measurement. This is used to generate a train of pulses to test the proposed reflection detection method.

Scenario Two: Simulated Spatial Room Impulse Responses

The second test scenario uses a simulated SRIR generated for a simple cuboid-shaped room using CATT-Acoustic v.9.1a. The dimensions of this test room are 4 m $\times$ 4 m $\times$ 3.5 m, and the simulation parameters can be seen in Table 6.1. A plot of the geometry of the room and the source and receiver locations can be found in Figure 6.11.

To generate a related set of candidate ToAs, reflections within the SRIR are estimated and identified using an implementation of the image-source method, as defined in [17]. This allows for the number of false-positive detections to be estimated, as well as errors in the ToA estimation.

Scenario Three: Real-World Spatial Room Impulse Responses

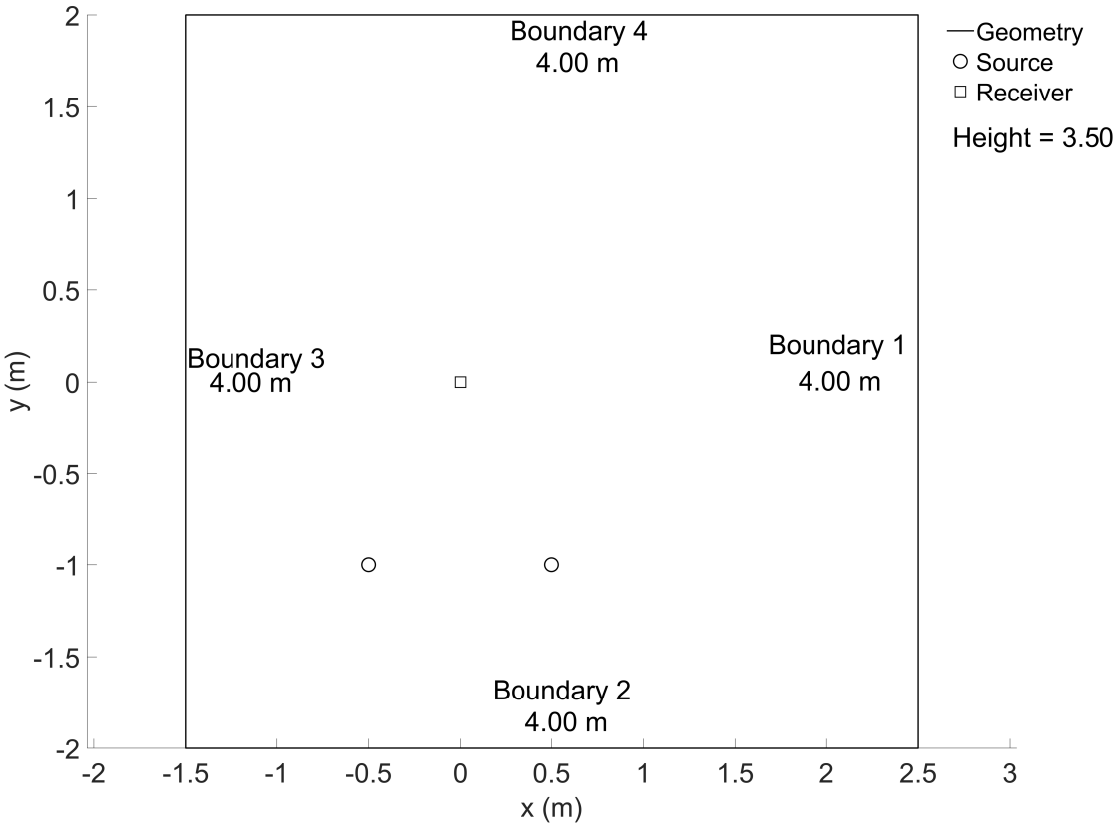

Figure 6.11: Geometry for the cuboid-shaped room used to render the CATT-Acoustic SRIR. Square marker denotes the receiver position and the circle markers denote the source positions.

Parameter	Value
Diffuse Reflections	Off
Number of Rays	10,000,000
Boundary Material	WOOD30
Source Directivity	Omnidirectional
Receiver Directivity	Omnidirectional
Rendering	Third-Order Ambisonic

Table 6.1: Simulation Parameters used to render the the CATT-Acoustic SRIR used to test the EDESAR Reflection Detection Method.

In the final scenario two real-world SRIRs measured using an EigenMike EM32 [13] and Genelec 8030 [139] loudspeaker both positioned at a height of 1.5 m from the floor to the centre of the microphone array and loudspeaker. To generate the SRIR the exponential sine sweep method from [140] is used, using a 20 s exponential sine sweep from 20 Hz to 20 kHz. As with the measurements in Chapter 5, an omnidirectional sound source is approximated by averaging the SRIR measurements over four loudspeaker orientations at 0°, 90°, 190°, and 270° as used in [141]. While this will not necessarily produce equal excitation across all angles and for all frequencies, and in particular for high frequencies where loudspeakers tend to be more directional, it has been

shown to produce a more uniform excitation of a space in similar circumstances [141]. The final SRIRs are then normalised to have a maximum sample value of ± 1, and converted to third-order spherical harmonic domain signals using MH Acoustics' EigenStudio [13].

The measurement room is cuboid-shaped with dimensions 10.35 m$\times$13.29 m$\times$4.19 m, and has a number of non-removable, adjustable, floor length curtains. As it was not possible to remove these curtains, they were positioned, as much as is possible, to limit their impact on the obtained SRIRs. Hence they were arranged in corners of the room, across windows, and, where possible, to cover features on the walls such as electrical outputs, as well as the computer and interface used for the measurements. While it is accepted that this is non-ideal, and could have some impact on the results, every effort has been made to minimize their potential influence on the measurements obtained, and ensure that the main reflective boundaries are exposed and clear from other possibly confounding features. Furthermore, the ceiling was covered in large metal piping connected to extractor fans and a layer of metal railing approximately 1 m from the ceiling. The noise floor in the room is measured as 60.2 dBA using an SPL meter and the room's temperature was 24.4°C. An image of the measurement set-up and environment can be seen in Figure 6.12, and the source and receiver positions can be seen in Figure 6.13.

As with Scenario 2 the image-source method [17] is used to generate a set of candidate ToAs. From these simulated arrival times, an approximation of where reflections could be present in the measured SRIR is obtained. It is, however, important to note that these candidate ToAs do not account for diffuse reflections that may be present, and so, detections made by the proposed method that do not align with an arrival computed using the image-source method are not necessarily false-positive detections.

6.5 Results

6.5.1 Scenario One: Randomly Generated Train of Pulses

In Figure 6.14 the results for the simulated train of impulses can be seen for EDESAR, CVLM, and DTW based matching pursuit methods. The results show that both the proposed method and the DTW based maximum likelihood method have no false-positive detections, and both only miss one reflection (one of the overlapping cases after the 400th sample). Given that these overlapping reflections have a maximum spacing of seven samples and DoA of $\theta = 30^{\circ}$ and $\theta = 80^{\circ}$ respectively, it is possible that during the summation process these reflections interact with each other as a result of phase differences introduced by the spherical harmonic vector for

Figure 6.12: Image of the room setup used in Scenario Three, showing the Genelec 8030 and EigenMike. As can be seen there is curtain coverage across the right wall which occludes the windows, and curtains positioned in the corners of the room hiding large electrical outlets. On the ceiling there are light fixtures, railing, extractor fans, and a series of large rectangular pipes.

each impulse, therefore, making them harder to detect individually.

From Table 6.2, it can be seen that the EDESAR and DTW approaches have comparable maximum and minimum ToA error, however, based on the RMS error, the EDESAR approach is generally more accurate. This result suggests that in this case the EDESAR methods is generally more accurate by 45.71 μs. The results also show that the CVLM method is least accurate with 18 false-positive detections, a maximum ToA error of 476.19 μs (249.43 μs more than EDESAR) and an RMS ToA error of 184.89 μs (137.39 μs more than EDESAR).

Method	Max ToA Error (μs)	Min ToA Error (μs)	RMS ToA Error (μs)	False-Positives
EDESAR	$\pm$226.76 μs	0 μs	$\pm$47.50 μs	0
CVLM	$\pm$476.19 μs	0 μs	$\pm$184.89 μs	18
DTW	$\pm$226.76 μs	0 μs	$\pm$93.21 μs	0

Table 6.2: Reflection detection results for the proposed EDESAR method, CVLM, and DTW based maximum likelihood for the randomly generated train of pulses. Results show the maximum time-of-arrival error, minimum time-of-arrival error, RMS time-of-arrival error, and number of false-positive detections.

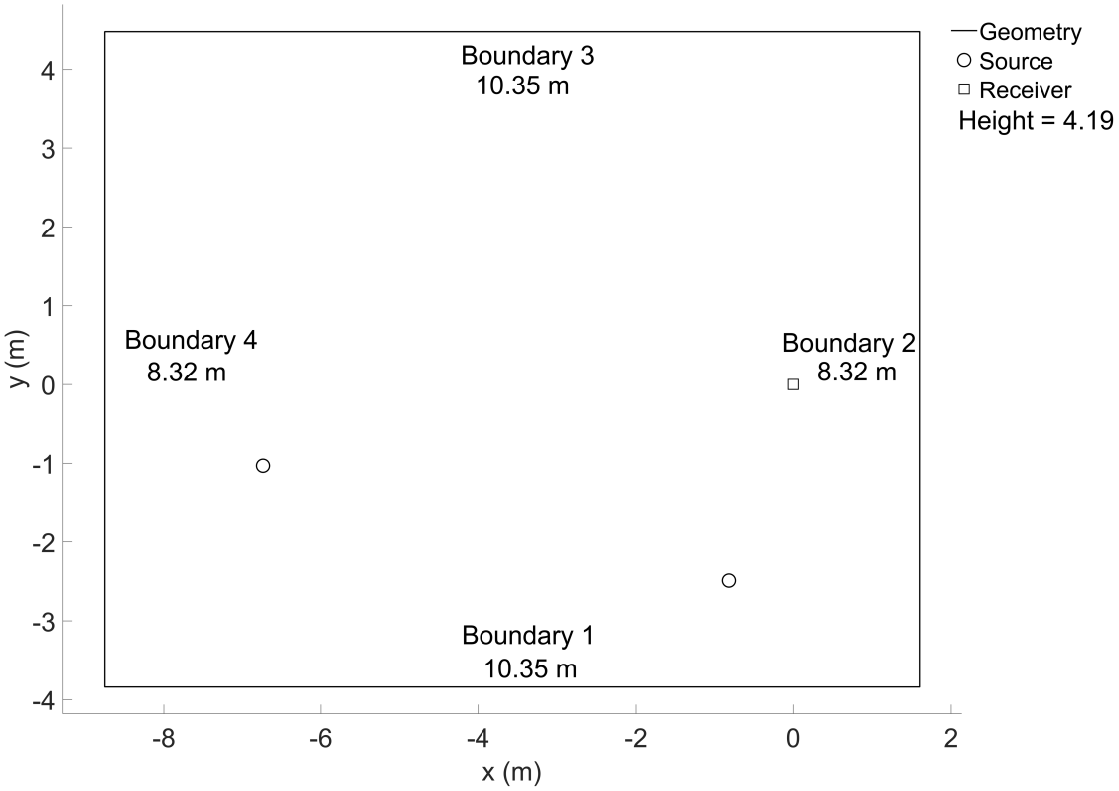

Figure 6.13: Geometry for the cuboid-shaped room used in the real-world measurements. Square marker denotes the receiver position and the circle markers denote the source positions.

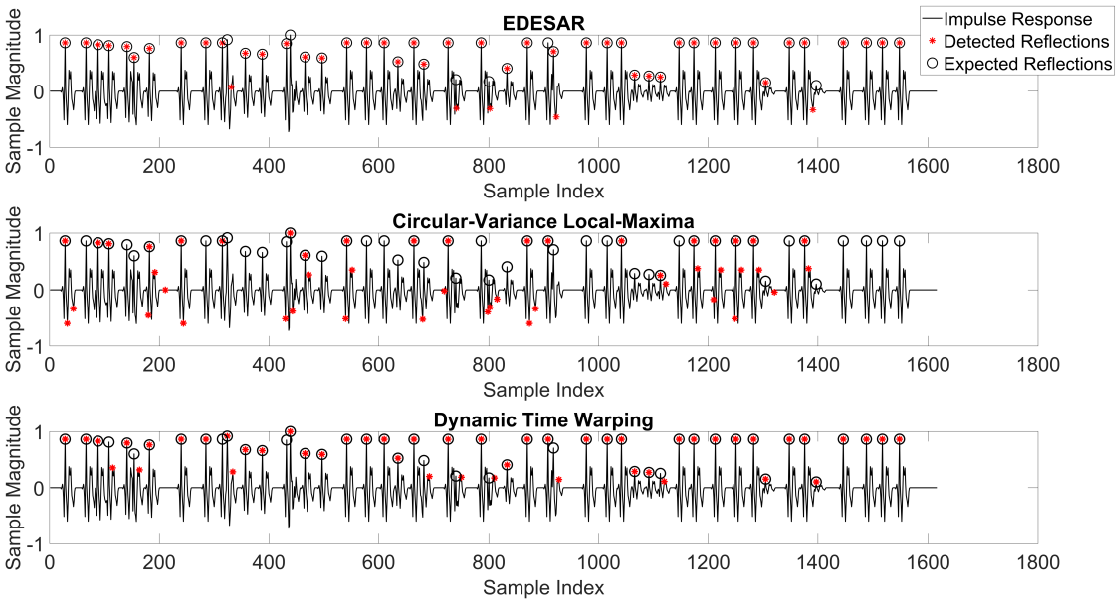

Figure 6.14: Comparison between proposed method (top), CVLM technique (middle), and DTW reflection detection technique (bottom), using the first randomly generated SRIR. The circles indicate the correct time-of-arrival for a reflection, and the red asterisks denote the estimated time-of-arrival.

6.5.2 Scenario Two: Simulated Spatial Room Impulse Responses

In Figure 6.15, the results for the CATT-Acoustic simulated SRIR can be seen for the EDESAR, CVLM, and DTW based matching pursuit methods. The results show that both the EDESAR and DTW approaches have detected the main reflections within the SRIR, while the CVLM approach has detected fewer reflections. It is interesting to note that while the detections at samples 563, 736, and 809, do not align with an expected ToA these regions in the SRIR have signals that look like previous reflections in the SRIR. CATT-Acoustic computes the SRIR using a combination of the image-source method and ray-tracing, with in this case 10,000,000 rays used to compute paths through an environment. Reflection paths that pass through a sphere around the receiver (dimensions of which are not known) define the arrival of a reflection [152], resulting in reflection paths that cannot be defined exactly using the image-source method alone.

From the results in Table 6.3, it can be seen that while the EDESAR method has the most correctly detected reflections, it has a larger RMS ToA error of 168.89 μs. Furthermore, the EDESAR method has detected all first order reflections with at most a 22.68 μs ToA error compared to 181.41 μs for the CVLM method, and 45.35 μs for the DTW approach. Across all of these methods the ToA estimates become more inaccurate as reflection order increases. The main benefit of the EDESAR method here is that it can disambiguate between simultaneous arrivals. Out of the 71 correct detections only 52 unique ToA values are present and this means that 28 of these detections belong to a reflection that arrive at the same time as at least one other reflection, as validated with the image-source method's predicted reflections. These simultaneously arriving reflections are detected as a single arrival with both the CVLM and DTW methods.

Method	Max ToA Error (μs)	Min ToA Error (μs)	RMS ToA Error (μs)	Correct Detections	False-Positives
EDESAR	$\pm$430.83 μs	0 μs	$\pm$168.89 μs	71	14
CVLM	$\pm$430.83 μs	0 μs	$\pm$144.58 μs	29	4
DTW	$\pm$362.81 μs	0 μs	$\pm$98.71 μs	60	66

Table 6.3: Reflection detection results for the proposed EDESAR method, CVLM, and DTW based maximum likelihood for the CATT-Acoustic simulation of a cuboid-shaped room. Results show the maximum time-of-arrival error, minimum time-of-arrival error, RMS time-of-arrival error, number of correct detections, and number of false-positive detections.

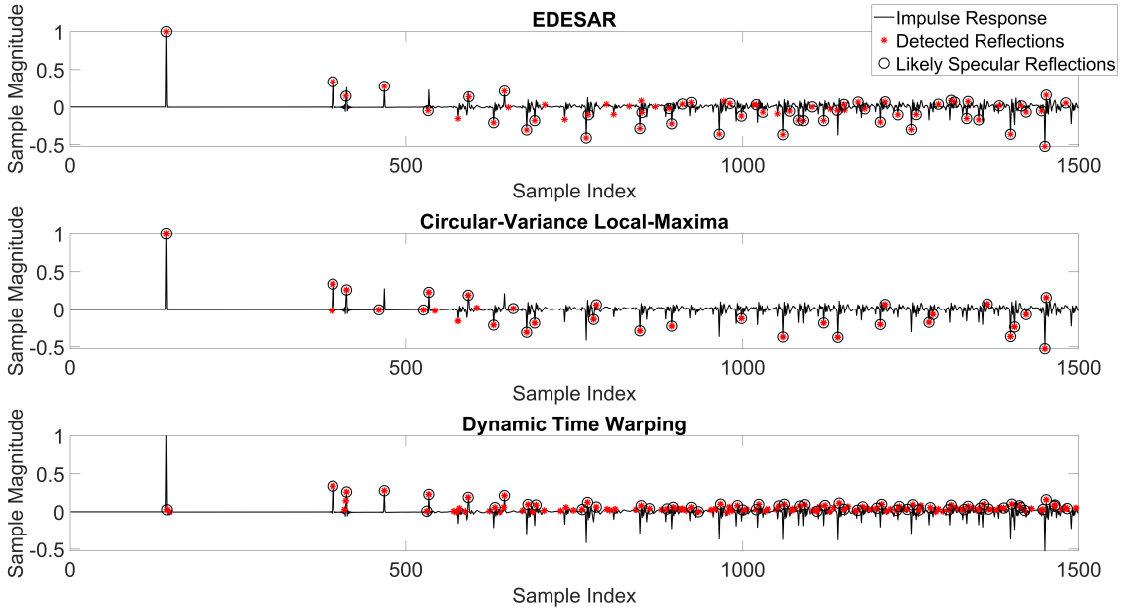

Figure 6.15: Comparison between proposed method (top), CVLM technique (middle), and DTW reflection detection technique (bottom), using a simulated SRIR. The black solid line is the omnidirectional zeroth order spherical harmonic domain channel of the SRIR, red asterisks denote a detection made by the methods, and the black circles denote the correctly detected reflections.

6.5.3 Scenario Three: Real-World Spatial Room Impulse Responses

For real-world measurements it becomes harder to compare between algorithms, is as a result of the reflected arrivals in the signal not being explicitly known. While acoustic modelling can be used to approximate the arrival of specular reflections, there is no guarantee that each modelled reflection arrives at the microphone array, and it does not account for diffuse reflections. The implication here being that false-positive detection could also be a diffuse reflection, as these reflection in the measured SRIR will not necessarily align with a candidate ToA.

In Figure 6.16 the results for the first real-world SRIR can be seen for the EDESAR, CVLM, and DTW based matching pursuit methods. It can be seen that both the proposed EDESAR method and the DTW approach have detected the main peaks in the SRIR. The EDESAR algorithm has false positive detections at samples 971, 984, 985, 986, 999, 1001, and 1044, which are as a result of the same reflection being detected multiple times. This is as a result of differences in the spatial-width of this reflection over time-frames, which produces a detected spatial region across time-frames that do not overlap by 80%. An example of this can be seen in Figure 6.17, for the third and fourth detection in the example SRIR where the fourth detection only has a 42.95% overlap in detected spatial region with the third detection. Each of the seven detections mentioned above have DoA estimates produced by EDESAR, and are result of two points of

reflection: the floor and possibly the pipes, extractor fans, or railing on the ceiling. The estimated DoA for these detections are $[\theta = 203°\phi = 116°]$, $[\theta = 202°\phi = 111°]$, $[\theta = 204°\phi = 114°]$, $[\theta = 208°\phi = 45°]$, $[\theta = 204°\phi = 109°]$, $[\theta = 203°\phi = 56°]$, and $[\theta = 210°\phi = 36°]$.

From the results in Table 6.4, it can be seen that the CVLM method has detected the most correct reflections, however, this method was unable to detect the direct sound and the first two first-order reflections, as a result of a large circular variance for these time-frames. Both the EDESAR and DTW methods have detected all first order reflections, with a maximum ToA error of 90.70 μs for EDESAR and 340.13 μs for the DTW approach. Out of the 48 detections made by the EDESAR method there are 45 unique ToAs, and therefore six of these detected reflections arrive at the same time as another reflection, which the other two methods have detected as a single arrival.

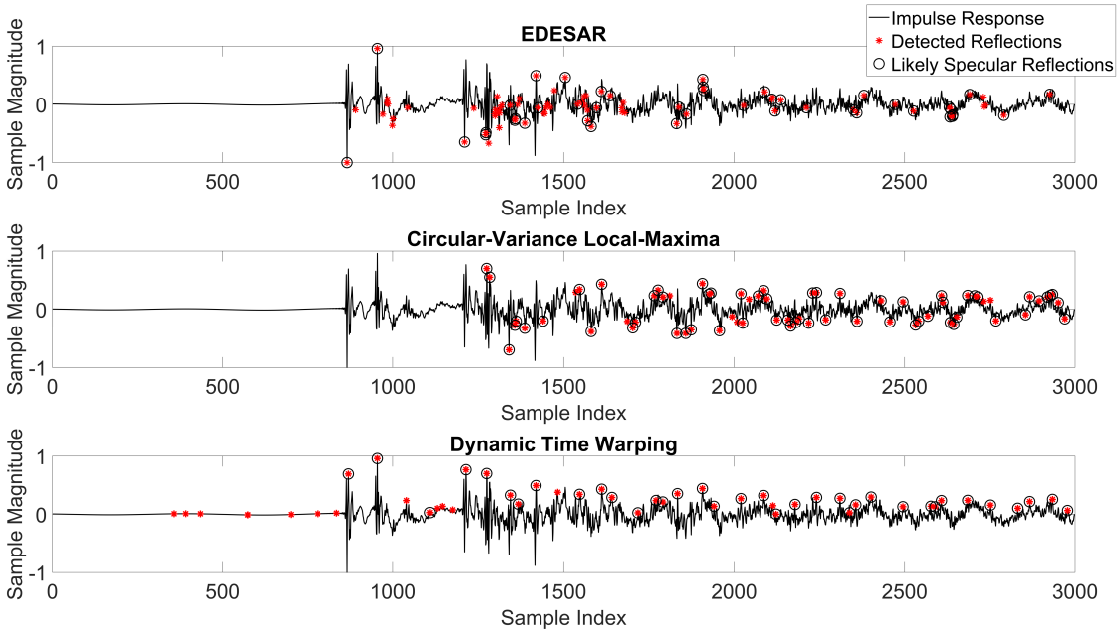

Figure 6.16: Comparison between proposed method (top), CVLM technique (middle), and dynamic time warping reflection detection technique (bottom), using the first real-world SRIR. The black solid line is the omnidirectional zeroth order spherical harmonic domain channel of the SRIR, red asterisks denote a detection made by the methods, and the black circles denote the correctly detected reflections

In Figure 6.18 the results for the second real-world SRIR can be seen for the EDESAR, CVLM, and DTW based matching pursuit methods. It can be seen that in this case the EDESAR method has detected the least correct reflections, with multiple detections of the same reflection as was seen in the previous case. The additional detections made by the CVLM and DTW methods are after the 1500 sample index, where the EDESAR method does not detect any reflections, as the normalised sample magnitude drops below the threshold of 0.01. Considering only the

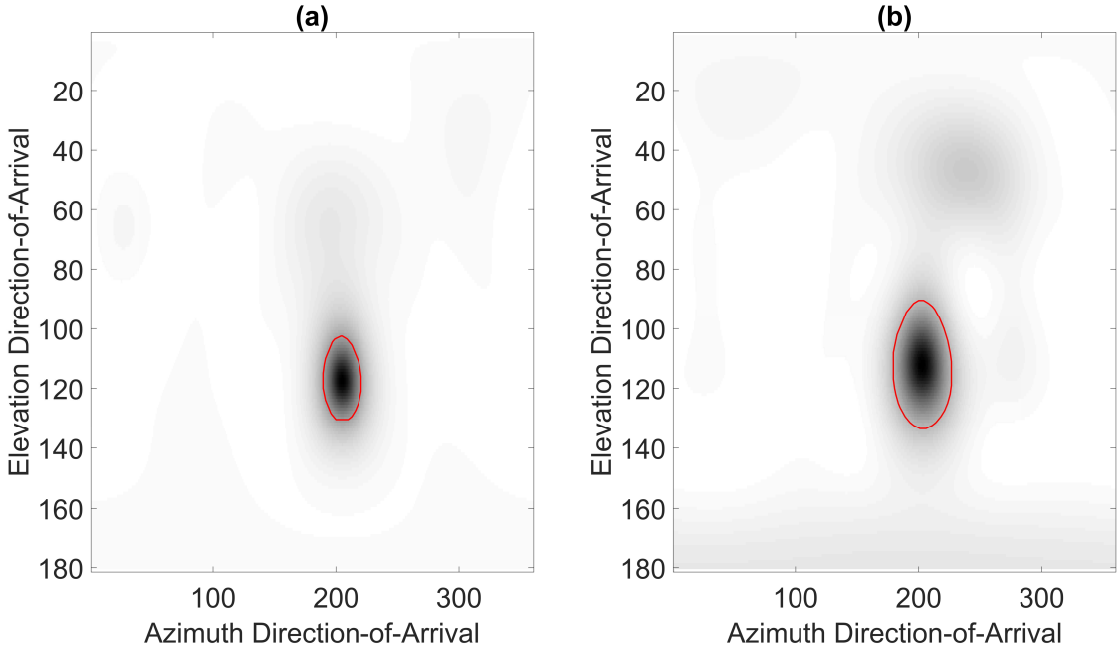

Figure 6.17: Comparison between (a) spatial region produced by the MVDR beamformer for the third detection made by the EDESAR method for the first real-world SRIR, and (b) the spatial region for the fourth detection. It can be seen that the detected spatial region, outlined in red, extracted for the fourth detection is larger than that of the third with only 42.95% overlap, causing these to be detected as two separate reflections.

Method	Max ToA Error (μs)	Min ToA Error (μs)	RMS ToA Error (μs)	Correct Detections	False-Positives
EDESAR	$\pm$476.19 μs	0 μs	$\pm$198.20 μs	48	43
CVLM	$\pm$476.19 μs	0 μs	$\pm$200.14 μs	61	8
DTW	$\pm$453.51 μs	0 μs	$\pm$206.17 μs	36	14

Table 6.4: Reflection detection results for the proposed EDESAR method, CVLM, and DTW based maximum likelihood for the first real-world measurement. Results show the maximum time-of-arrival error, minimum time-of-arrival error, RMS time-of-arrival error, number of correct detections, and number of false-positive detections.

detections within the first 1500 samples, the CVLM method has 17 correct detections, but does not detect the direct sound, and the DTW approach made 12 correct detections, which are fewer than the number of correct detections made by the proposed EDESAR method. All three approaches have detected all first-order reflections, however, the EDESAR method in this case has the largest ToA error for first-order reflections 340.13 μs, compared to 45.35 μs and 90.70 μs for the CVLM and DTW methods respectively. Out of the 21 detections made by the EDESAR method in this case, four detections belong to simultaneously arriving reflections that have been individually detected, and validated using the image-source method.

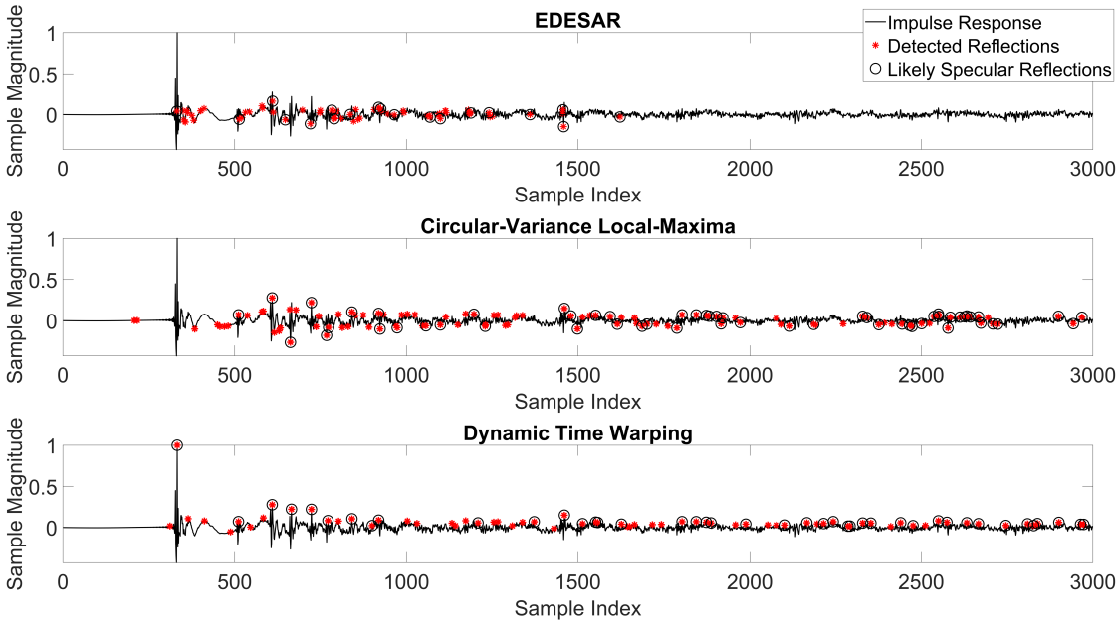

Figure 6.18: Comparison between EDESAR (top), CVLM technique (middle), and dynamic time warping reflection detection technique (bottom), using a second real-world SRIR. The black solid line is the omnidirectional zeroth order spherical harmonic domain channel of the SRIR, red asterisks denote a detection made by the method in question, and the black circles denote correctly detected reflections.

Method	Max ToA Error (μs)	Min ToA Error (μs)	RMS ToA Error (μs)	Correct Detections	False-Positives
EDESAR	$\pm$362.81 μs	0 μs	$\pm$184.73 μs	21	57
CVLM	$\pm$476.19 μs	0 μs	$\pm$176.16 μs	56	56
DTW	$\pm$476.19 μs	0 μs	$\pm$218.25 μs	42	28

Table 6.5: Reflection detection results for the proposed EDESAR method, CVLM, and DTW based maximum likelihood for the second real-world measurement. Results show the maximum time-of-arrival error, minimum time-of-arrival error, RMS time-of-arrival error, number of correct detections, and number of false-positive detections.

6.6 Discussion

The results presented in this chapter have shown that the proposed EDESAR method outperforms both the DTW based matching pursuit and CVLM when analysing simulated SRIRs, with a larger number of correctly detected reflections, and the lowest error in ToA estimates for first-order reflections. Both the proposed EDESAR method, the CVLM, and DTW based reflection detection methods decrease in accuracy when analysing real-world data. While the CVLM method detected the most correct reflections for both real-world scenarios, it was unable to detect the direct sound in both examples and two first-order reflections in the first test SRIR. For both real-world SRIRs the EDESAR method produced better or comparable estimates of ToA than the other two approaches with only an 8.57 μs difference in RMS ToA error for the second

real-world SRIR. The main benefit of the EDESAR method is the ability to detect simultaneously arriving reflections, as discussed for Scenario 2 and 3 where 28, 6, and 4 simultaneously arriving reflections were detected as individual discrete arrivals. These simultaneously arriving reflections were always detected as a single arrival by both the CVLM and DTW methods.

The key issue with the EDESAR method is that, for real-world measurements, reflections are sometimes detected multiple times. As such these detections will have the same DoA, and so assigned to the same boundary when performing geometry inference, and is not therefore so much of a concern when applied in this way. Furthermore, as a result of the generally improved or comparable ToA estimation accuracy, fewer detrimental false-positive detections, and the ability to detect simultaneously arriving reflections, which as a consequence relaxes constraints on source and receiver positioning, the EDESAR method is a more appealing option for geometry inference.

6.7 Conclusions

In this chapter the Eigenbeam Detection and Evaluation of Simultaneously Arriving Reflections (EDESAR) method for reflection detection was presented. This method performs spatiotemporal decomposition by applying a spherical beamformer to short time-frames of a SRIR. This generates a heat map of directional intensity - the directional spectrum. Any reflections present in a time-frame are then detected as regions of high-intensity within this directional spectrum. Comparisons between the proposed method and implementations of two state-of-the-art techniques, the circular variance local maxima and dynamic time warping based matching pursuit methods, showed that generally the proposed method produced more accurate estimations of ToA with a maximum RMS error of 198.20 μs across the two real-world measurements compared to 200.14 μs for the circular variance local maxima method and 206.17 μs for the dynamic time warping based approach. One key issue with the EDESAR method when analysing real-world measurements was the detection of the same reflection multiple times. As these detections belonged to the same reflection, the estimated direction-of-arrival are all similar, and therefore these detections will all be assigned to the same boundary when performing geometry inference, and as such do not present a problem. The main benefit of this approach, compared to existing reflection detection methods, is its ability to detect reflections that arrive simultaneously at the microphone array as individual reflections. This maximises the number of reflections that are extractable from a single SRIR, and therefore, relaxes constraints imposed on the source and receiver positioning needed to ensure a first-order reflection from each boundary is detectable.

The EDESAR method will be used in the next chapter of this thesis to detect reflections for use in geometry inference problems.

Chapter 7

Geometry Inference of Convex and Non-Convex Rooms using Compact Microphone Arrays

7.1 Introduction

In the previous chapter a spatiotemporal decomposition based reflection detection method was outlined. This performs beamforming on short time-frames of a SRIR, to detect the arrival of directional signals, reflections, at a microphone array receiver. The results presented showed that the proposed method generally produces more accurate estimates of ToA for reflections when compared to implementations of the circular variance local maxima [14] and the dynamic time warping based maximum likelihood [15] reflection detection methods. However, when analysing real-world measurements the proposed method occasionally detected the same reflection multiple times as a result of change in the area of the spatial region that the reflection occupied over subsequent time-frames. It was argued that in context of the geometry inference this does not pose an issue, as these detections will be assigned to the same boundary.

Geometry inference refers to the inverse problem of estimating the locations of reflective boundaries within an environment from the reflections captured across a number RIRs. This analysis technique exploits the temporal, and sometimes spatial, information contained within these (spatial) room impulse response to estimate reflection paths, and therefore boundary locations. Current-state of the art methods constrain the problem to that of a convex-shaped room, simpli-

fying the problem, as a result of requiring a reflection from each boundary being detectable in all or a subset of Spatial Room Impulse Responses (SRIRs) measured at different source and receiver positions. The consequence of this is that these methods are not applicable to non-convex rooms, as in some cases it is not possible to position source/receiver combinations to detect multiple reflections from a boundary. The problem of geometry inference for non-convex-shaped rooms is therefore more complex, as no assumptions can be made about the number of boundaries or shape of the room. Therefore, to infer the geometry of non-convex-shaped rooms only a single source and receiver position should be used to infer a boundary's location. This can be achieved through the use of a compact microphone array capable of capturing both ToA and DoA for a reflection, which can be used to estimate the location of a candidate image-source. Furthermore, by relaxing assumptions on the shape of the room and required positioning of measurement locations, no constraints can be imposed on the number of reflections extracted from a single SRIR, as commonly done within the literature.

This chapter will be presented as follows: Section: 7.2 will present the problem domain, Section: 7.3 will present the Acoustic Reflection Cartographer (ARC) method, Section: 7.4 will describe the testing procedures, Section: 7.5 will present the findings, Section: 7.6 will discuss the results in the context of the literature, and Section: 7.7 will conclude the paper.

7.2 Problem Formulation

As discussed in Chapter 2, the image-source method computes the ToA of reflections in a RIR by computing the locations of an image-source by mirroring the source, and subsequent image-sources, perpendicularly across each boundary within the room. The distance between the image-source and the receiver then defines the ToA. These image-sources $\widetilde{\mathbf{s}}$ can be computed using the location of the source/image-source $\mathbf{s}$, a point on the boundary, $\mathbf{b}$ and the boundary's unit normal $\mathbf{n}$ as in [6],

$$\widetilde{\mathbf{s}} = \mathbf{s} + 2 < \mathbf{b} - \mathbf{s}, \mathbf{n} > \mathbf{n} \tag{7.1}$$

where $< ., . >$ denotes dot product. Image-reversion techniques exploit this relationship by estimating the image-source $\widetilde{\mathbf{s}}$ from the measured RIR. The boundary location can then be estimated from the image-source, and the previous-source that was mirrored in the boundary to produce the image-source, by exploiting the properties of the image-source method. That is, as

an image-source is produced by mirroring the previous-source perpendicularly across a boundary, the distance from previous-source-to-boundary and boundary-to-image-source are equal, and the line between the previous-source and image-source is parallel to the boundary's normal, as seen in Figure 7.1. A point on the boundary $\widetilde{\mathbf{b}}$ and the boundary's normal $\widetilde{\mathbf{n}}$ can therefore, from [6], be estimated as,

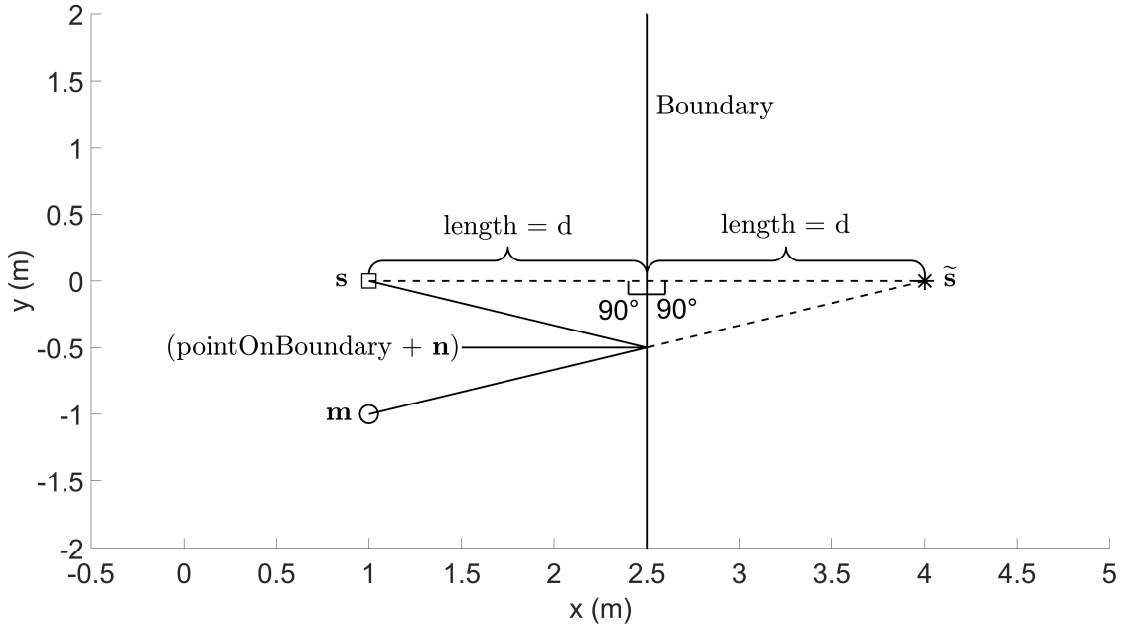

Figure 7.1: Example of an image-source produced by mirroring the source perpendicularly across the boundary. As can be seen the reflection path produced is specular with the angle of reflection relative to the normal of the plane equal to the angle of incidence.

$$\widetilde{\mathbf{b}} = \frac{\widetilde{\mathbf{s}} + \mathbf{s}}{2} \tag{7.2}$$

$$\widetilde{\mathbf{n}} = \frac{\widetilde{\mathbf{s}} - \mathbf{s}}{||\widetilde{\mathbf{s}} - \mathbf{s}||} \tag{7.3}$$

In practice not every image-source is defined using the location of the source $\mathbf{s}$, as such, one of the stages in the image-source reversion process is to find the most likely previous-source, which is substituted for $\mathbf{s}$ in (7.2). This process produces a set of candidate boundaries for the room, which either define the geometry of the room or require refining to a subset of candidate boundaries that define the room.

As discussed in Chapter 4 geometry inference methods impose certain constraints to simplify the problem and based on the work discussed in Chapter 4 the following assumptions are made with the geometry inference method presented in this chapter.

- The relative position of all source and receivers are known.

- It is assumed that the source-to-receiver distance is known *a priori* to account for any measurement system latency.

- Knowledge or room temperature is known to allow speed-of-sound to be estimated.

- It is assumed that the walls are perpendicular to the floor and ceiling, and the floor and ceiling are parallel to each other.

- That all reflections have a dominant specular component allowing their reflection paths to be traced.

- Each boundary has at least one first-order reflection assignable to and detectable in at least one SRIR.

- In this study an empirically defined minimum source/receiver-boundary distance of 50 cm is used (half that of the minimum recommended distance of 1 m in [122] to allow for analysis of smaller/complex rooms). This constraint is imposed to ideally improve the methods robustness to false-positive detections, where boundaries inaccurately inferred close to the source or receiver can lead to desired boundaries being invalidated by the proposed boundary validation process.

- The inferred boundaries define a closed geometry.

In this chapter it is assumed that a method such as the Eigenbeam Detection and Evaluation of Simultaneously Arriving Reflections (EDESAR) method for reflection detection as (see Chapter 6) has been used to analyse reflections present in the SRIR. Therefore, the presentation of the proposed method assumes that a set of candidate reflections with estimated ToA and DoA have already been detected. These candidate detections are first organised in descending ToA, and in the case of multiple measurement positions all candidate detections across the SRIRs are grouped together and sorted - making sure that each reflection's receiver location is stored.

7.3 Method

The proposed Acoustic Reflection Cartographer (ARC) method is an image-source reversion method, consisting of two processing steps, image-source reversion and geometry validation.

The ToA and DoA for each candidate reflection is used to compute the location of the image-source that produces a given reflection. It is important to note that any error in the ToA and DoA estimates will result in a less accurate estimate of the boundary location, the proposed method assumes that these estimated values are accurate and does not attempt to account for estimation error. This is an iterative process that is performed in reverse order, prioritising the first arrivals at the receiver, and making it easier to remove false-positive detections without disrupting the loop iterator. From these estimated image-source locations, the most-likely previous-sources are searched for, and a set of candidate boundaries defined using these image-source/previous-source pairs. The geometry validation process is then used to refine the candidate boundaries to ideally retain only those that pertain to the given measurement environment. An overview of this whole process can be seen in Figure 7.2

7.3.1 Image-source Reversion

For each candidate detection an estimated ToA and DoA value will be extracted from the SRIR. Assuming that the first arrival at the microphone array belongs to the direct sound and all subsequent detections are reflections, the source location (if not known *a priori*) and image-source locations can be defined using directional cosines, from [153] as,

$$\widetilde{\mathbf{s}}_i = \mathbf{m} + d_i \begin{bmatrix} \sin(\phi_i)\cos(\theta_i) \\ \sin(\phi_i)\sin(\theta_i) \\ \cos(\phi_i) \end{bmatrix} \tag{7.4}$$

where $\mathbf{m}$ is the $[x, y, z]$ coordinate for the receiver and d_i is the distance travelled by the i^{th} detection, computed as ToA $* c$, where c is the speed of sound. To define the parameters for the candidate boundaries (7.2), the most-likely previous source for each image-source needs to be found.

When searching for the most-likely previous sources it is important to consider that each image-source is either produced by a first-order reflection from a new or existing boundary, a higher-order reflection from an existing boundary, or a false-positive detection. Following the definition of the SRIR in Chapter 2, it can be assumed that the first two detections that produce a valid boundary based on the above assumptions, can be defined as first-order reflections, this caveat is used to bootstrap the process. Furthermore, it is assumed that the first detection that can produce either the floor or ceiling for each source/receiver pair is first-order and that the mean

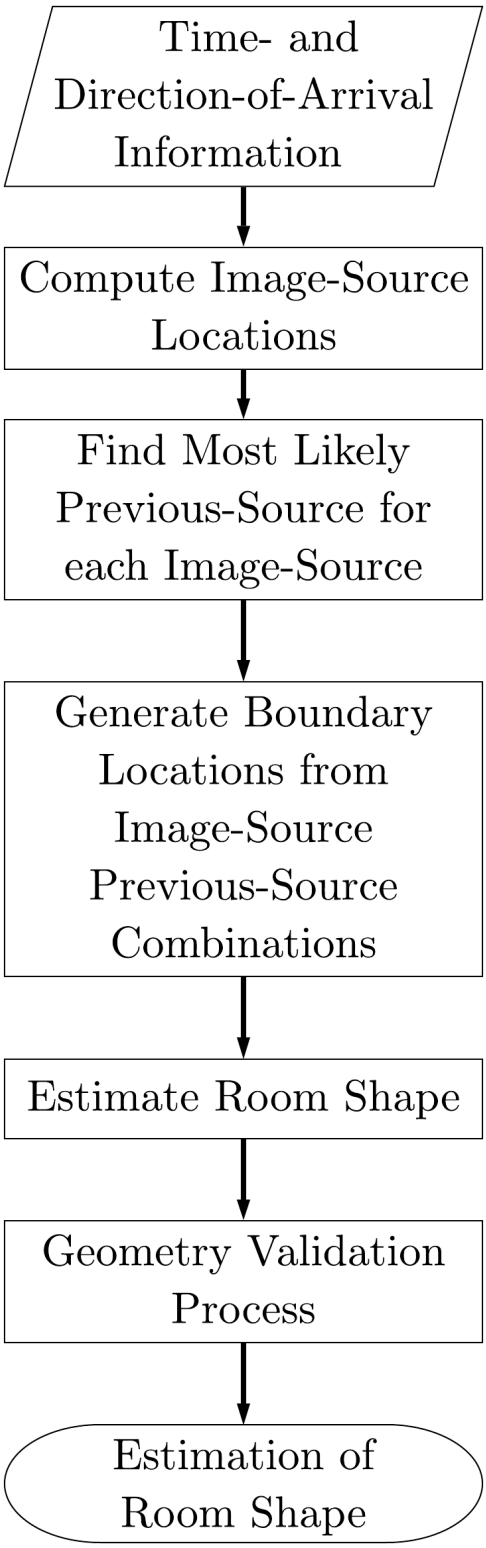

Figure 7.2: Flowchart presenting an overview of the proposed geometry inference process.

boundary position, across SRIR measurements, for these is assumed to be the floor and ceiling location. For subsequent reflections, the assumption of first-order does not hold, as the first arriving second-order reflection will likely arrive before the last first-order [6]. Therefore, for

subsequent detections in the SRIR the previous-source can either be the loudspeaker or an image-source – which is produced by a detection with a ToA less than that of the detection being analysed – and therefore needs to be searched for. The aforementioned assumptions are used to limit the image-source reversion process, by only considering previous-sources that produce boundaries that adhere to the proposed assumptions.

The first consideration in the process is to ascertain whether the image-source is as a result of a reflection from a known boundary. This is tested for the source and all image-sources ($\widetilde{\mathbf{s}}_k$) with a ToA less than that of $\widetilde{\mathbf{s}}_i$ as,

$$\text{previousSource} = \widetilde{\mathbf{s}}_k, \text{ if } ||(\widetilde{\mathbf{s}}_k + 2\,\langle \widetilde{\mathbf{b}}_l - \widetilde{\mathbf{s}}_k, \hat{\mathbf{n}}_l \rangle \hat{\mathbf{n}}_l) - \widetilde{\mathbf{s}}_i|| \leq \epsilon_{\widetilde{\mathbf{s}}} \tag{7.5}$$

where $\widetilde{\mathbf{s}}_k$ is the image-source for the k^{th} reflection, $l = 1 : L$ is the number of inferred boundaries defined by first-order reflections, $\langle ., . \rangle$ denotes dot product, and $\epsilon_{\widetilde{\mathbf{s}}}$ is an empirically defined threshold value chosen to allow for inaccuracies in ToA and DoA estimation. If any of these image-sources tested produce an image-source location close to the actual image-source ($\widetilde{\mathbf{s}}_i$) it is assumed to be the most-likely previous-source.

If no existing boundaries defined by a first-order reflection are attributable to $\widetilde{\mathbf{s}}_i$, then a new boundary is defined. As with the previous work in the literature [6, 8, 110] an image-source that cannot be defined using existing boundaries is assumed to be first-order. However, contrary to these works a set of constraints are imposed to remove image-sources that are as a result of false-positive detections, these constraints are,

- The difference in propagation distance Δl between the image-source-to-receiver path and source-to-boundary to receiver path should be within a defined threshold such that $\Delta l \leq \epsilon_l$, where ϵ_l is the threshold

- The inferred boundary is perpendicular to the floor, defined using the z-axis coefficient for the boundary's normal $\Delta\widetilde{\mathbf{n}}$, which should be 0 for a boundary perpendicular to the floor and ceiling, constrained as $\Delta\widetilde{\mathbf{n}} \leq \epsilon_{\widetilde{\mathbf{n}}}$, where $\epsilon_{\widetilde{\mathbf{n}}}$ is the threshold value.

- The inferred boundary is at least 50 cm away from the source and receiver, as defined by imposed minimum source-to-boundary and receiver-to-boundary distances.

- The specular reflection produced by the path from source-to-boundary should have x and y directional cosines close to that of the actual reflection path from image-source-to-receiver, such that $\Delta\angle \leq \epsilon_{\angle}$ where $\epsilon_{\angle}$ is the threshold value used and $\Delta\angle$ is calculated as $||[\widetilde{\alpha}, \widetilde{\beta}] - [\alpha, \beta]||$ and $\widetilde{\alpha}$, $\widetilde{\beta}$ are calculated, from [153], as,

$$\widetilde{\alpha} = \widetilde{\alpha}_{prev} - 2\cos(\upsilon)\mu \tag{7.6}$$

$$\widetilde{\beta} = \widetilde{\beta}_{prev} - 2\cos(\upsilon)\eta \tag{7.7}$$

where $\widetilde{\alpha}$ and $\widetilde{\beta}$ are the directional cosines along the x and y axes respectively, α_{prev} and β_{prev} are the directional cosines computed for a line going from the previous-source to the point where the line from image-source-to-receiver intersects the boundary, υ is the angle of incidence, and μ and η are the directional cosines of the normal vector of the plane along the x and y axes, respectively. The implication of defining a reflection that is not attributable to an existing inferred boundary as being first-order is that any higher-order reflections defined as a first-order reflection will produce a boundary distant from the desired boundary location, and therefore, a geometry validation process is required to refine the inferred boundaries. Furthermore, second-order reflections that are produced by interactions between perpendicular boundaries will produce an angled boundary that will impact the inferred shape of the room. Therefore, attempts are made to find the correct previous-source for image-sources produced by these perpendicular reflections.

To attempt to find the correct previous-source for an image-source defined by a reflection produced by interactions between perpendicular boundaries, the properties of the image-source method are once again exploited. Given that an image-source is generated by mirroring its previous-source perpendicularly across a boundary, for the case of a reflection between perpendicular boundaries, the relationship between the image-source, previous-source, and the previous-source of the previous-source, can be expressed as a rotation of these image-sources around a point in space. An example of this can be seen in Figure 7.3 for both a second- and third-order reflection. From this relationship, this point of rotation must be equidistant from both the image-source, the previous-source, and the previous-source for the previous source. Using the point on boundary equation in (7.2), the point of rotation $\mathbf{p_r}$ can be expressed as,

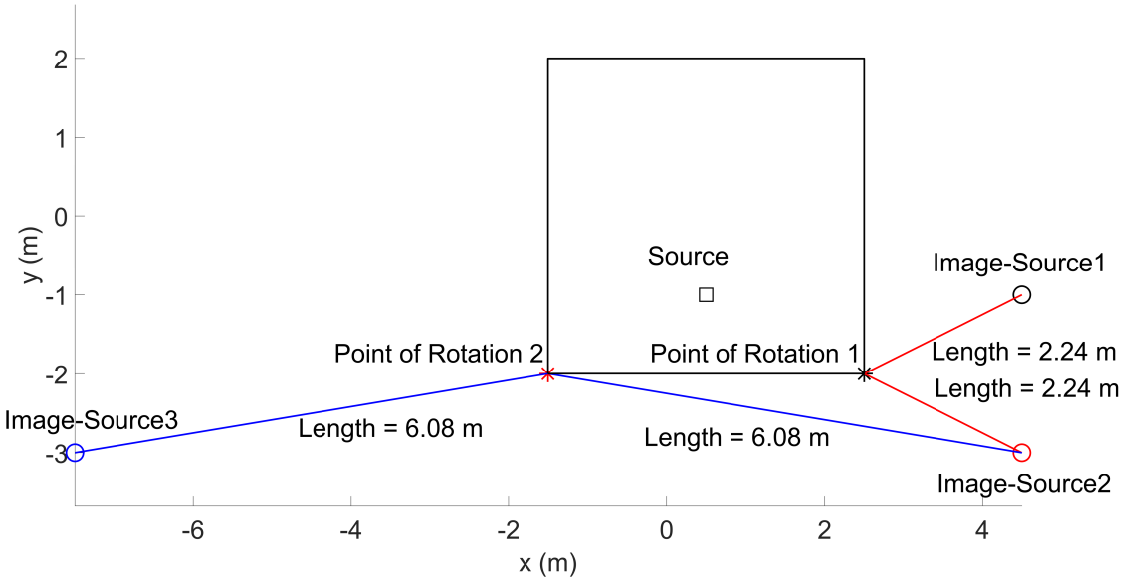

Figure 7.3: Diagram showing the rotational relationship between the image-source and its previous source, in this case Image-Source2 with Image-Source1 and Image-Source3 with Image-Source2. Image-Source1 is produced by mirroring the source in the boundary on the right side of this simple, square, 2D geometry, Image-Source2 is produced by mirroring Image-Source1 in the lower boundary, and Image-Source3 is produced by mirroring Image-Source2 in the left boundary. Point of Rotation 1 is the mid-point between Image-Source2 and the Source location, and Point of Rotation 2 is the mid-point between Image-Source3 and Image-Source1.

$$\mathbf{p_r} = \frac{\widetilde{\mathbf{s}_i} + \widetilde{\mathbf{s}_j}}{2} \tag{7.8}$$

where $\widetilde{\mathbf{s}_i}$ is the image-source being analysed and $\widetilde{\mathbf{s}_j}$ is the previous-source of the previous-source. The image-source produced for a reflection between perpendicular boundaries can therefore be detected if the image-source and previous-source are equidistant from this point of rotation as,

$$\text{previousSoure} = \widetilde{\mathbf{s}}_k, \text{ if } |\ ||\widetilde{\mathbf{s}_i} - \mathbf{p_r}|| - ||\widetilde{\mathbf{s}_k} - \mathbf{p_r}||\ | \leq \epsilon_o \tag{7.9}$$

If more than one previous-source can be defined using this relationship then the previous-source with the smallest error in reflection path is used as,

$$\min(\Delta l + \Delta\angle + \Delta\widetilde{\mathbf{n}}) \tag{7.10}$$

In the case that none of these steps produce a valid candidate previous-source, the image-source

in question is assumed to be as a result of a false-positive detection made by the reflection detection method. An overview of this process can be seen in Algorithm 6–7.

For each detected image-source and previous-source combination the boundary, $\widetilde{\mathbf{B}}$, between them can be defined as the four corners of a quralateral patch using the point-on-boundary and boundary normal (7.2), from [154], as,

$$\widetilde{B}_x = \widetilde{\mathbf{b}}_x + \bar{\mathbf{W}}_{1,1}\begin{bmatrix}1 & -1 & -1 & 1\end{bmatrix}^T + \bar{\mathbf{W}}_{1,2} * \begin{bmatrix}-1 & -1 & 1 & 1\end{bmatrix}^T \tag{7.11}$$

$$\widetilde{B}_y = \widetilde{\mathbf{b}}_y + \bar{\mathbf{W}}_{2,1}\begin{bmatrix}1 & -1 & -1 & 1\end{bmatrix}^T + \bar{\mathbf{W}}_{2,2} * \begin{bmatrix}-1 & -1 & 1 & 1\end{bmatrix}^T \tag{7.12}$$

$$\widetilde{B}_z = \widetilde{\mathbf{b}}_z + \bar{\mathbf{W}}_{3,1}\begin{bmatrix}1 & -1 & -1 & 1\end{bmatrix}^T + \bar{\mathbf{W}}_{3,2} * \begin{bmatrix}-1 & -1 & 1 & 1\end{bmatrix}^T \tag{7.13}$$

where $\widetilde{\mathbf{b}}_x$, $\widetilde{\mathbf{b}}_y$, and $\widetilde{\mathbf{b}}_z$ are the x, y, and z coordinates for the point on the boundary, $\bar{\mathbf{W}}$ is the $[3 \times 2]$ matrix containing two points that are orthogonal to the boundary normal computed from the orthonormal null space of the plane normal, and the two vectors $\begin{bmatrix}-1 & -1 & 1 & 1\end{bmatrix}$ and $\begin{bmatrix}1 & -1 & -1 & 1\end{bmatrix}$ are used to define a plane that is 2 m in length. The initial length of the boundary is arbitrary, as it has no bearing on the final inferred geometry.

While these proposed steps aim to reduce the impact of incorrectly inferred boundaries, it is not infallible as in some cases the correct previous-source was not observed. Furthermore, errors in the estimated ToA and DoA for higher-order reflections can result in an image-source that cannot be attributable to a corresponding boundary, and this produces additional boundaries outside of the desired room's geometry. Therefore, to accurately infer the shape of a given room, further steps are required to remove any erroneously inferred boundaries.

Algorithm 6: Pseudocode for image-source reversion process considering a single source and receiver (part 1)

```
Generate image-sources
number of detections = 0
for ii = number of reflection : -1 : 1
    if imageSource(ii,:) within 1 meter of source or receiver then
        remove imageSource(ii,:) as it cannot produce a valid boundary
        continue to next loop iteration
    end
    if number of detections < 2 then
        if norm(((imageSource(ii,:) + source)/2) - source) > 0.5 and
          norm(((imageSource(ii,:) + source)/2) - receiver) > 0.5 then
            previousSource(ii,:) = source
            number of detections++;
            continue to next loop iteration
        end
    end
    for ll = 1 : number of first-order boundaries do
        for kk = ii + 1 : number of reflections do
            if norm((imageSource(kk,:) + 2 * dot(pointOnBoundary(ll,:) -
              imageSource(kk,:), boundaryNormal(ll,:))-imageSource(ii,:)) < ε_s then
                previousSource(ii,:) = imageSource(kk,:)
                number of detections++
                continue
            end
        end
    end
```

7.3.2 Geometry Validation

From Figure 7.4, it can be seen that there are three types of potentially erroneous boundary detections:

- Boundaries positioned on the corners of the desired geometry as a result of not detecting the correct previous-source for a second-order reflection between perpendicular boundaries or additional inferred angled boundaries, as seen in examples b, c and e.

- Boundaries positioned immediately after another boundary, which are likely to be a product of either noise, or a single reflection being detected as multiple separate arrivals, as seen in examples a and f.

- Boundaries positioned far outside of the desired geometry, which are as a result of higher-

Algorithm 7: Pseudocode for image-source reversion process considering a single source and receiver (part 2)

```
// Continuation of for loop from Algorithm 6
1
2     Define new boundary using source as previous-source
3     if Δl < ϵ_l and Δ∠ < ϵ_∠ and Δñ < ϵ_ñ then
4         possiblePreviousSource = source;
5         if bounday is the first that can define the floor or ceiling then
6             previousSource(ii,:) = source
7             number of detections = number of detections + 1
8             continue to next loop iteration
9         end
10    end
11    store = 1
12    for kk = ii + 1 : number of reflections do
13        if imageSource(kk,:) and (imageSource(ii,:) have a difference in distance <
           ϵ_o to the point of rotation then
14            possiblePreviousSource(store, :) = imageSource(kk,:)
15            store = store + 1
16        end
17    end
18    if length(possiblePreviousSource) == 0 then
19        remove imageSource(ii,:)
20        continue
21    end
22    if length(possiblePreviousSource) > 1 then
23        [ , minIndex] = min(Δl + Δ∠ + Δñ)
24        previousSource(ii,:) = possiblePreviousSource(minIndex, :)
25        number of detections = number of detections + 1
26    else
27        previousSource(ii,:) = possiblePreviousSource
28        number of detections = number of detections + 1
29    end
30 end
```

order reflections being defined as first-order, as seen across all six examples.

The latter two of these potentially erroneous boundary conditions will be considered here, as they will have the largest impact on the accuracy of the geometry inference process.

Ahead of the next step, boundaries that are coincident are removed until only one remains, reducing the number of boundaries to be tested and improve computational efficiency of this approach. Two boundaries are defined as being coincident if the boundary normals $\widetilde{\mathbf{n}}_1$ and $\widetilde{\mathbf{n}}_2$

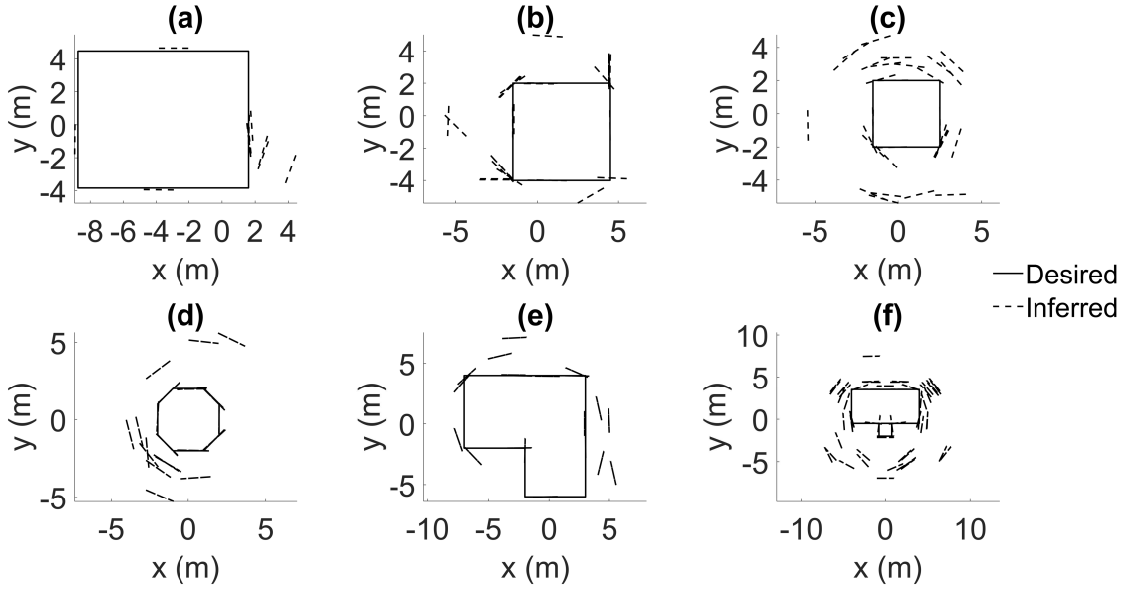

Figure 7.4: Example inferred boundaries (dashed lines) and the desired geometry (solid lines) for six different test cases, (a) Real-world measurements of a cuboid-shaped room, (b) CATT-Acoustic simulated measurements for a cuboid-shaped room, (c) CATT-Acoustic simulated measurements for a second cuboid-shaped room, (d) CATT-Acoustic simulated measurements for an octagonal-shaped room, (e) CATT-Acoustic simulated measurements for a L-shaped room, and (f) CATT-Acoustic simulated measurements for a T-shaped room. Each figure shows outlier boundaries outside of the desired geometry produced by incorrect assignment of previous-source.

are parallel and the inferred point on the boundaries $\widetilde{\mathbf{b}}_1$ and $\widetilde{\mathbf{b}}_2$, where $_1$ and $_2$ denotes different boundaries, exists on both boundaries [155], such that,

$$||\widetilde{\mathbf{n}}_1 \times \widetilde{\mathbf{n}}_2|| \leq \epsilon_{par} \tag{7.14}$$

$$\text{and } | < \widetilde{\mathbf{n}}_1, \widetilde{\mathbf{b}}_1 - \widetilde{\mathbf{b}}_2 > | \leq \epsilon_{point} \tag{7.15}$$

where ϵ_{par} and ϵ_{point} are empirically defined threshold values to account for small variations in boundary position as a result of ToA and DoA errors. An additional constraint is required to account for non-convex-shaped rooms, as multiple distinct boundaries can be co-planar, as seen for boundaries 1 and 2 in Figure 7.5. Therefore, the distance between the points on plane $\widetilde{\mathbf{b}}_1$ and $\widetilde{\mathbf{b}}_2$ must be less that the minimum parallel plane distance of 1 m (as defined by the minimum source/receiver to boundary distance).

To perform the boundary validation process, an approximate estimation of the room's inferred geometry, based on the nearest intersection points between non-parallel inferred boundaries, is

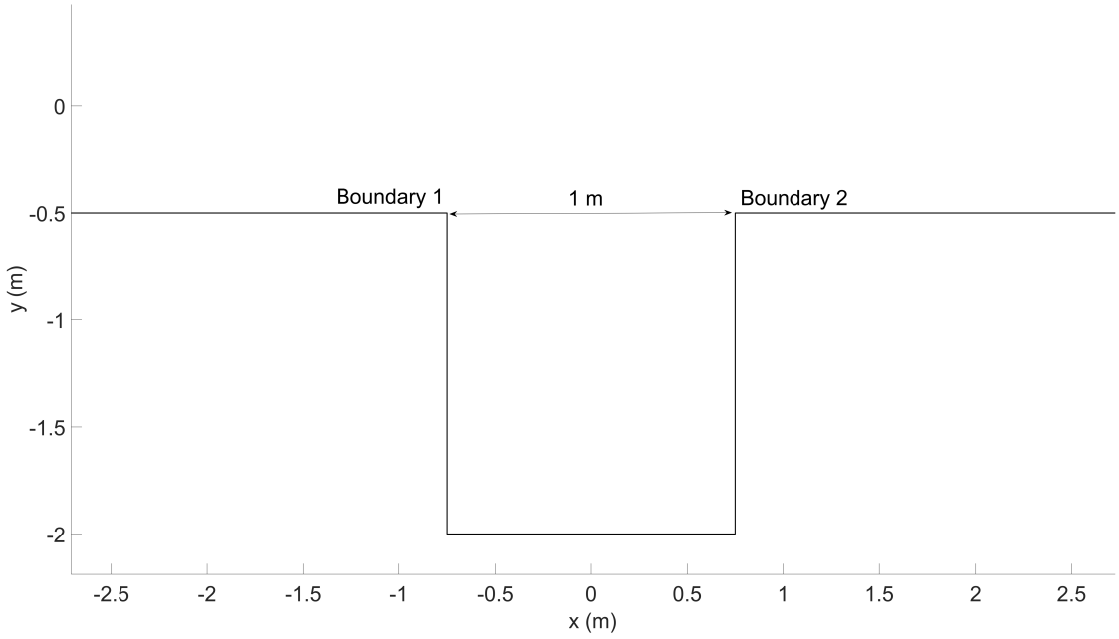

Figure 7.5: Example non-convex T-shaped room where boundaries 1 and 2 are mathematically coincident, but belong to two separate boundaries.

generated. Under the assumption that all walls are perpendicular to the floor and ceiling, the boundary-to-boundary intersection of interest are non-parallel boundaries that intersect along the x and y axes, *i.e.* walls, which from [156] are,

$$x = \frac{d_k \widehat{\mathbf{n}}_{i,y} - d_i \widehat{\mathbf{n}}_{k,y}}{\mathbf{u}_z} \tag{7.16}$$

$$y = \frac{d_i \widehat{\mathbf{n}}_{k,x} - d_k \widehat{\mathbf{n}}_{i,x}}{\mathbf{u}_z} \tag{7.17}$$

$$z = 0 \tag{7.18}$$

where subscript x, y, and z denote the Cartesian coordinates, and the coefficient d and the intersection direction vector $\mathbf{u}$ are, from [156], computed as,

$$d_i = - < \widehat{\mathbf{n}}_i, \mathbf{p}_i > \tag{7.19}$$

$$d_k = - < \widehat{\mathbf{n}}_k, \mathbf{p}_k > \tag{7.20}$$

$$\mathbf{u} = \widehat{\mathbf{n}}_i \times \widehat{\mathbf{n}}_k \tag{7.21}$$

where $<,>$ denotes dot product and $\times$ denotes cross product. If any intersection is further than 100 m from the receiver locations, it is assumed to be the intersection point between two nearly-parallel boundaries and is ignored. The resulting inferred boundary $\mathbf{B}_i$ is computed from the nearest intersecting non-parallel boundaries, on either side of the boundary $\widetilde{\mathbf{B}}_i$, for notation purposes these are referred to as $\widetilde{\mathbf{B}}_k$ and $\widetilde{\mathbf{B}}_j$. The boundary, $\widetilde{\mathbf{B}}_i$, can then be constrained based on these points of intersection as,

$$\mathbf{B}_i = \begin{bmatrix} \begin{bmatrix} x(\widetilde{\mathbf{B}}_i, \widetilde{\mathbf{B}}_j) & y(\widetilde{\mathbf{B}}_i, \widetilde{\mathbf{B}}_j) & z(\widetilde{\mathbf{B}}_i, \widetilde{\mathbf{B}}_j) \end{bmatrix} + \begin{bmatrix} u_{j,x} & u_{j,y} & \frac{\min(z)}{u_{j,z}} \end{bmatrix} \\ \begin{bmatrix} x(\widetilde{\mathbf{B}}_i, \widetilde{\mathbf{B}}_k) & y(\widetilde{\mathbf{B}}_i, \widetilde{\mathbf{B}}_k) & z(\widetilde{\mathbf{B}}_i, \widetilde{\mathbf{B}}_k) \end{bmatrix} + \begin{bmatrix} u_{k,x} & u_{k,y} & \frac{\min(z)}{u_{k,z}} \end{bmatrix} \\ \begin{bmatrix} x(\widetilde{\mathbf{B}}_i, \widetilde{\mathbf{B}}_k) & y(\widetilde{\mathbf{B}}_i, \widetilde{\mathbf{B}}_k) & z(\widetilde{\mathbf{B}}_i, \widetilde{\mathbf{B}}_k) \end{bmatrix} + \begin{bmatrix} u_{k,x} & u_{k,y} & \frac{\min(z)}{u_{k,z}} \end{bmatrix} \\ \begin{bmatrix} x(\widetilde{\mathbf{B}}_i, \widetilde{\mathbf{B}}_j) & y(\widetilde{\mathbf{B}}_i, \widetilde{\mathbf{B}}_j) & z(\widetilde{\mathbf{B}}_i, \widetilde{\mathbf{B}}_j) \end{bmatrix} + \begin{bmatrix} u_{j,x} & u_{j,y} & \frac{\min(z)}{u_{j,z}} \end{bmatrix} \end{bmatrix} \tag{7.22}$$

where $\begin{bmatrix} x(\widetilde{\mathbf{B}}_i, \widetilde{\mathbf{B}}_j) & y(\widetilde{\mathbf{B}}_i, \widetilde{\mathbf{B}}_j) & z(\widetilde{\mathbf{B}}_i, \widetilde{\mathbf{B}}_j) \end{bmatrix}$ are the intersection points between boundaries $\widetilde{\mathbf{B}}_i$ and $\widetilde{\mathbf{B}}_j$, $\begin{bmatrix} x(\widetilde{\mathbf{B}}_i, \widetilde{\mathbf{B}}_k) & y(\widetilde{\mathbf{B}}_i, \widetilde{\mathbf{B}}_k) & z(\widetilde{\mathbf{B}}_i, \widetilde{\mathbf{B}}_k) \end{bmatrix}$ are the intersection points between boundaries $\widetilde{\mathbf{B}}_i$ and $\widetilde{\mathbf{B}}_k$, and $\min(z)$ and $\max(z)$ are the z coordinate for the floor and ceiling respectively, computed during the image-source reversion process, which scales the intersection direction vector u to produce the correct z coordinates. An example of the constrained boundaries from Figure 7.4 can be seen in Figure 7.6, where it can be seen that for examples (a)-(e) the boundaries that define the room have been constrained to the right shape, but there exist inferred boundaries that are not part of the desired geometry as a result of incorrectly assigned previous-source candidate for higher-order reflections. To then remove the aforementioned inferred boundaries that are positioned outside of the desired geometry of the room, a three step geometry validation process is proposed. These three steps are as follows, **reflection path validation**, **line-of-sight boundary validation**, and **closed geometry validation**.

Step 1: Reflection Path Validation

The first step is to check if the reflection path from the image-source-to-receiver is obstructed by additional boundaries that are closer to the receiver than the boundary inferred by this image-source. This process will remove the majority of the additional boundaries seen in Figure 7.6, and is performed by defining a line from the image-source that produced the boundary being tested to the receiver, and computing the intersection $\begin{bmatrix} x & y & z \end{bmatrix}$ between the line and every other boundary $\mathbf{B}$ from [157] as,

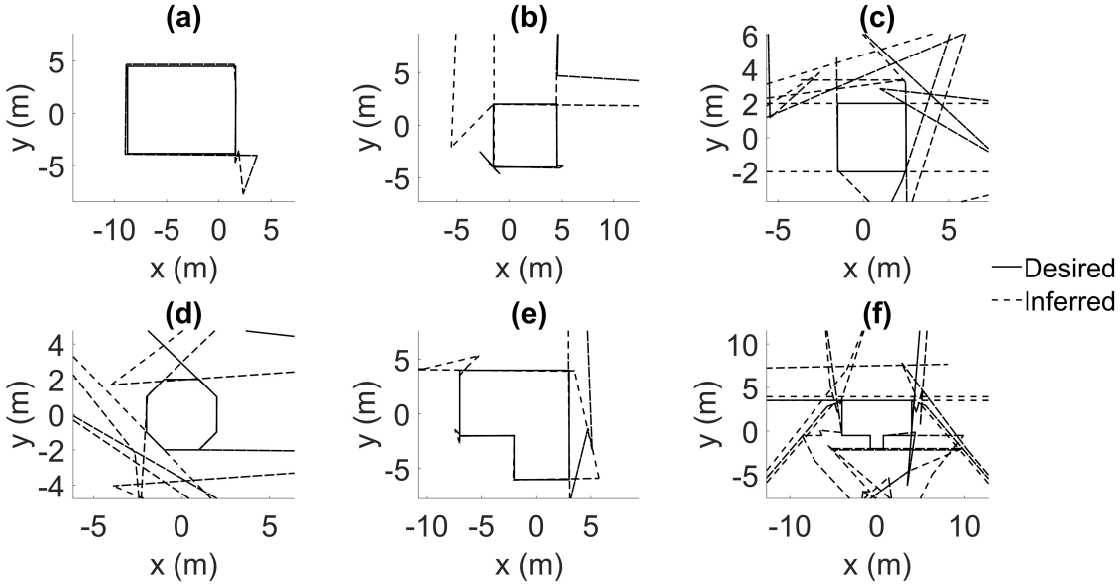

Figure 7.6: Example inferred room shape (dashed lines) and the desired geometry (solid lines) for six different test cases as considered previously in Figure 7.4. The approximate shape of the room exists in all cases, but as a result of outlier boundaries there are incorrect boundaries.

$$\Xi = \frac{\begin{vmatrix} 1 & 1 & 1 & 1 \\ \mathbf{B}_{1,x} & \mathbf{B}_{2,x} & \mathbf{B}_{3,x} & \mathbf{m}_x \\ \mathbf{B}_{1,y} & \mathbf{B}_{2,y} & \mathbf{B}_{3,y} & \mathbf{m}_y \\ \mathbf{B}_{1,z} & \mathbf{B}_{2,z} & \mathbf{B}_{3,z} & \mathbf{m}_z \end{vmatrix}}{\begin{vmatrix} 1 & 1 & 1 & 0 \\ \mathbf{B}_{1,x} & \mathbf{B}_{2,x} & \mathbf{B}_{3,x} & \widetilde{\mathbf{s}}_{i,x} - \mathbf{m}_x \\ \mathbf{B}_{1,y} & \mathbf{B}_{2,y} & \mathbf{B}_{3,y} & \widetilde{\mathbf{B}}_{i,y} - \mathbf{m}_y \\ \mathbf{B}_{1,z} & \mathbf{B}_{2,z} & \mathbf{B}_{3,z} & \widetilde{\mathbf{s}}_{i,z} - \mathbf{m}_z \end{vmatrix}} \tag{7.23}$$

$$x = x_4 + (x_5 - x_4)\Xi \tag{7.24}$$

$$y = y_4 + (y_5 - y_4)\Xi \tag{7.25}$$

$$z = z_4 + (z_5 - z_4)\Xi \tag{7.26}$$

where $\widetilde{\mathbf{s}}_{i,x}$, $\widetilde{\mathbf{s}}_{i,y}$, and $\widetilde{\mathbf{s}}_{i,z}$ are the Cartesian coordinates for the image-source that produces the i^{th} wall, $\mathbf{m}_x$, $\mathbf{m}_y$, and $\mathbf{m}_z$ are the Cartesian coordinates for the receiver position, $\mathbf{B}_{1,x}$ refers to the x axis coordinate of the first corner of the boundary, and $\left|.\right|$ is the determinant of the matrix. As this equation assumes a boundary of infinite length the resulting point of intersection is checked to ensure that it lies on the defined boundary. If any other boundary has an intersection closer

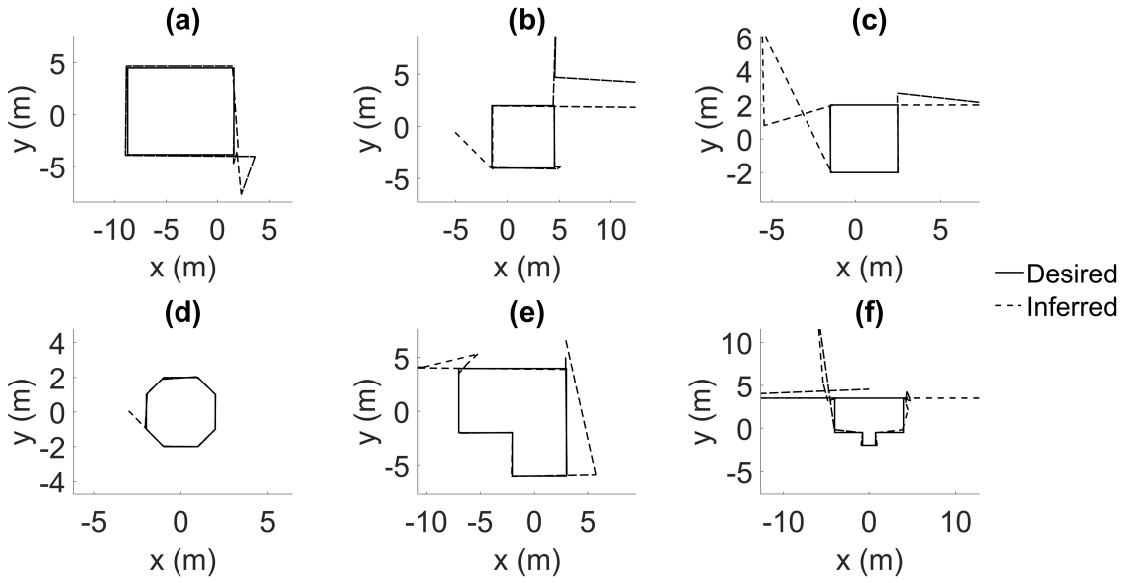

Figure 7.7: Example inferred room shape (dashed lines) and the desired geometry (solid) lines) for the six different test cases presented previously in Figure 7.4 after the reflection path validation process.

to the receiver than the i^{th} boundary being tested, the i^{th} boundary is removed. It is important to note that as a result of the inferred shape, the line between image-source and receiver may not intersect with the boundary it produces. In this case the i^{th} boundary cannot be invalidated and is kept. Once all boundaries have been tested, the shape of the room is inferred from the remaining boundaries and the process is repeated until no further boundaries are removed. An example of the resulting inferred geometry after this step can be seen in Figure 7.7, which shows that the majority of additional boundaries presented in Figure 7.6 have been removed.

Step 2: Line-of-Sight Boundary Validation

While the majority of incorrect boundaries have now been removed, there are still non-valid boundaries that remain as a result of the line between image-source and receiver not intersecting with the boundary. To remove these remaining unwanted boundaries, a line-of-sight test is performed to see if each inferred boundary is visible to at least one receiver position. Any boundaries that are not in line-of-sight of the receiver could not have produced a reflection that arrives at the receiver. To test line-of-sight a set of rays are defined with $0 \leq \theta \leq 359$ and $\phi = 90$ using (7.4) with $d_i = 1$. The value of d_i (the length of ray) is arbitrary as the line-plane intersection equations assume a line of infinite length. The line-plane intersections are then computed using (7.23), substituting $\widetilde{\mathbf{s}}_i$ with the point on the ray defined using (7.4). The first boundary that intersects with each of these rays is considered valid. An example of the resulting inferred boundaries after this step has been performed can be seen in Figure 7.8. As can be seen

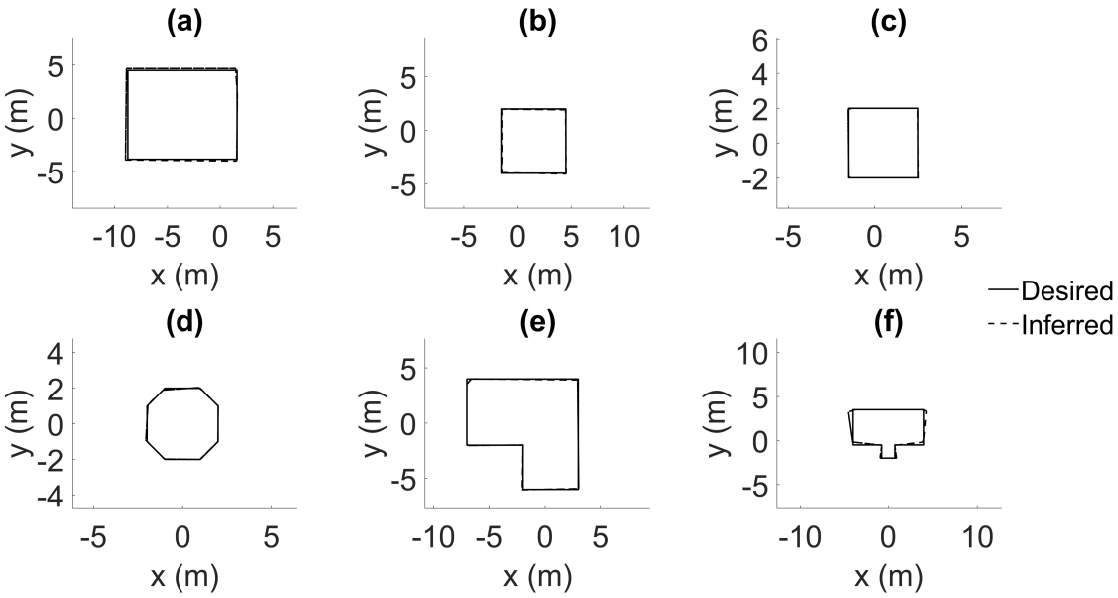

Figure 7.8: Example inferred room shape (dashed lines) and the desired geometry (solid lines) the six different test cases presented previously in Figure 7.4 after removing any boundaries that are not in line-of-sight of the receiver. The results show that all of the remaining external boundaries have now been removed.

all remaining additional boundaries from Figure 7.7 have been removed.

Step 3: Closed Geometry Test

These first two steps will have refined candidate boundaries to that of the desired room for the majority of cases. The final step is to ensure that the inferred geometry of the room produces a closed shape. As with the previous two stages the geometry of the room is first inferred, then any constrained boundaries that do not intersect with two adjacent boundaries, one on each side, are removed.

An overview of this whole geometry validation process can be seen in Algorithm 8–9. The remaining boundaries are then considered valid, and the shape of the room can be inferred from them.

Algorithm 8: Pseudocode for the three step geometry validation process. Step 1 checks to see if the reflection paths is obstructed by additional boundaries, Step 2 checks that each boundary is in line-of-sight of at least one receiver position, and Step 3 checks that the inferred room's shape produces a closed geometry. (Part 1)

```
while changesMade ~= 0 do
    Infer geometry using plane-plane intersections.
    changeHappened = 0
    Step 1: Check reflection path for multiple boundary intersections
    for ii = 1 : numberOfPlanes do
        for kk = 1 : numberOfPlanes do
            if boundary kk intersects line between point of incidence on boundary ii
             and the receiver then
                remove boundary ii
                changeHappened = 1;
            end
        end
    end
    if changeHappened == 0 then
        changesMade = 0
    end
end
Step 2: line-of-sight test
Infer geometry - image-source-to-receiver path must intersect boundary
for ii = 1 : noReceivers do
    for θ = 1 : 359 do
        Define ray in azimuth direction θ from receiver ii
        for kk = 1 : numberOfPlanes do
            if ray intersects boundary kk && intersection is not on the boundary edge
             then
                boundaryIsValid(ii) = 1
            end
        end
    end
end
```

7.4 Testing

Three sets of tests are used to test the proposed method under different measurement conditions. The first test case will test the proposed method with seven CATT-Acoustic [16] simulated enclosed spaces of different sizes, shapes, and complexity, detailing the accuracy of the model under highly controlled measurement conditions. The second scenario consists of 33 source/receiver combinations across two different L-shaped rooms, testing the performance of the method across different measurement set-ups. The final scenario consists of two sets of real-world mea-

Algorithm 9: Pseudocode for the three step geometry validation process. Step 1 checks to see if the reflection paths is obstructed by additional boundaries, Step 2 checks that each boundary is in line-of-sight of at least one receiver position, and Step 3 checks that the inferred room's shape produces a closed geometry. (Part 2)

```
Remove boundaries where boundaryIsValid == 0
Infer geometry using plane-plane intersections.
Step 3: Closed Geometry test
for ii = 1 : numberOfPlanes do
    Compute the distance between boundary ii and adjacent boundaries
    if boundaries do not connect and distance between boundaries is < 0.1 then
        remove boundary ii
    end
end
Infer geometry.
```

surements for a cuboid-shaped room, testing the robustness of the method to real-world implementation.

Preliminary Testing: Ground-Truth

This scenario assesses the accuracy of the geometry inference method when presented with ground-truth data, that is an exact measurement of the time- and direction-of-arrival for each reflection and an exact simulation of the SRIR where each reflection is represented as a single peak within the SRIR and no additional processing, such as filtering, is applied). The ground-truth data is generated for a cuboid-shaped room with dimensions [4 m $\times$ 4 m $\times$ 3.5 m], boundary absorption coefficient of 0.02, speed-of-sound defined as 344 m/s, and using simulated reflection information for two different source positions. The data is simulated using an adapted version of the image-source code [158], which outputs a third-order spherical harmonic domain SRIR and ground-truth time- and direction-of-arrival values for each reflection. Five different test scenarios are presented considering: I) Ground-truth ToA and DoA for all reflections; II) Ground-truth ToA and DoA with randomly generated and normally distributed errors added to the ToA values for 21 different magnitudes 22.67 μs – 476.19μs (from one sample at 44.1 kHz up to maximum error reported for the EDESAR method); III) Ground-truth ToA and DoA with randomly generated and normally distributed errors added to the azimuth DoA for 10 different magnitudes 0°–10°; IV) Ground-truth ToA and DoA with randomly generated and normally distributed errors added to the elevation DoA for 10 different magnitudes 0°–10°; and V) Ground-truth SRIR sampled at 44.1 kHz with additive noise at signal-to-noise ratios (SNRs) of no noise and from

60 dB to 0 dB in steps of 5 dB. The additive noise used in the SNR tests is generated by adding randomly generated Gaussian white noise to each channel of the SRIR. These tests will assess the accuracy of the method when presented with exact data, assess the robustness of the method to time- and direction-of-arrival estimation errors, and assess the method's robustness to inteferring noise. The geometry of the room and source and receiver locations can be seen in Figure 7.9.

Test Case One

The first example in this scenario consist of five different simulated environments, two cuboid-shaped rooms (126 m^3 and 56 m^3), one octagonal-shaped room (42 m^3), one L-shaped room (240 m^3), and one T-shaped room (137 m^3). All but the T-shaped room consist of two measurement positions, which used three. All of these rooms are simulated using CATT-Acoustic v.9.1a [152]. To ensure that the resulting SRIRs consist of more than a sparse set of reflections 10,000,000 rays are used producing sufficient coverage throughout the environment, and for testing purposes diffuse reflections are turned off. Furthermore, for all scenarios the boundaries are defined as being made of wood, using CATT-Acoustic's WOOD30 material [16]. Across all tests the source is defined as being 1.5 m off the floor.The resulting SRIRs are rendered out as third-order spherical harmonic domain signals. The dimensions, source and receiver positions, room shapes, and impulse-responses can be seen in Figure 7.9-7.13, and the SRIR and the detected reflection locations can be seen in Appendix A Figures A.1–A.5

An additional two sets of simulated SRIRs for different source and receiver locations for a third cuboid-shaped room ($\sim$504.63 m^3) with small recessed windows is also used. This example tests the performance of the method when there are features of the room present that cannot be inferred using the proposed method. As before 10,000,000 rays are used producing sufficient coverage throughout the environment, with diffuse reflections turned off. The floor is linoleum (LINOLEUM30), ceiling is defined as perforated metal (METAL_PERF), and the walls are brick (BRICK_WALL1). The dimensions, source and receiver positions, room shape, and impulse-responses can be seen in Figures 7.14-7.15, and the SRIR and the detected reflection locations can be seen in Appendix A Figures A.6–A.7.

Test Case Two

This test consists two L-shaped rooms, with volumes 320 m^3 and 360 m^3, simulated in CATT-Acoustic using the same parameters as the L-Shaped room in Scenario One. These rooms are simulated using a single receiver positioned in line-of-sight of every boundary, and 14 and 15

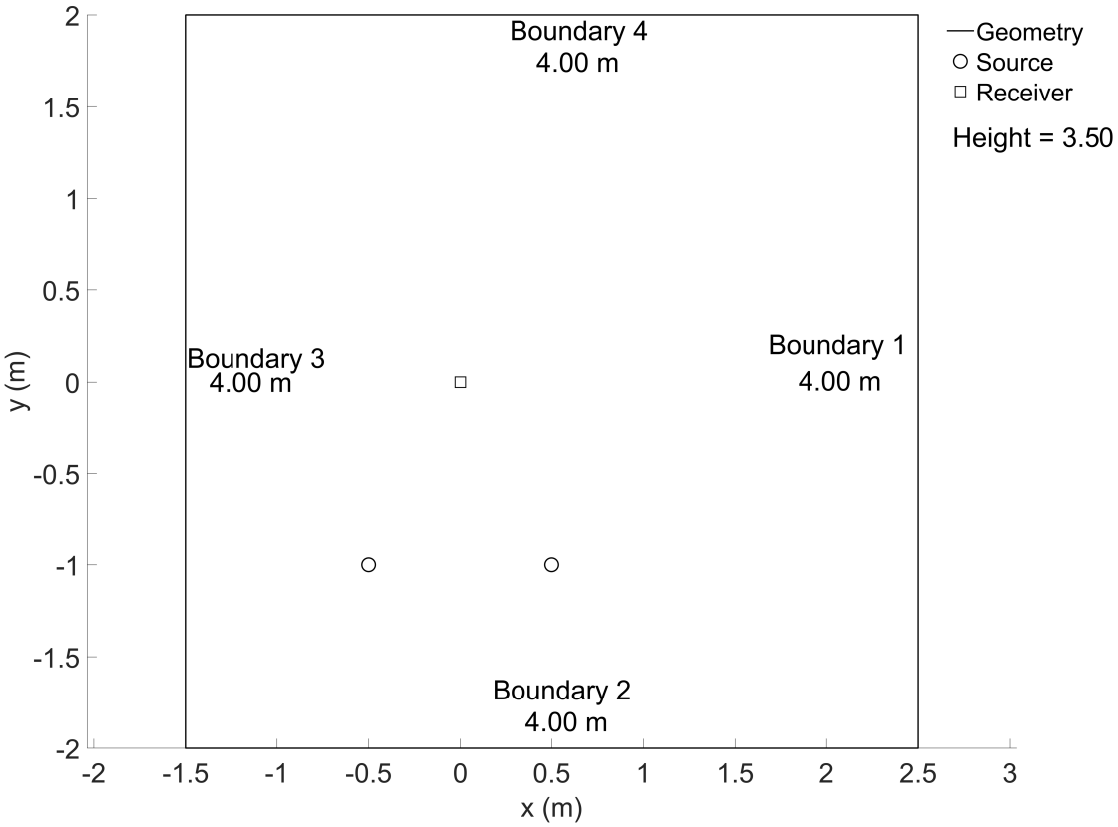

Figure 7.9: Geometry for Ground Truth testing and Scenario One First Cuboid-Shaped Room. Square marker denotes the receiver position and the circle markers denote the source positions.

randomly selected source positions for room one and two across the two segments of the room. From these two sets of 15 source positions for each L-shaped room, a selection of 33 source combinations that ensure a first-order reflection from each boundary, are used to test the proposed method. This example tests the variability of the performance of the method, quantifying any difference in estimation accuracy between the two rooms. The source positions and room shape can be seen in Figures 7.16-7.17 and the combinations of sources used can be seen in Tables 7.1-7.2.

Test Case Three

This test consists of two sets of SRIRs measured in a real world space, with a volume of 360.11 m^3, with each measurement set using different source and receiver locations. The receiver used is the EigenMike EM32 [159], a spherical microphone array with 32- spatially distributed channels across the sphere, and the source used is a Genelec 8030 [139] loudspeaker. The test signal used to capture the response of the room is an exponential sine-sweep [140], 20 s in length with a frequency range of 100 Hz-20 kHz, using the inverse-filter of the original sine-sweep to produce the SRIR. To better approximate an omnidirectional source, the mean

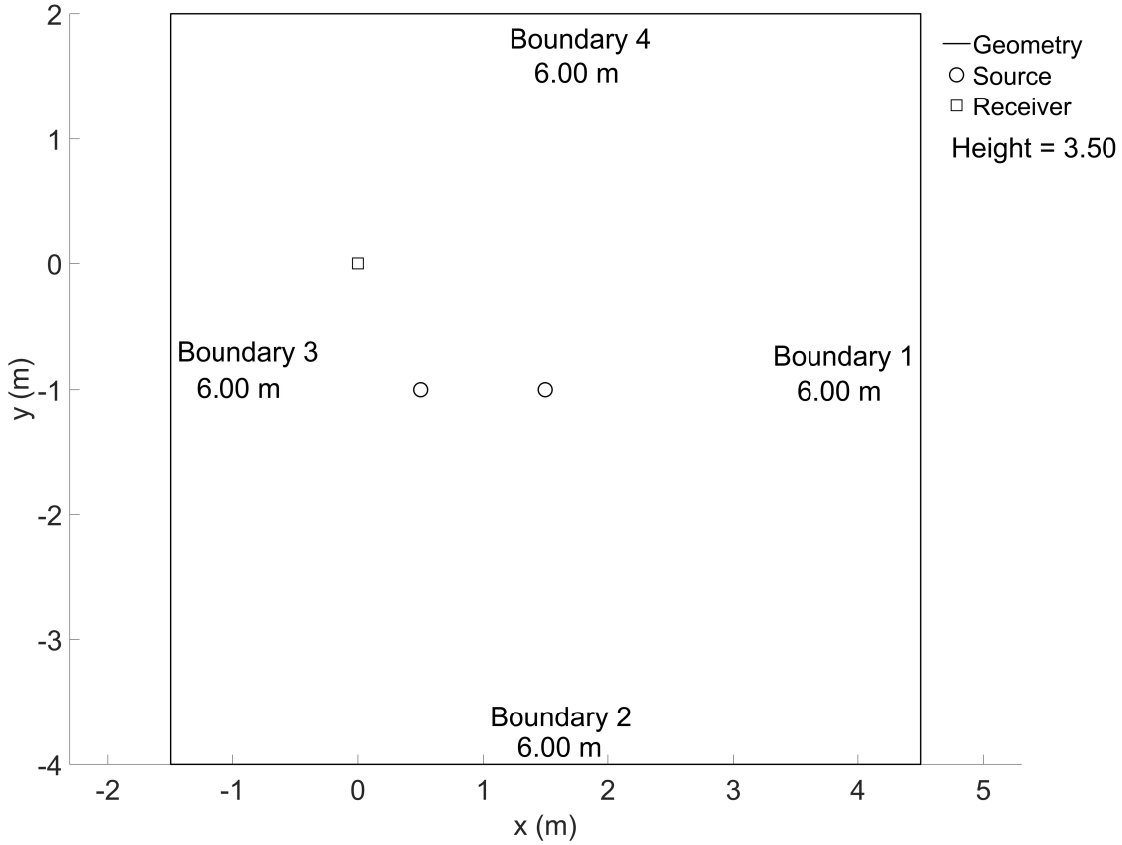

Figure 7.10: Geometry for Scenario One Second Cuboid-Shaped Room. Square marker denotes the receiver position and the circle markers denote the source positions.

of the SRIRs measured at four speaker orientations ($0°$, $90°$, $180°$, and $270°$) is taken, as used in [141]. The final SRIRs are then normalised to have a maximum sample value of ± 1, and converted to third-order spherical harmonic domain signals using MH Acoustics' EigenStudio [13]. The measurement room is cuboid-shaped with dimensions 10.35 m$\times$13.29 m$\times$4.19 m, and has a number of non-removable, adjustable, floor length curtains. As it was not possible to remove these curtains, they were positioned, as much as is possible, to limit their impact on the obtained SRIRs. Hence they were arranged in corners of the room, across windows, and, where possible, to cover features on the walls such as electrical outputs, as well as the computer and interface used for the measurements. While it is accepted that this is non-ideal, and could have some impact on the results, every effort has been made to minimize their potential influence on the measurements obtained, and ensure that the main reflective boundaries are exposed and clear from other possibly confounding features. Furthermore, the ceiling was covered in large metal piping connected to extractor fans and a layer of metal railing approximately 1 m from the ceiling. The noise floor in the room is measured as 60.2 dBA using an SPL meter and the room's temperature was $24.4°$C, and hence the speed of sound is estimated as 346.97 m/s [160]. The room's geometry, loudspeaker and receiver positions, and impulse-responses can be seen in

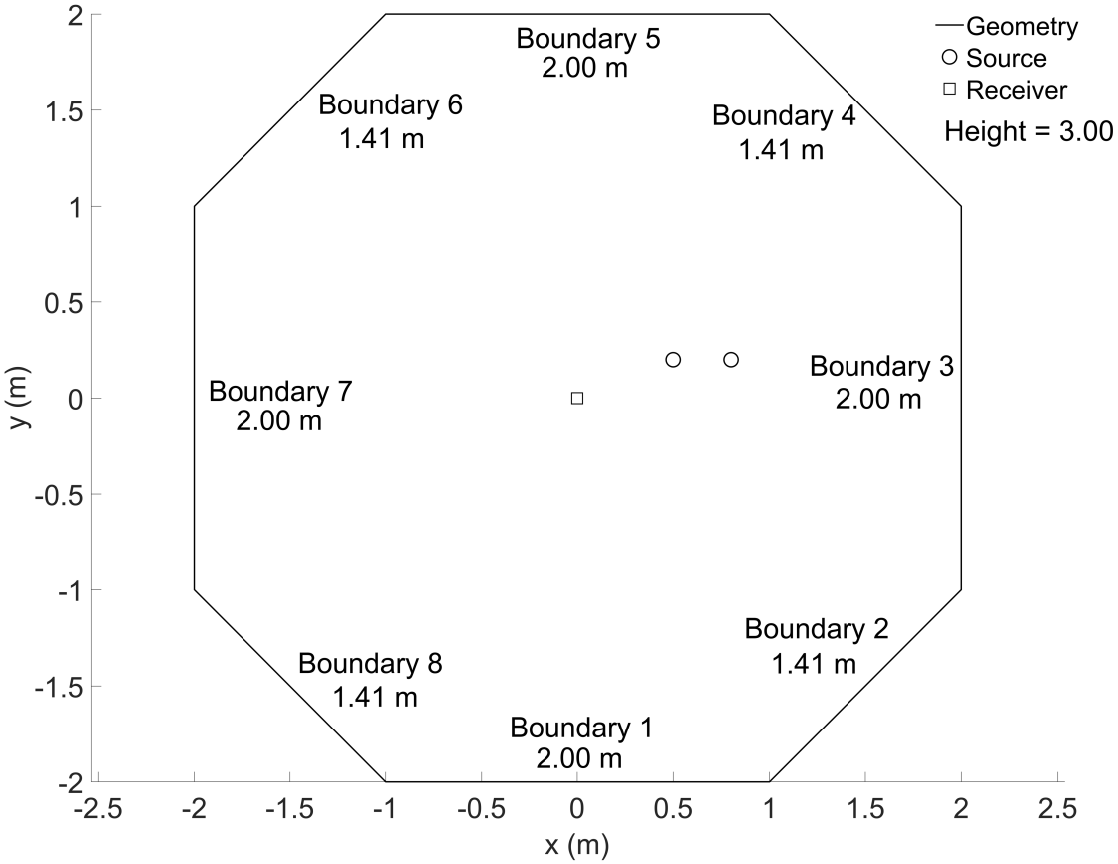

Figure 7.11: Geometry for Scenario One Octagonal-Shaped Room. Square marker denotes the receiver position and the circle markers denote the source positions.

Figure 7.18 - 7.19, a picture of the measurement environment can be seen in Figure 7.20, and the SRIR and the detected reflection locations can be seen in Appendix A Figures A.8–A.9.

The number of measurements used for each test case is equal to that required to ensure a first-order reflection for each boundary is captured. In practice any number of SRIRs can be used, for different source and receiver positions, but fewer is more computationally efficient, as fewer SRIR need to be analysed, resulting in fewer candidate image-sources, and consequently fewer boundaries that need to be validated. The SRIRs being analysed are truncated to 1500 samples for the first two cuboid- and the octagonal-shaped rooms in Scenario One, 2000 samples for the L- and T-shaped rooms in Scenario One and all cases in Scenario Two, and 3000 for the third cuboid-shaped room in Scenario One and all sets in Scenario Three. The truncation lengths are chosen to allow candidate reflection up to fourth-order [161] to be detected, which from [18, 36], is defined as,

$$T_{r_o} = \frac{4V}{cS}(r_o + 1) \tag{7.27}$$

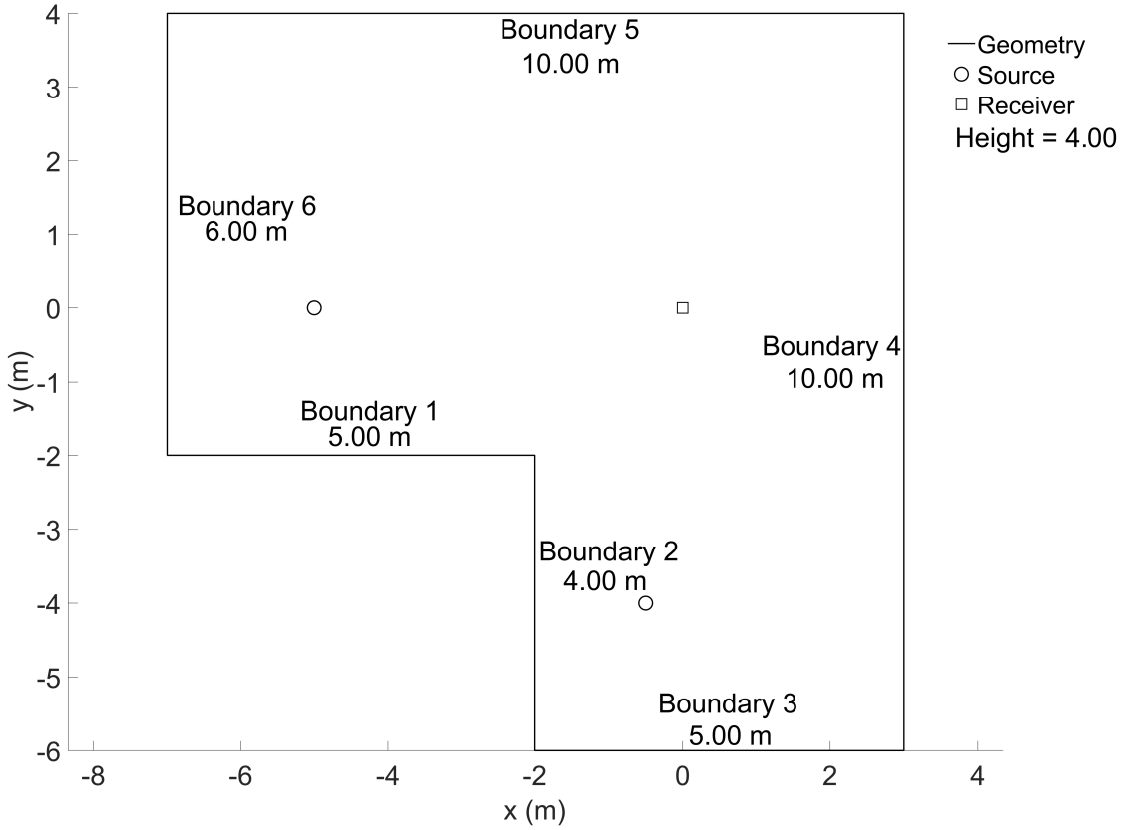

Figure 7.12: Geometry for Scenario One L-Shaped Room. Square marker denotes the receiver position and the circle markers denote the source positions.

where T_{ro} is the estimtaed arrival time of the first arriving fourth-order reflection, V is the volume of the room, S is the surface area of the room, c is the speed of sound, and r_o is the reflection order (in this case four). The resulting values is then rounded up to the nearest multiple of 500. For all scenarios every detection made by the EDESAR reflection detection method is used, and no candidate detections have been manually removed.

The threshold values, $\epsilon_{\widetilde{\mathbf{s}}}$, ϵ_o, ϵ_l, $\epsilon_\angle$, $\epsilon_{\widetilde{\mathbf{n}}}$, ϵ_{par}, and ϵ_{point}, discussed in Section 7.3 have been derived empirically through examination of results obtained for the different Scenarios used for testing, and chosen so all first-order reflections are assigned to the correct boundaries, while reducing the number of inaccurately inferred boundaries due to false-positive detections, the same values for these are used across all test cases. These are shown in Table 7.3.

To present the accuracy of the proposed method, four error metrics are used to analyse resulting inferred boundaries:

- ΔPosition - the RMS of the distance between desired and inferred boundaries [6, 8, 9] measured at 10 cm intervals along the length of the target boundary at z = 0.

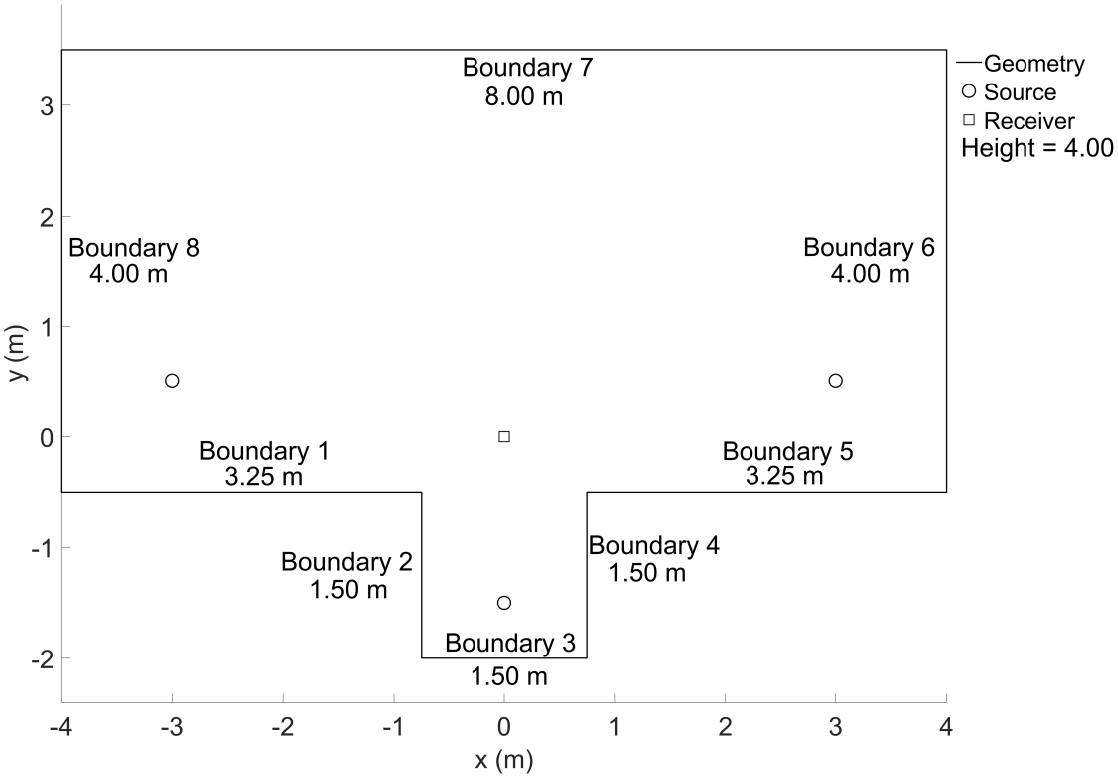

Figure 7.13: Geometry for Scenario One T-Shaped Room. Square marker denotes the receiver position and the circle markers denote the source positions.

- Dihedral Angle [9, 117] - the angle between desired and inferred boundaries, if there is more than one inferred boundary then the average and weighted average (weighted based on the length of each boundary compared to the summed length of the inferred boundaries) is taken over the inferred boundaries .

- ΔLength [6] - the difference in length between desired and inferred boundary.

- δLength - the relative error of the inferred boundary's length to that of the desired boundary, computed as, $\frac{\Delta\text{Length}}{\text{desired length}} * 100$

7.5 Results

As in Chapter 6, the following steps have been performed on the SRIRs prior to inferring boundary locations. Firstly, the SRIRs are temporally adjusted to ensure that the ToA of the direct sound is the same as would be expected given the speed of sound and source-to-receiver distance, removing latency introduced by the measurement or simulation system. The estimated azimuth θ DoA for each reflection is then shifted by the difference between the estimated and expected DoA for the direct sound, ensuring that $\theta = 0^\circ$ is aligned with the positive going x-

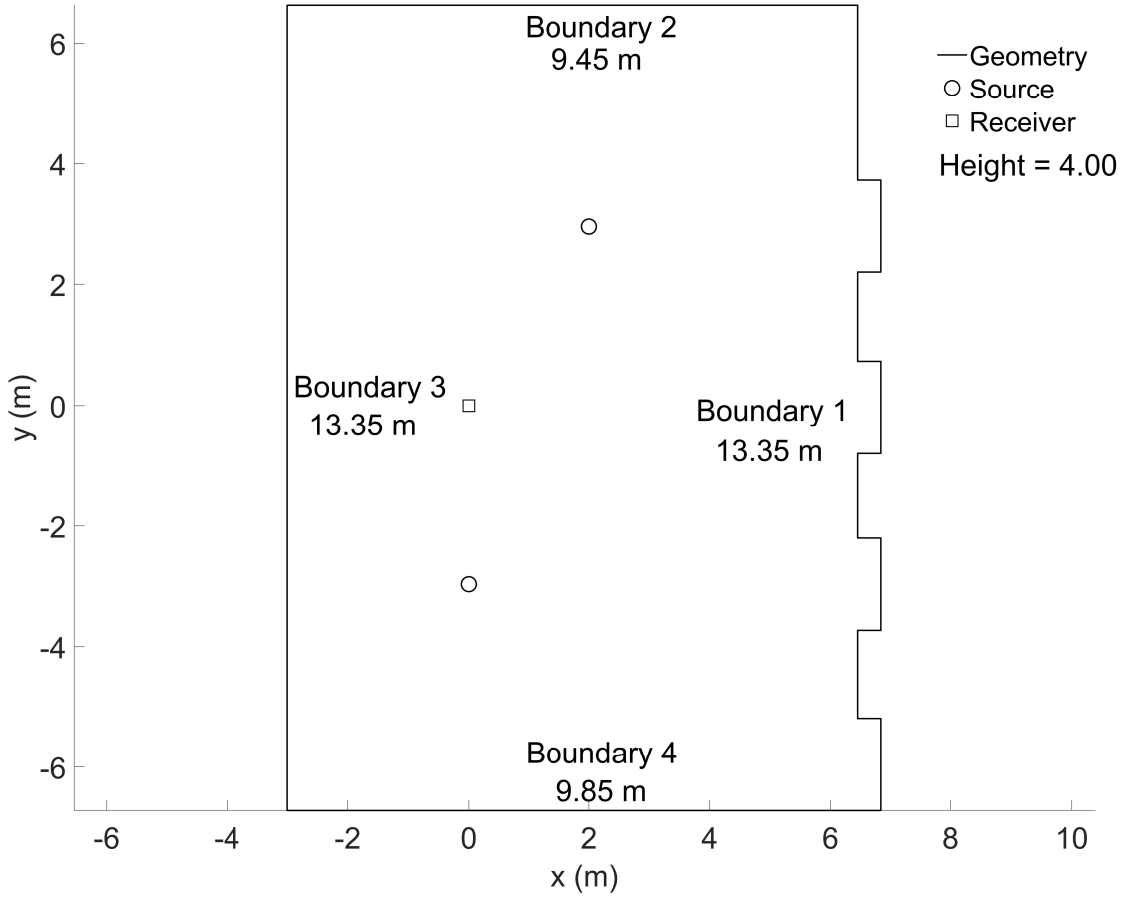

Figure 7.14: First source/receiver positions for Scenario One Third Cuboid-Shaped Room. Square marker denotes the receiver position and the circle markers denote the source positions.

axis. As discussed in Chapter 6 the steered response power map version of the EDESAR method is used to analyse the reflections in the simulated SRIR and the MVDR beamformer version is used for the real-world measurements.

7.5.1 Preliminary Testing: Ground-Truth

From the results in Table 7.4, it can be seen that when the geometry inference method is presented with ground truth values of ToA and DoA, it produces an exact estimate of the room's geometry – even when using all reflections within the first 32.01 ms (311 reflections in total). This result would suggest that any errors within the estimated geometry are more likely as a consequence of inaccuracies within the reflection detection and evaluation step. Furthermore, it would be expected that errors as a result of time- and direction-of-arrival estimation inaccuracies will vary as a result of where these inferred boundaries intersect with neighbouring boundaries. The resulting inferred geometry when using all reflections can be seen in Figure 7.21.

In Table 7.5, the results for exact DoA and ToA with normally distributed randomised error is

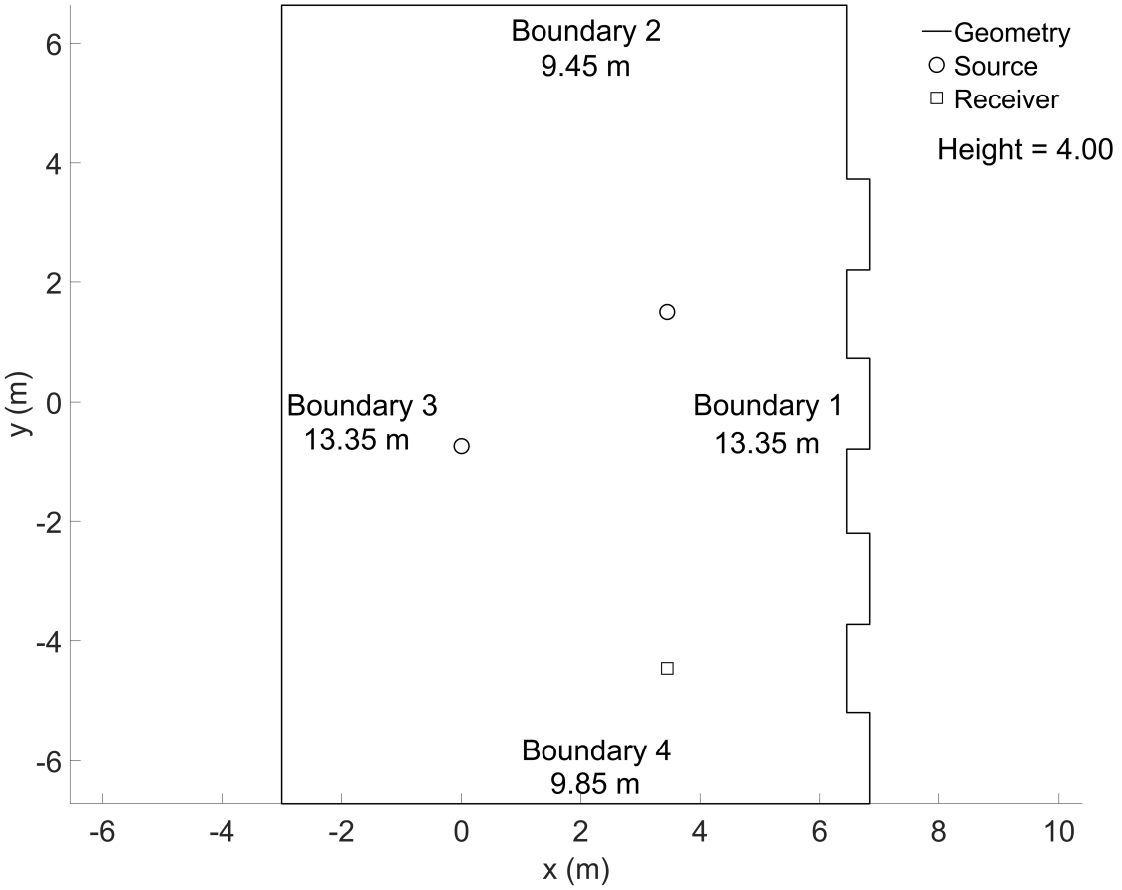

Figure 7.15: Second source/receiver positions for Scenario One Third Cuboid-Shaped Room. Square marker denotes the receiver position and the circle markers denote the source positions.

presented. Errors in ToA estimation will result in an under or overestimation of the distance between the receiver and the estimated image-source location. This consequently will result in a boundary parallel to the desired boundary being inferred, leading to increased ΔPosition and ΔLength errors. From the results presented in Table 7.5 and Figure 7.22, it can be seen that as ToA error increases so too does the positional error of the boundary. It can be seen that the resulting estimation error does not linearly increase as ToA error increases. This is as a consequence of using all reflections, irrelevant of reflection order, to estimate the shape of the room, resulting in multiple boundaries being inferred for a given boundary, and, as a consequence of the geometry validation process, generally only the closest boundary to the receiver being used. Furthermore, when a larger ToA error is observed for higher-order reflections the proposed method typically defines the previous-source position as being the source, which results in an inferred boundary positioned outside of the desired geometry and is generally invalidated during the boundary validation process. However, when this is not the case, typically as a result of a reflection between perpendicular boundaries, an increased Dihedral Angle or larger ΔLength is observed due to angled boundaries inferred at the corners of the room, such as with $\pm$362.81 μs, $\pm$408.16 μs,

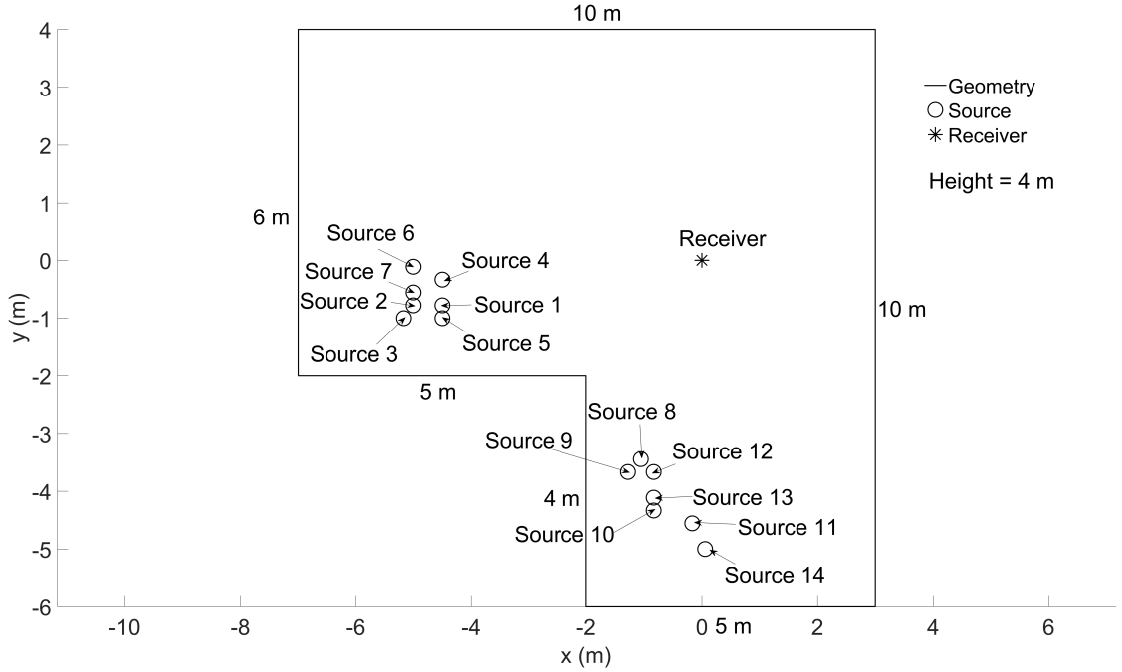

Figure 7.16: Geometry for Scenario Two L-Shaped Room One, image shows the 14 different source positions (Circle marker) and the receiver position (Square Marker) used when testing the proposed geometry inference method.

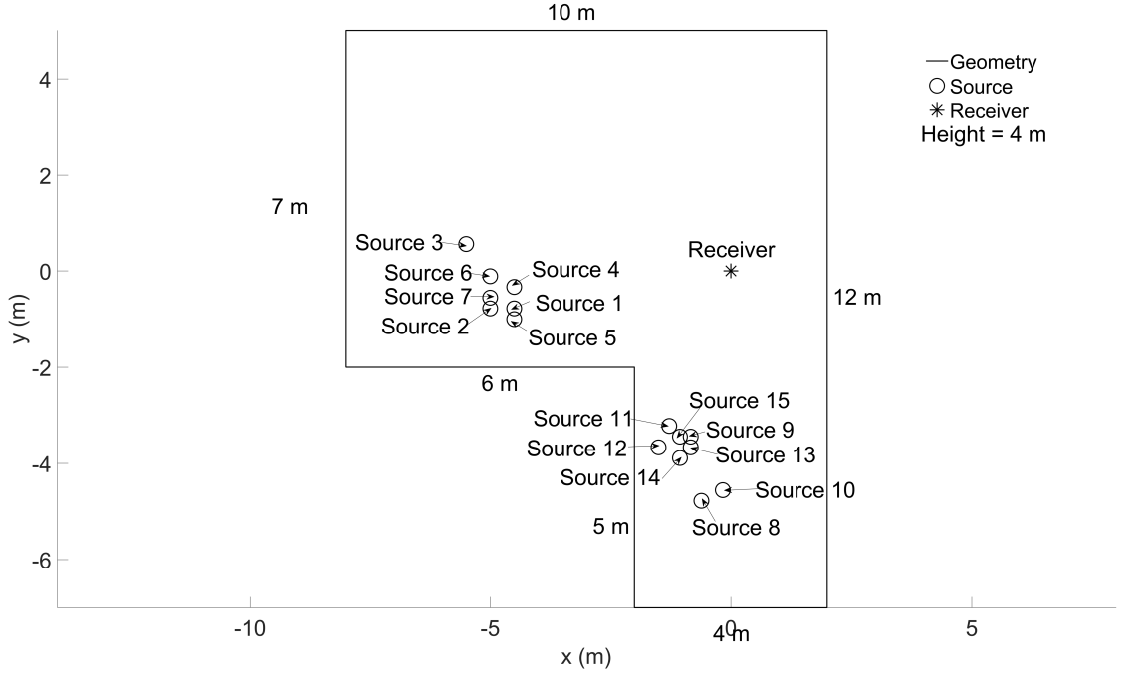

Figure 7.17: Geometry for Scenario Two L-Shaped Room Two, image shows the 15 different source (Circle marker) positions and the receiver (Square Marker) location used when testing the proposed geometry inference method.

$\pm$430.84 μs, and $\pm$476.19 μs. The best case, and cases with the worst ΔPosition, Dihedral Angle, and ΔLength, can be seen in Figure 7.23.

Measurement Set	Source Positions 1	Source Position 2
1	1	9
2	1	8
3	2	12
4	2	13
5	3	9
6	3	10
7	3	11
8	3	8
9	4	9
10	4	10
11	4	11
12	4	12
13	4	13
14	4	14
15	4	8
16	5	9
17	5	11
18	5	12
19	5	13
20	5	14
21	5	8
22	6	9
23	6	10
24	6	11
25	6	12
26	6	13
27	6	14
28	6	8
29	7	10
30	7	11
31	7	12
32	7	13
33	7	8

Table 7.1: Combinations of source positions used for each measurement set used in Scenario Two, L-Shaped Room One.

In Table 7.6, the results for exact ToA and elevation DoA, and azimuth DoA with normally distributed randomised errors can be seen. Any errors in azimuth DoA will result in a horizontal rotation of the inferred image-source's position around the receiver. In this case, the inferred boundary will be horizontally angled when compared to the desired boundary, resulting in a ΔPosition, Dihedral Angle, and ΔLength error. As with the ToA error, the severity of the room estimation error introduced as a result of under or overestimation of azimuth DoA ultimately depends on how all inferred boundaries intersect. From the results in Table 7.6 and Figure 7.24, it can be seen that, as with the ToA tests, there is not a linear relationship between azimuth DoA error and the proposed error metrics, again as a consequence of using all reflections for

Measurement Set	Source Positions 1	Source Position 2
1	1	10
2	1	13
3	2	10
4	2	13
5	4	10
6	4	13
7	5	13
8	5	15
9	6	13
10	7	15
11	7	13
12	7	15
13	5	12
14	6	12
15	7	12
16	1	11
17	1	14
18	2	9
19	2	11
20	2	14
21	4	8
22	4	9
23	4	11
24	5	11
25	5	14
26	6	9
27	6	11
28	6	14
29	7	9
30	7	11
31	7	14
32	3	9
33	3	11

Table 7.2: Combinations of source positions used for each measurement set used in Scenario Two L-Shaped Room Two.

which a previous-source has been found. This is particularly noticable for the $\pm 9^{\circ}$ angular error case, where a more accurate estimate of the room's geometry, and consequently lower ΔPosition, Dihedral Angle, and ΔLength, is observed compared to cases with lower angular errors, such as the $\pm 8^{\circ}$, 6°, 5°, 2°, and 1°. In addition to this, it is important to note that the cases where larger errors in ΔLength are observed, typically are as a result of multiple boundaries being inferred for a given desired boundary. The best case and cases with the worst ΔPosition, Dihedral Angle, and ΔLength can be seen in Figure 7.25. For the case of the largest error in ΔPosition and Dihedral Angle, while the correct boundaries have been inferred, six boundaries have been inferred at the corners between boundaries one and four, four between

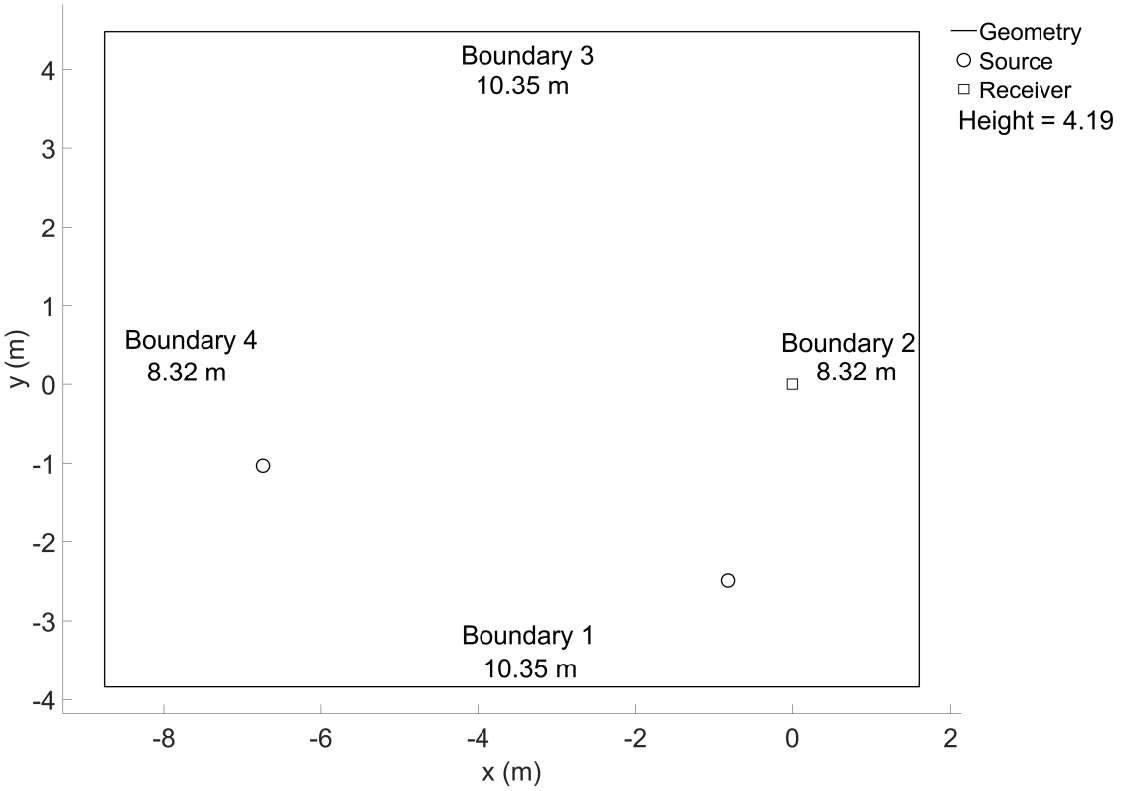

Figure 7.18: Geometry for the Scenario Three cuboid-shaped, room measurement set one. Square marker denotes the receiver position and the circle markers denote the source positions.

Threshold	Value
$\epsilon_{\widetilde{\mathbf{s}}}$	30 cm
ϵ_o	15 cm
ϵ_l	10 cm
$\epsilon_\angle$	0.05
$\epsilon_{\widetilde{\mathbf{n}}}$	0.05
ϵ_{par}	0.1
ϵ_{point}	0.1

Table 7.3: The empirically defined values of $\epsilon_{\widetilde{\mathbf{s}}}$, ϵ_o, ϵ_l, $\epsilon_\angle$, $\epsilon_{\widetilde{\mathbf{n}}}$, ϵ_{par}, and ϵ_{point}, used when testing the proposed geometry inference method. These are defined to reduce the number of inaccurately inferred boundaries while ensured all first-order reflections are assigned to the correct boundaries.

Test Case	**ΔPosition**	**Weighted Dihedral angle**	**ΔLength**
First-Order	0.00 cm	0.00°	0.00 cm
All Reflections	0.00 cm	0.00°	0.00 cm

Table 7.4: Analysis of geometry inference method when presented with exact time- and direction-of-arrival values for 311 reflections. Results presented as the RMS ΔPosition, weighted dihedral angle, and ΔLength.

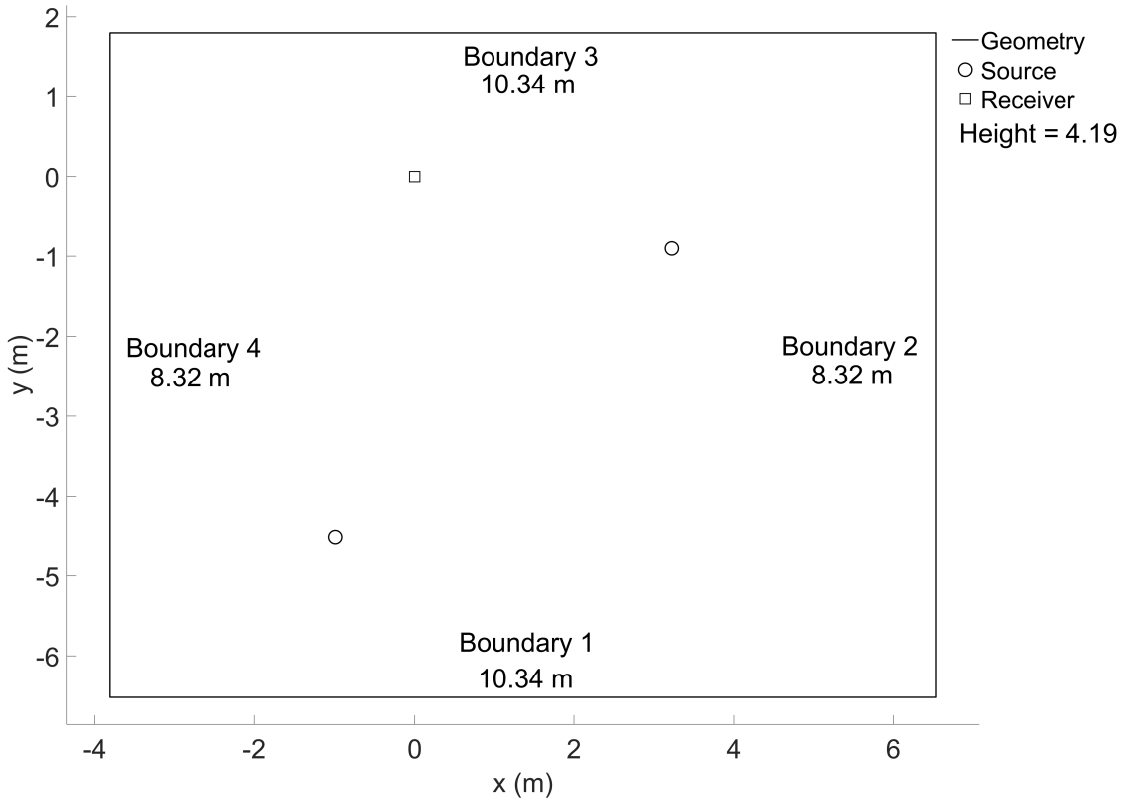

Figure 7.19: Geometry for the Scenario Three cuboid-shaped, room measurement set two. Square marker denotes the receiver position and the circle markers denote the source positions.

four and three, and three between three and two, as a result of being unable to find the correct previous source for higher-order reflections. These erroneously inferred corner boundaries have consequently resulted in inaccurate estimation of the boundaries location as a result of defining the room's shape based on intersections between the available boundaries.

In Table 7.7, the results for exact ToA and azimuth DoA, and elevation DoA with normally distributed randomised errors can be seen. Any errors in elevation will result in a vertical rotation of a given image-source around the receiver. The consequence of this being that any inferred boundaries will be vertically angled, and potentially invalidated by the image-reversion process. Therefore, elevation errors can potentially result in boundaries produced by first-order reflections being ignored as a result of the assumption that all walls are perpendicular to the floor and ceiling. It would, therefore, be expected that larger errors in the ΔPosition and Dihedral Angle will be observed when an elevation error has resulted in the first-order reflection for a given boundary across all measurement positions being ignored. From the results in Table 7.7 and Figure 7.26, it can be seen that the estimation error of the geometry inference method is generally lower than that observed for azimuth DoA errors. The two cases when a ΔPosition error

Figure 7.20: Image of the room setup for Scenario Three, showing the Genelec 8030 and EigenMike. As can be seen there is curtain coverage across the right wall which occludes the windows, and curtains positioned in the corners of the room hiding large electrical outlets. On the ceiling there are light fixtures, railing, extractor fans, and a series of large rectangular pipes.

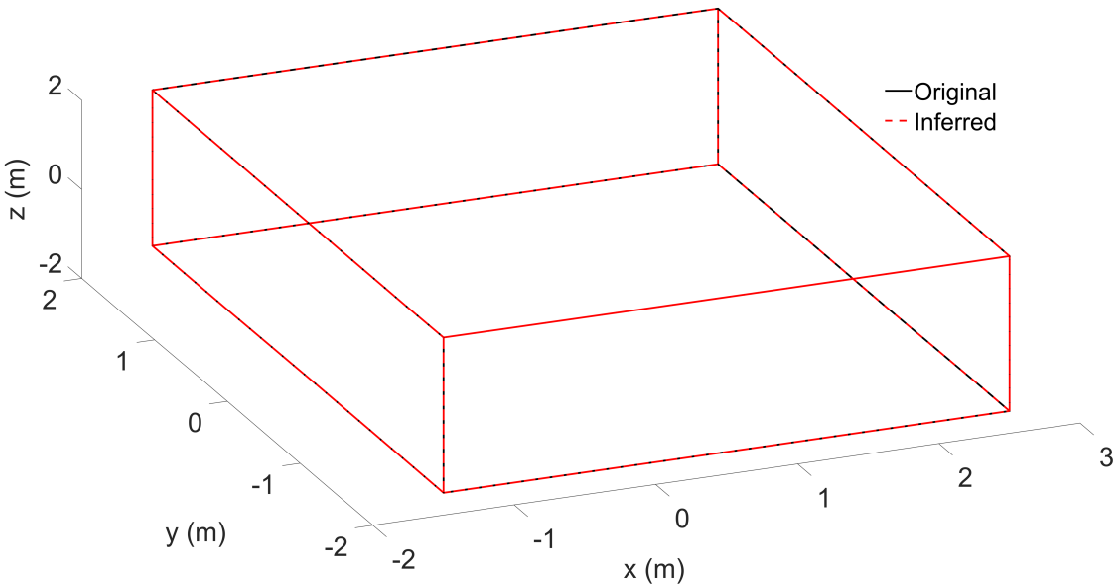

Figure 7.21: Inferred geometry (dashed red line) and desired geometry (solid line) for Ground-Truth test.

greater than 10 cm is observed, $\Delta\phi = 7°$ and $8°$, correspond to cases when an elevation error of $\Delta\phi \geq \pm 5°$ is observed for a boundary's first-order reflections in both measurement positions (see Table 7.8), and a higher-order reflection between perpendicular boundaries is not accurately

ToA Error	**ΔPosition**	**Weighted Dihedral angle**	**ΔLength**
±0 μs	0.00 cm	0.00°	0.00 cm
±22.68 μs	0.00 cm	0.00°	0.00 cm
±45.35 μs	0.58 cm	0.00°	1.41 cm
±68.03 μs	0.91 cm	0.00°	1.58 cm
±90.70 μs	1.41 cm	0.00°	3.16 cm
±113.38 μs	1.53 cm	0.00°	3.00 cm
±136.05 μs	1.78 cm	0.00°	4.30 cm
±158.73 μs	1.58 cm	0.00°	3.54 cm
±181.41 μs	2.12 cm	0.00°	5.15 cm
±204.08 μs	0.75 cm	0.40°	3.21 cm
±226.76 μs	1.35 cm	0.00°	2.55 cm
±249.43 μs	1.00 cm	0.00°	2.24 cm
±272.11 μs	2.89 cm	0.00°	6.71 cm
±294.78 μs	2.61 cm	0.00°	6.04 cm
±317.46 μs	3.92 cm	0.00°	9.06 cm
±340.14 μs	5.00 cm	0.00°	10.78 cm
±362.81 μs	3.98 cm	3.27°	7.84 cm
±385.49 μs	3.19 cm	0.00°	5.52 cm
±408.16 μs	1.97 cm	0.24°	5.51 cm
±430.84 μs	5.52 cm	0.32°	12.34 cm
±453.51 μs	1.96 cm	0.00°	4.61 cm
±476.19 μs	4.20 cm	0.00°	358.56 cm

Table 7.5: Analysis of geometry inference method when presented with time- and direction-of-arrival values for 311 reflections with randomly generated and normally distributed errors introduced to the time-of-arrival values. Results presented as the RMS ΔPosition, weighted dihedral angle, and ΔLength.

assigned a previous-source. However, in the case of $\Delta\phi = 6°$, while both first-order reflections for boundary two are ignored, a second-order reflection between perpendicular boundaries results in the boundary being more accurately inferred. As with the azimuth DoA error, cases with larger ΔLengths ($\pm5°$, $\pm7° - -10°$) correspond to cases when the proposed geometry inference method has inferred additional boundaries outside of the desired room's geometry for a given boundary. These results would suggest that the method is more robust to elevation estimation inaccuracies than azimuth, as long as $\Delta\phi < \pm5°$ for at least one of a given boundary's first-order reflections. The best case, and cases with the worst ΔPosition, Dihedral Angle, and ΔLength, can be seen in Figure 7.27.

In Table 7.9 and Figure 7.28, analysis of the proposed geometry inference method and reflection detection method is presented for decreasing SNR (increases in the level of inteferring noise). In this scenario, the geometry inference process is performed using reflections detected within the simulated SRIR. These results show that comparable accuracy is achieved for ΔPosition

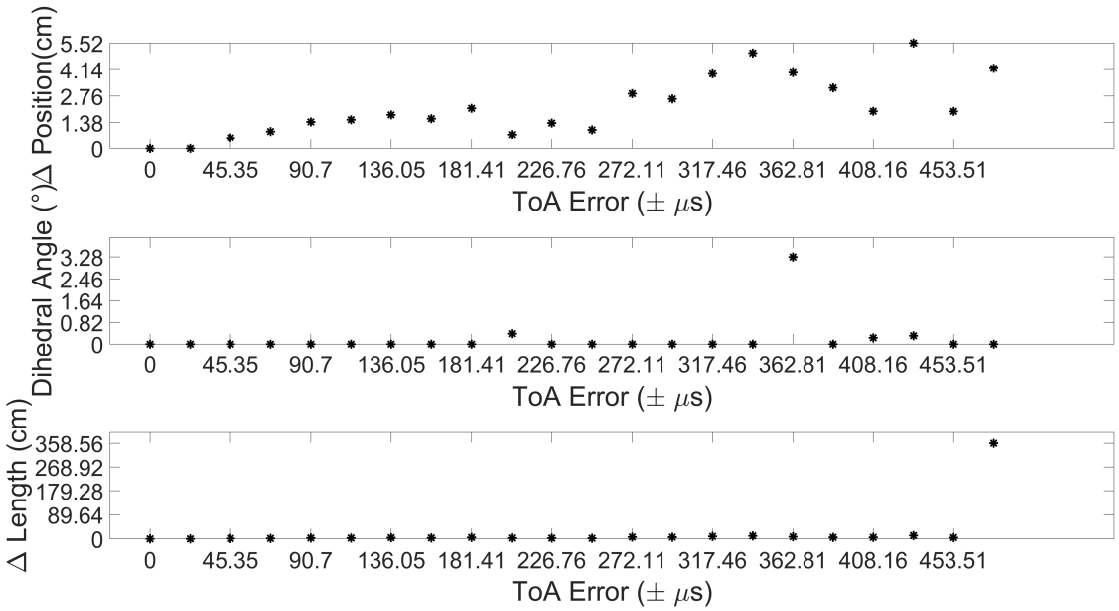

Figure 7.22: Comparison of the ΔPosition, Weighted Dihedral Angle, and ΔLength compared to the magnitude of the introduced time-of-arrival errors.

Azimuth DoA Error $\Delta\theta$	**ΔPosition**	**Weighted Dihedral angle**	**ΔLength**
±0	0.00 cm	0.00°	0.00 cm
±1	29.27 cm	6.83°	920.05 cm
±2	11.05 cm	2.42°	6653.42 cm
±3	3.28 cm	2.39°	8.75 cm
±4	4.20 cm	1.89°	9.02 cm
±5	9.33 cm	3.65°	24.70 cm
±6	12.65 cm	4.63°	34.26 cm
±7	8.15 cm	5.20°	51.24 cm
±8	54.76 cm	26.16°	96.05 cm
±9	5.80 cm	3.50°	19.98 cm
±10	26.01 cm	11.11°	98.50 cm

Table 7.6: Analysis of geometry inference method when presented with time- and direction-of-arrival values for 311 reflections with randomly generated and normally distributed errors introduced to the azimuth direction-of-arrival values. Results presented as the RMS ΔPosition, weighted dihedral angle, and ΔLength.

and Dihedral Angle while SNR is greater than or equal to 20 dB. For these cases a maximum differece of 0.88 cm is observed for ΔPosition and a maximum difference of 1.04° for Dihedral Angle. Furthermore, except for the 30 dB and 25 dB cases, a comparable ΔLength is observed, in both outlier cases an additional boundary has been inferred for Boundary two, which exceeds the desired boundary length. Once the SNR drops below 20 dB, the proposed reflection detection method produces false-positive detections within the noise floor of the SRIR. These

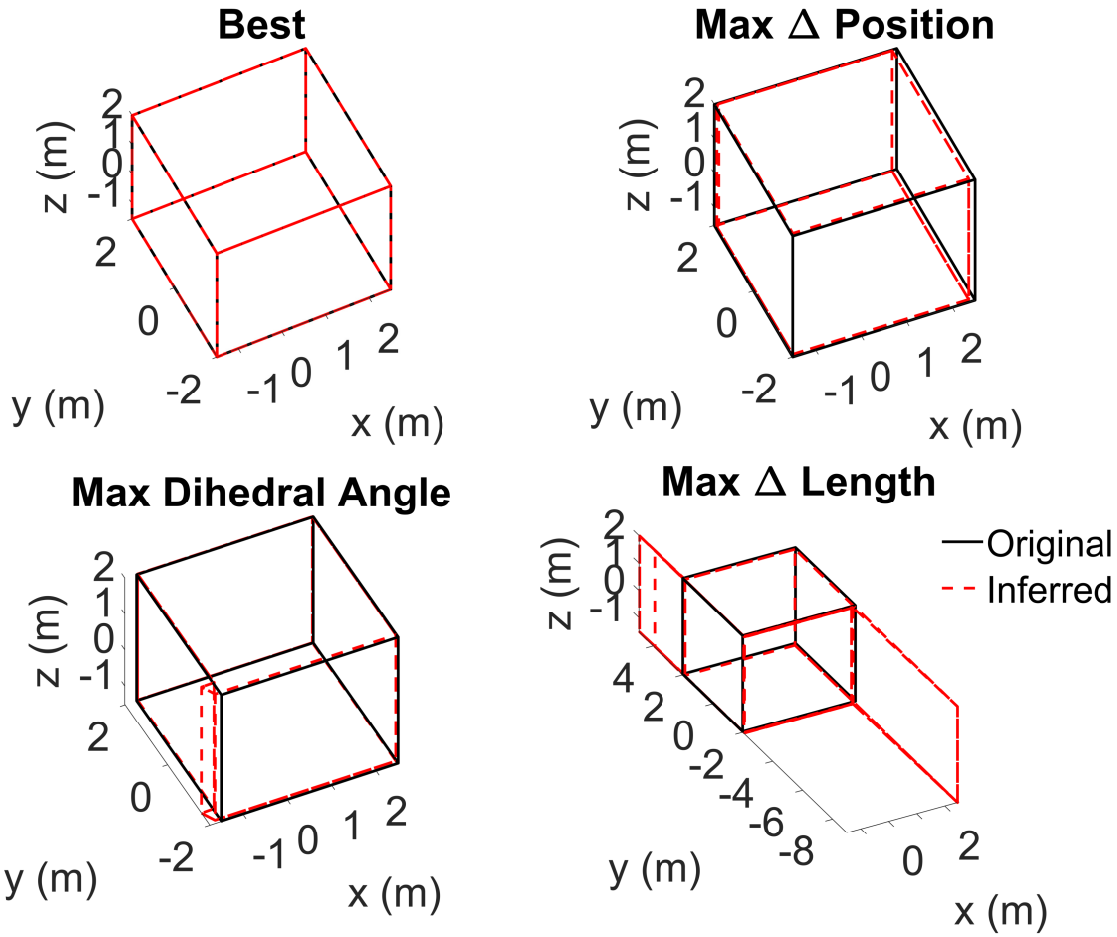

Figure 7.23: Inferred geometry (dashed red line) and desired geometry (solid line) for Ground-Truth test with randomly generated and normally distributed errors added to the ToA values. The best case is 0 μs, largest ΔPosition error is 430.84 μs, largest Dihedral Angle error is 362.81 μs, and largest Δ length is 476.19 μs.

false-positive detections result in an underestimation of the floor locations by 21 cm for all cases when SNR $\leq$ 15 dB, Boundary three is incorrectly inferred for SNRs of 15, 5, and 0 dB, and Boundaries two and four are incorrectly inferred for the case when SNR was 10 dB. In practice, as the SNR decreases below 20 dB, and consequently the number of false-positive detections increases, the proposed geometry inference method is more likely to produce erroneous estimates of the room's geometry. However, the severity of this error depends on the location of any inferred boundaries produced by false-positive detections, and how they intersect with neighboring boundaries. The best case, and cases with the worst ΔPosition, Dihedral Angle, and ΔLength, can be seen in Figure 7.29.

When analysing the accuracy of the reflection detection process, the DoA for all first-order reflections are estimated to within 2° of the expected value up to a SNR of 10dB, and except for the ceiling, the ToA error is less than one sample, and therefore, is more likely a product of

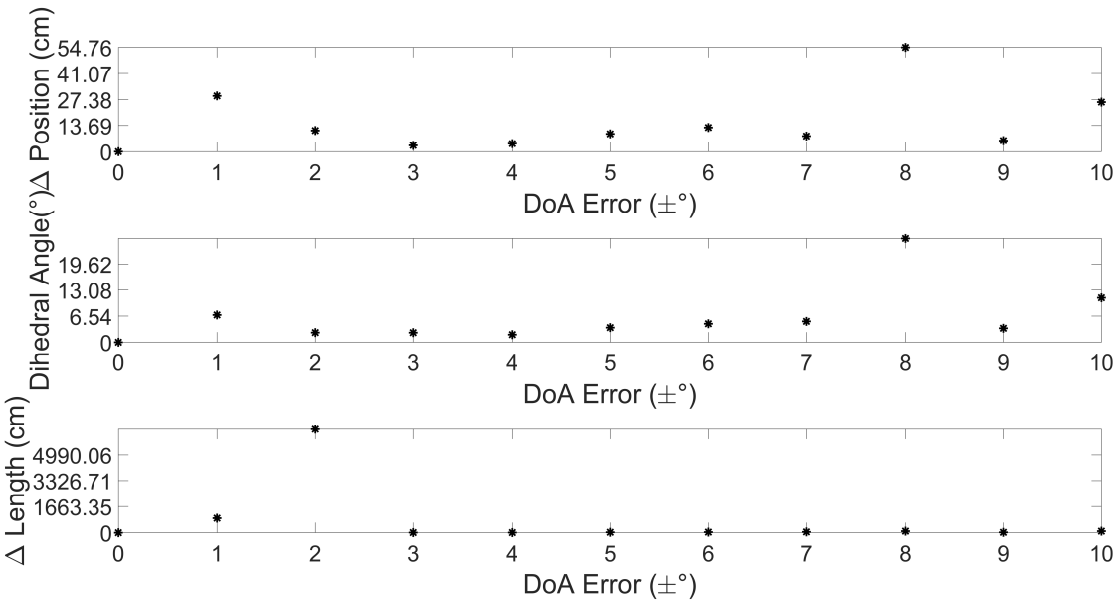

Figure 7.24: Comparison of the ΔPosition, Weighted Dihedral Angle, and ΔLength compared to the magnitude of the introduced azimuth direction-of-arrival errors.

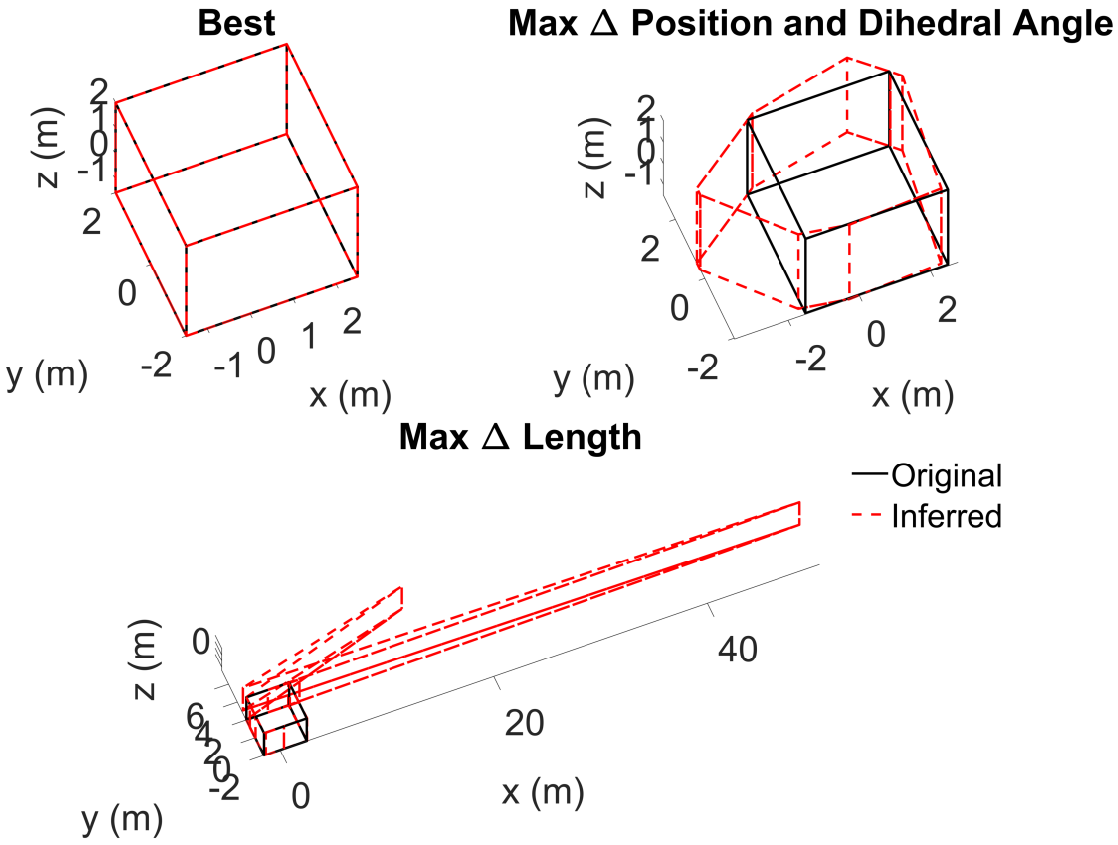

Figure 7.25: Inferred geometry (dashed red line) and desired geometry (solid line) for Ground-Truth test with randomly generated and normally distributed errors added to the Azimuth DoA values. The best case is 0°, the case with the largest ΔPosition error and Dihedral Angle error is 8°, and the case with the largest Δ length is 2°.

Elevation DoA Error $\Delta\phi$	**ΔPosition**	**Weighted Dihedral angle**	**ΔLength**
±0	0.00 cm	0.00°	0.00 cm
±1	1.38 cm	0.28°	3.20 cm
±2	2.74 cm	0.23°	4.24 cm
±3	8.39 cm	0.86°	18.87 cm
±4	5.48 cm	0.00°	9.49 cm
±5	2.06 cm	1.97°	392.07 cm
±6	2.89 cm	0.00°	5.00 cm
±7	34.25 cm	14.12°	33.23 cm
±8	48.33 cm	14.71°	60.79 cm
±9	5.31 cm	0.04°	125.36 cm
±10	0.91 cm	0.12°	815.03 cm

Table 7.7: Analysis of geometry inference method when presented with time- and direction-of-arrival values for 311 reflections with randomly generated and normally distributed errors introduced to the elevation direction-of-arrival values. Results presented as the RMS ΔPosition, weighted dihedral angle, and ΔLength.

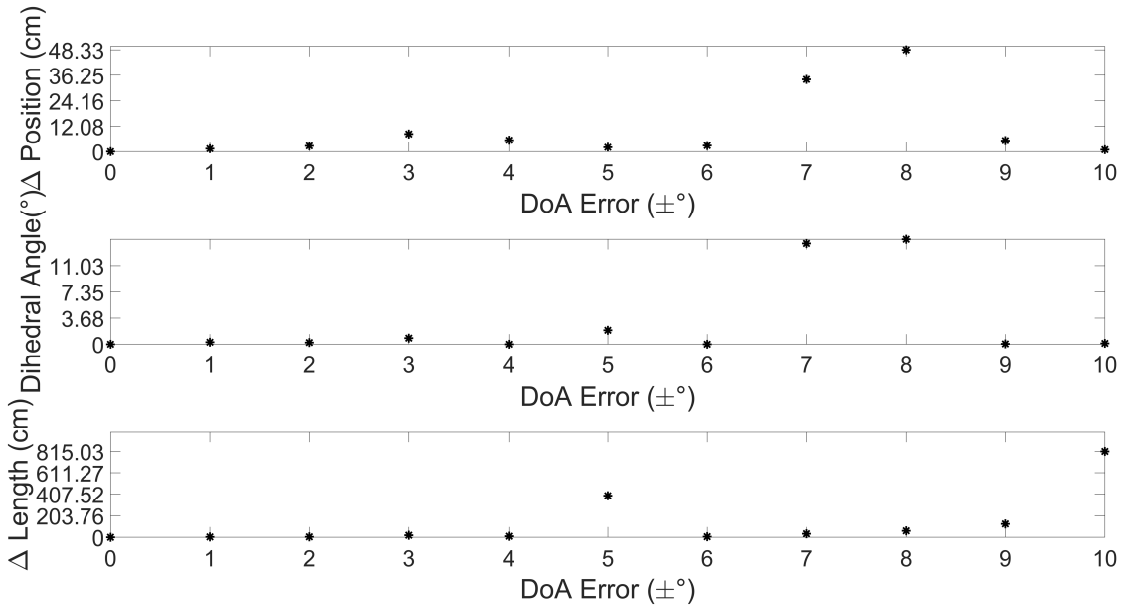

Figure 7.26: Comparison of the ΔPosition, Weighted Dihedral Angle, and ΔLength compared to the magnitude of the introduced elevation direction-of-arrival errors.

the discrete sampling of a continuous-time signal. Furthermore, it is important to note that the larger ToA and DoA estimates are generally observed when analysing simultaneously arriving reflections and as reflection density increases.

7.5.2 Test Case One

The result for the first cuboid room, as presented in Figure 7.30 and Table 7.10, show that the the general shape of the room has been inferred, with all but the ceiling having a boundary ΔPosition

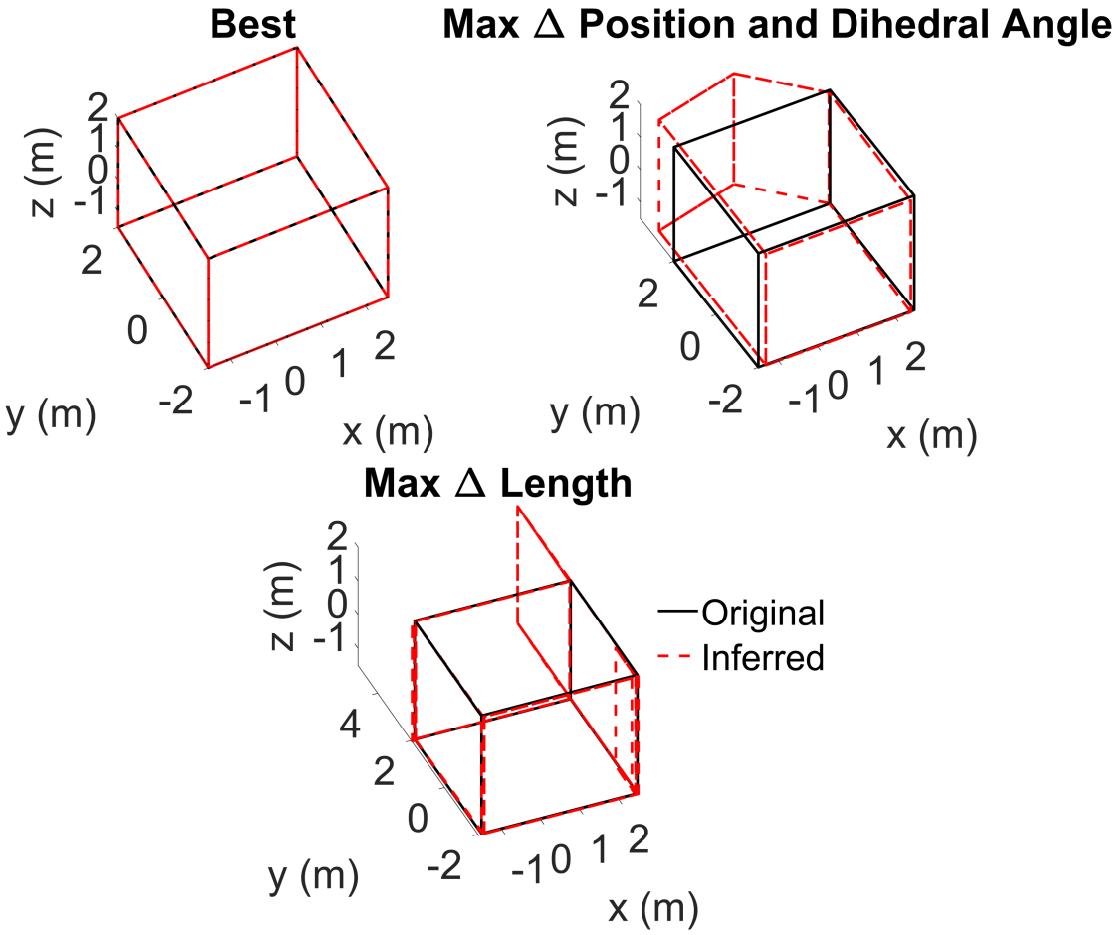

Figure 7.27: Inferred geometry (dashed red line) and desired geometry (solid line) for Ground-Truth test with randomly generated and normally distributed errors added to the Elevation DoA values. The best case is 0°, the case with the largest ΔPosition error and Dihedral Angle error is 8°, and the case with the largest Δ length is 5°.

	One		Two		Three		Four		Ceiling		Floor	
	S1	**S2**	**S1**	**S2**	**S1**	**S2**	**S1**	**S2**	**S1**	**S2**	**S1**	**S2**
±1°	0.75°	0.03°	0.20°	0.52°	0.17°	0.02°	0.81°	0.02°	0.52°	0.17°	0.90°	0.89°
±2°	1.09°	0.96°	0.99°	0.77°	0.04°	1.63°	0.29°	1.31°	0.73°	0.04°	1.09°	1.82°
±3°	2.99°	1.64°	1.10°	1.49°	1.18°	2.30°	1.52°	0.12°	2.59°	1.18°	2.43°	2.61°
±4°	3.41°	1.72°	2.60°	2.91°	0.39°	2.05°	2.63°	1.64°	3.60°	0.39°	3.92°	3.27°
±5°	2.82°	2.95°	2.62°	4.21°	2.98°	3.23°	1.88°	2.51°	0.84°	2.98°	3.73°	4.11°
±6°	0.51°	5.86°	5.59°	5.75°	2.51°	0.89°	3.35°	4.59°	2.11°	2.51°	4.55°	1.27°
±7°	5.00°	5.10°	1.53°	2.35°	4.99°	1.11°	2.14°	6.21°	4.24°	4.99°	3.74°	2.21°
±8°	4.66°	2.60°	6.54°	3.14°	6.41°	3.00°	6.34°	6.45°	4.51°	6.41°	1.91°	4.78°
±9°	8.17°	4.53°	7.38°	0.73°	5.72°	1.59°	4.19°	8.50°	5.07°	5.72°	4.98°	3.20°
±10°	1.82°	0.56°	9.32°	8.64°	9.04°	1.18°	0.83°	5.81°	5.63°	9.04°	2.75°	6.62°

Table 7.8: Absolute value of elevation direction-of-arrival error introduced to the first-order reflections from each boundary for each source (**S1** and **S2**). Values in red indicate a first-order reflection that is ignored and the boundary is not inferred correctly, and the values in blue indicate a first-order reflection that is ignored, but the boundary is defined using a higher-order reflection.

SNR	ΔPosition	Weighted Dihedral angle	ΔLength	ΔToA	$\Delta\theta$	$\Delta\phi$	False
∞	3.79 cm	0.96°	9.57 cm	171.13 μs	32.01°	25.45°	1
60 dB	3.79 cm	0.96°	9.57 cm	156.26 μs	37.17°	26.48°	1
55 dB	3.79 cm	0.96°	9.57 cm	156.26 μs	37.17°	26.48°	1
50 dB	3.79 cm	0.96°	9.57 cm	156.26 μs	37.17°	26.48°	1
45 dB	3.79 cm	0.96°	9.57 cm	156.26 μs	37.17°	26.49°	1
40 dB	3.46 cm	1.01°	7.82 cm	156.26 μs	37.19°	26.49°	1
35 dB	3.46 cm	1.01°	7.82 cm	157.16 μs	35.72°	26.71°	1
30 dB	3.90 cm	1.44°	99.91 cm	151.49 μs	33.22°	25.76°	1
25 dB	4.18 cm	1.97°	100.02 cm	151.36 μs	34.47°	26.85°	2
20 dB	4.34 cm	2.00°	11.05 cm	155.11 μs	29.24°	24.30°	1
15 dB	118.28 cm	25.84°	1149.79 cm	174.85 μs	41.26°	27.29°	17
10 dB	114.43 cm	6.86°	3386.19 cm	198.52 μs	49.41°	30.58°	30
5 dB	118.27 cm	24.58°	969.02 cm	204.73 μs	44.00°	26.96°	28
0 dB	50.61 cm	24.07°	637.80 cm	222.79 μs	46.49°	26.52°	26

Table 7.9: Analysis of geometry inference method when presented with the ground-truth SRIR with noise added as SNR of no noise and 60 dB to 0 dB in 5 dB steps. Results presented as the RMS ΔPosition, weighted dihedral angle, ΔLength, the RMS time-of-arrival error across all detections ΔToA, RMS azimuth direction-of-arrival error across all detections ($\Delta\theta$), RMS elevation direction-of-arrival error across all detections ($\Delta\phi$), and number false-positives (False).

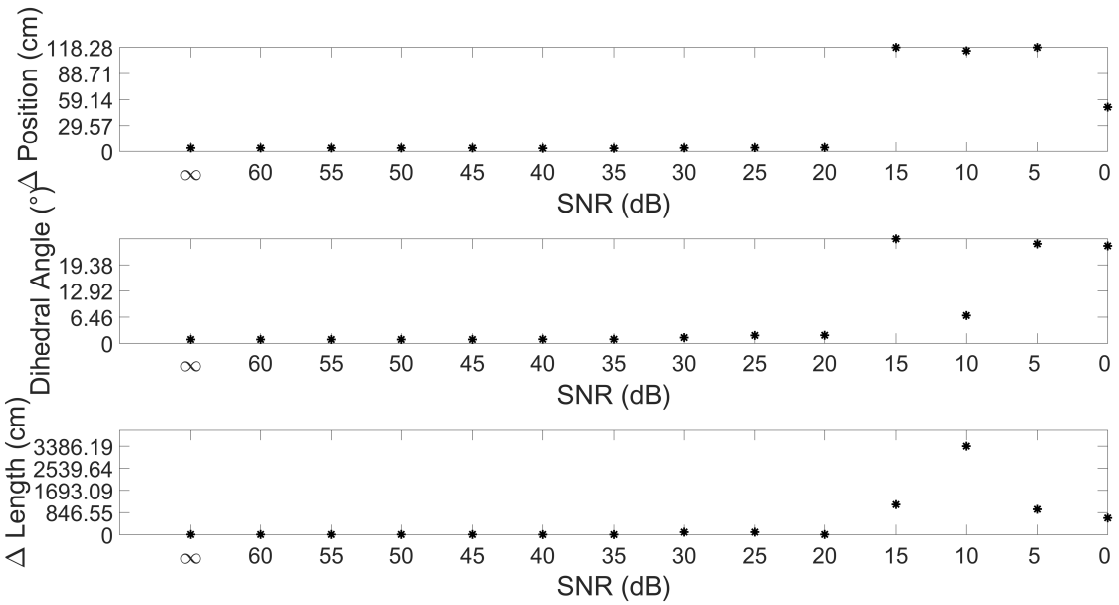

Figure 7.28: Comparison of the ΔPosition, Weighted Dihedral Angle, and ΔLength over different signal-to-noise ratios.

$\leq$ 2 cm. The higher ΔPosition for the ceiling is likely as a result of underestimation of the ToA for the corresponding reflection (1.4 ms and 0.57 ms for source one and two respectively). Furthermore, boundaries Two and Three have additional boundaries inferred on their corners as a result of incorrectly assigned previous-source candidates for their second-order reflections.

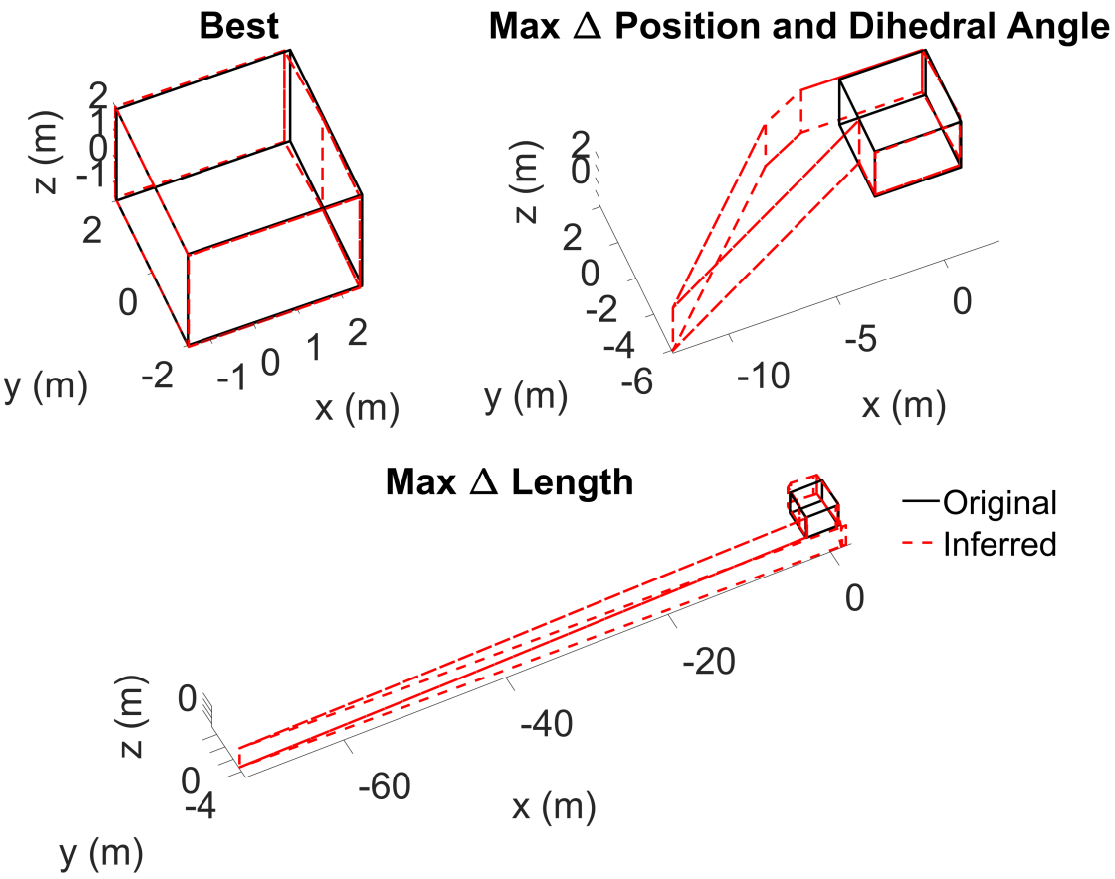

Figure 7.29: Inferred geometry (dashed red line) and desired geometry (solid line) for SNR tests. The best case is 60 dB, the case with the largest ΔPosition error and Dihedral Angle error is 15 dB, and the case with the largest Δ length is 10 dB.

This results in a slightly larger dihedral angle for these boundaries, and impacts the ΔLength error for surrounding boundaries. However, the weighted dihedral angles are still close to that of the other inferred boundaries.

Boundary	**ΔPosition**	**Dihedral Angle**	**Weighted Dihedral Angle**	**ΔLength**	**δLength**
One	1.18 cm	0.57°	0.57°	0.02 cm	0.00%
Two	0.00 cm	22.50°	0.95°	1.49 cm	0.37%
Three	1.52 cm	13.58°	0.90°	11.51 cm	2.88%
Four	0.90 cm	7.63°	0.87°	1.80 cm	0.45%
Floor	2.00 cm	N/A	N/A	N/A	N/A
Ceiling	8 cm	N/A	N/A	N/A	N/A
Mean	2.27 cm	19.00°	0.82°	3.71 cm	N/A

Table 7.10: Results for Scenario One: Cuboid Room One, presenting the four error metrics: difference in position (ΔPosition), dihedral angle, weighted dihedral angle, difference in boundary length (Δlength), and relative error of the inferred boundaries length (δLength).

The results for the larger second cuboid room, as presented in Figure 7.31 and Table 7.11, again

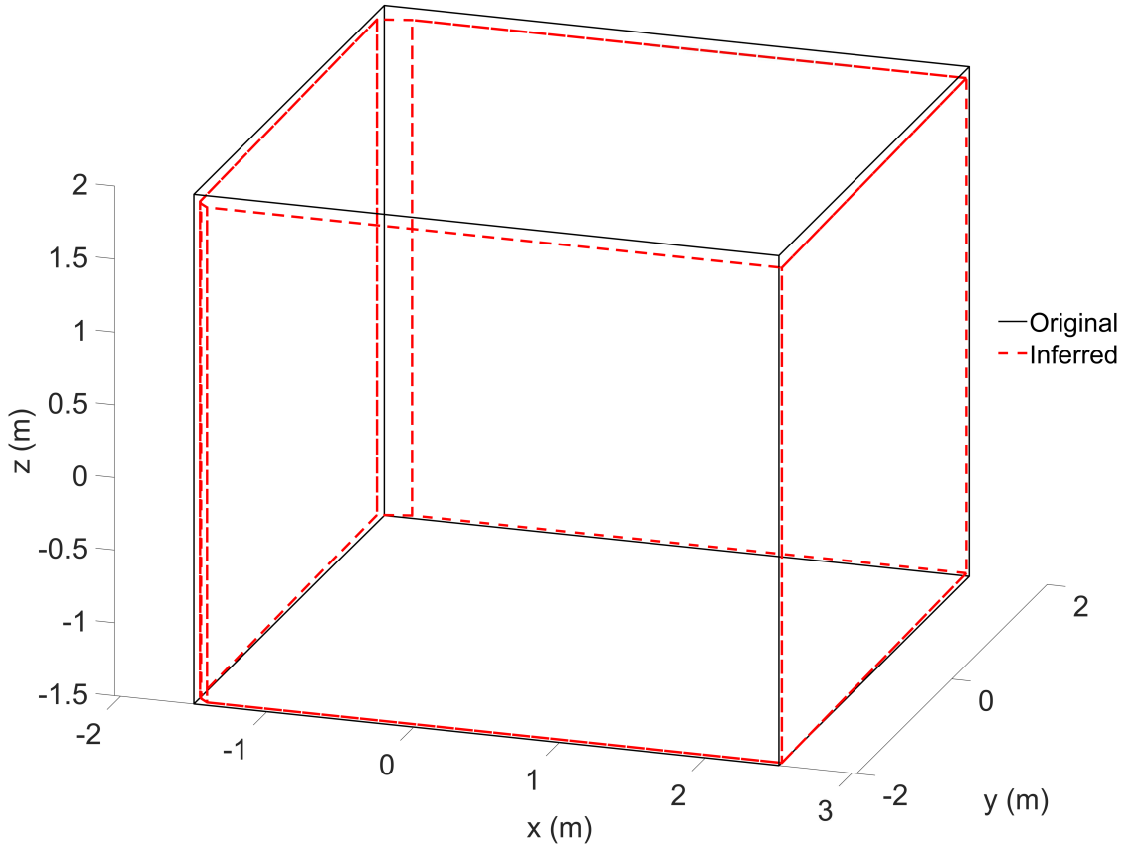

Figure 7.30: Inferred geometry (dashed red line) and desired geometry (solid line) for Scenario One - Cuboid One.

show that the general shape of the room has been inferred. However, there is a marginally larger boundary positional error for the walls when compared to the smaller cuboid room, with a maximum difference of 6.00 cm. As with the smaller cuboid room, there are additional, angled, inferred boundaries in the corners of the room, in this case at the point of intersection between boundaries Three and Four, producing a slightly larger dihedral angle for Boundary Four. However, the weighted dihedral angles are still close to that of the other inferred boundaries.

As with the previous two cases the general shape of the Octagonal Prism has been inferred, as presented in Figure 7.32 and Table 7.12, with a maximum ΔPosition of only 4.88 cm. In this example both boundaries Five and Seven are slightly angled, with a dihedral angle of 3.98° and 3.43° respectively, and this is as a result of a DoA estimation error of 3.92° and 3.65° for the first-order reflections from boundary Five and Seven. The consequence of this being that the length of boundary Six has been underestimated with a ΔLength error of 22.59 cm.

In Figure 7.33 and Table 7.13 the results for the first non-convex example, the L-shaped Room, can be seen. These again show that the shape of the room has been correctly inferred, however, there is an increase in the boundary positional error with a maximum error of 7.62 cm, although

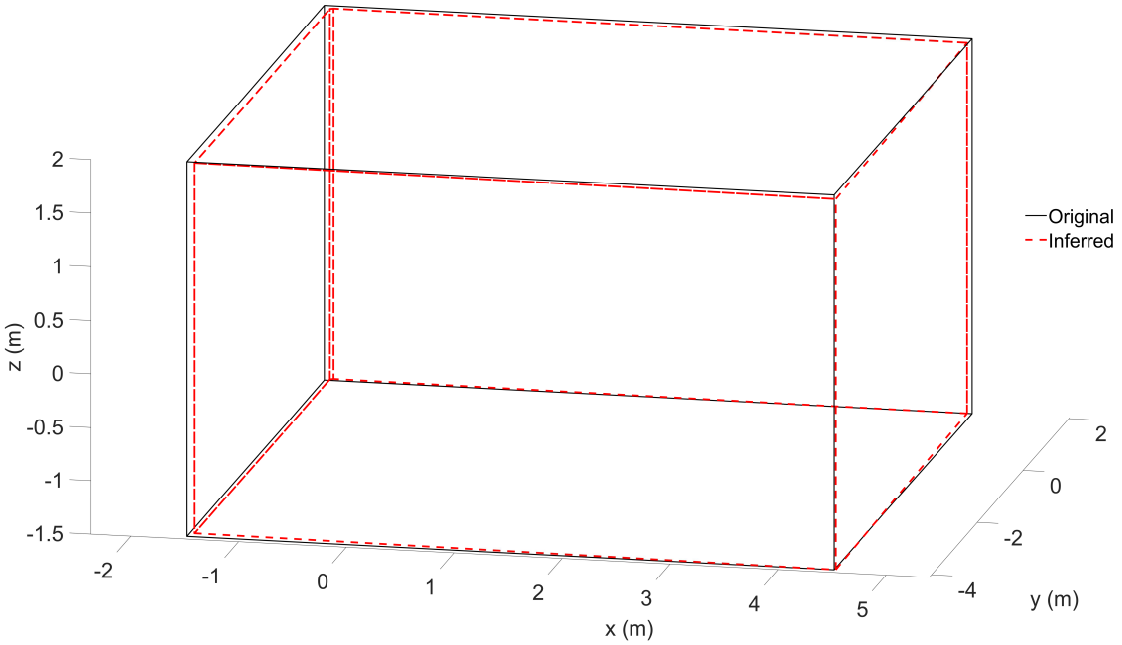

Figure 7.31: Inferred geometry (dashed red line) and desired geometry (solid line) for Scenario One - Cuboid Two.

Boundary	**ΔPosition**	**Dihedral Angle**	**Weighted Dihedral Angle**	**ΔLength**	**δLength**
One	1.73 cm	0.57°	0.57°	0.97 cm	0.16%
Two	3.63 cm	1.15°	1.15°	2.88 cm	0.48%
Three	6.00 cm	0.00°	0.00°	9.00 cm	1.50%
Four	5.26 cm	17.14°	0.79°	8.36 cm	1.39%
Floor	2.00 cm	N/A	N/A	N/A	N/A
Ceiling	2.00 cm	N/A	N/A	N/A	N/A
Mean	3.44 cm	4.72°	0.63°	5.30 cm	N/A

Table 7.11: Results for Scenario One: Cuboid Room Two, presenting the four error metrics: difference in position (ΔPosition), dihedral angle, weighted dihedral angle, difference in boundary length (Δlength), and relative error of the inferred boundaries length (δLength).

the overall performance is still comparable. As with the two cuboid rooms, the L-shaped Room has additional, angled, inferred boundaries at the corners, in this case at the points of intersection between boundaries One and Six and boundaries Five and Six. This produces a larger averaged dihedral angle for boundaries One and Five, and as a result boundary Six has a larger ΔLength error. However, the weighted dihedral angles are still close to that of the other inferred boundaries.

The T-shaped Room has the largest error values observed for the examples in Scenario One, as seen Figure 7.34 and Table 7.14, with a maximum boundary position error of 31.02 cm.

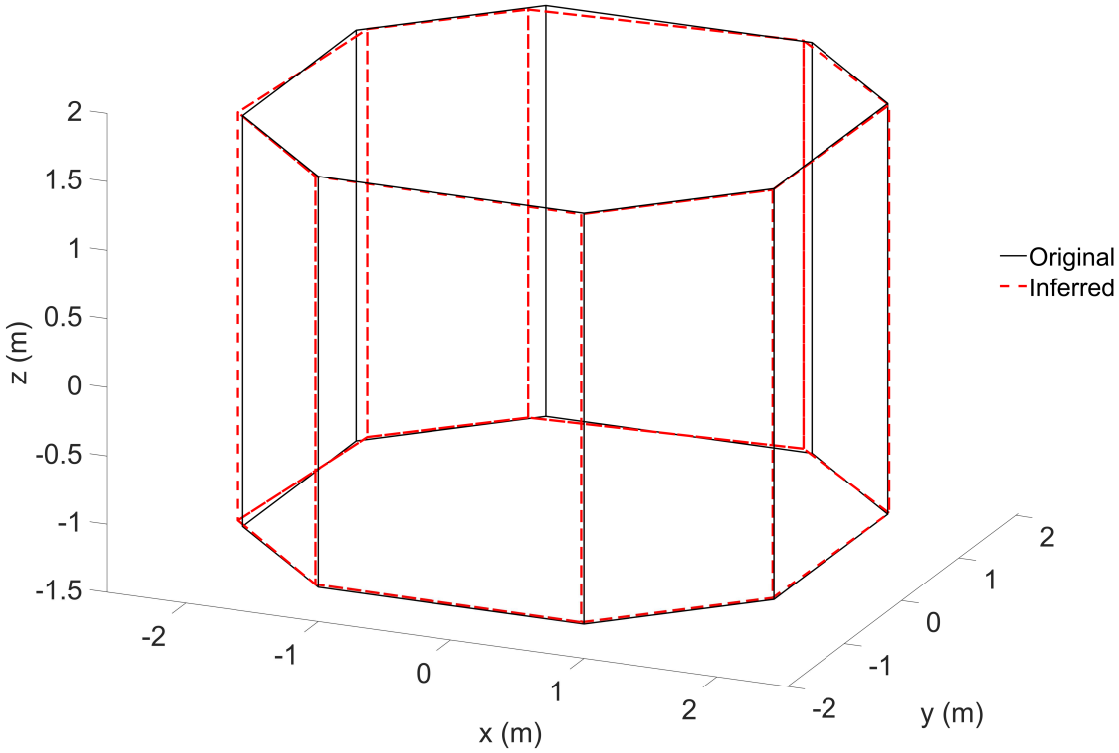

Figure 7.32: Inferred geometry (dashed red line) and desired geometry (solid line) for Scenario One - Octagonal Room.

Boundary	**ΔPosition**	**Dihedral Angle**	**Weighted Dihedral Angle**	**ΔLength**	**δLength**
One	1.15 cm	0.57°	0.57°	1.01 cm	0.50%
Two	1.08 cm	0.29°	0.29°	0.71 cm	0.50%
Three	0.61 cm	0.57°	0.57°	0.01 cm	0%
Four	1.41 cm	1.34°	1.34°	9.23 cm	6.53%
Five	4.88 cm	3.98°	3.98°	1.49 cm	0.74%
Six	1.11 cm	1.36°	1.36°	22.59 cm	15.98%
Seven	3.65 cm	3.43°	3.43°	0.36 cm	0.18%
Eight	0.39 cm	0.55°	0.55°	7.08 cm	5.00%
Floor	1.00 cm	N/A	N/A	N/A	N/A
Ceiling	1.00 cm	N/A	N/A	N/A	N/A
Mean	1.63 cm	1.51°	1.51°	5.31 cm	N/A

Table 7.12: Results for Scenario One: Octagonal Room presenting the four error metrics: difference in position (ΔPosition), dihedral angle, weighted dihedral angle, difference in boundary length (Δlength), and relative error of the inferred boundaries length (δLength).

The shape of the room has been inferred to some extent, however, only four boundaries have a Δposition under 10 cm and the larger dihedral angles, in this case, are a result of the inferred boundaries produced by first-order reflections being angled. This could be due to either the complexity of the room being considered or the requirement for more measurement positions to ensure all first-order reflections are captured, increasing the chance of erroneous boundaries

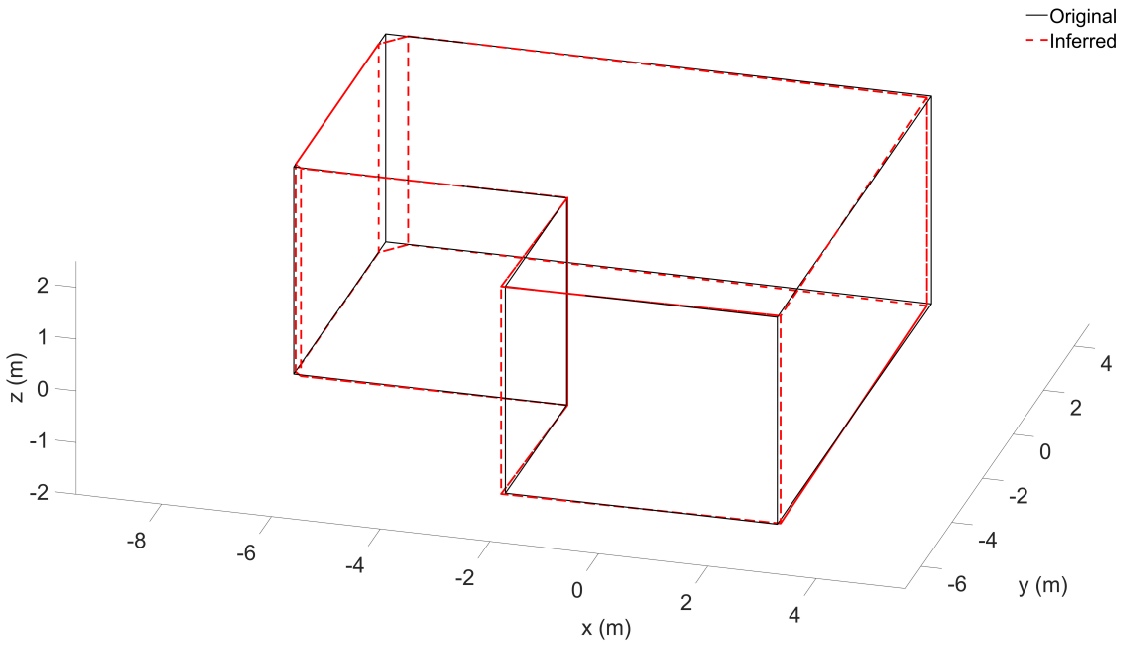

Figure 7.33: Inferred geometry (dashed red line) and desired geometry (solid line) for Scenario One - L-Shaped Room.

Boundary	**ΔPosition**	**Dihedral Angle**	**Weighted Dihedral Angle**	**ΔLength**	**δLength**
One	2.27 cm	23.78°	2.46°	7.54 cm	1.51%
Two	3.80 cm	1.13°	1.13°	5.08 cm	1.27%
Three	3.52 cm	1.12°	1.12°	11.10 cm	2.22%
Four	3.07 cm	0.58°	0.58°	14.95 cm	1.49%
Five	7.62 cm	24.72°	3.51°	15.41 cm	1.54%
Six	0.00 cm	0.00°	0.00°	58.00 cm	9.67%
Floor	1.00 cm	N/A	N/A	N/A	N/A
Ceiling	0.00 cm	N/A	N/A	N/A	N/A
Mean	2.66 cm	8.56°	1.47°	18.68 cm	N/A

Table 7.13: Results for Scenario One: L-Shaped Room presenting the four error metrics: difference in position (ΔPosition), dihedral angle, weighted dihedral angle, difference in boundary length (Δlength), and relative error of the inferred boundaries length (δLength).

being estimated. Furthermore, it is possible that these angled boundaries could be as a result of interactions between near-simultaneously arriving reflections, resulting in a less accurate estimate of DoA for these reflections, as seen for the simultaneously arriving reflections from boundaries two and four – where a DoA estimation error of 8° for each reflection was observed. This observation agrees with similar findings when analysing the ground-truth SRIRs, where typically larger errors in DoA estimation are observed when analysing simultaneously arriving reflections.

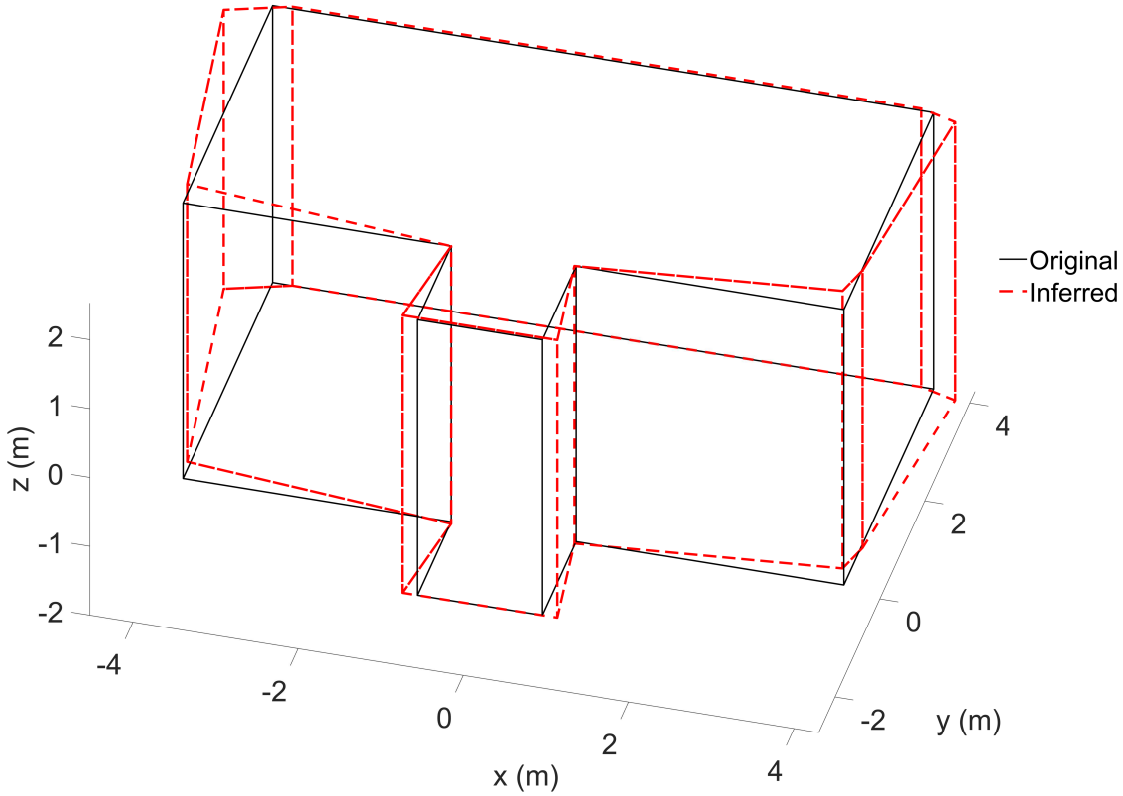

Figure 7.34: Inferred geometry (dashed red line) and desired geometry (solid line) for Scenario One - T-Shaped Room.

Boundary	ΔPosition	Dihedral Angle	Weighted Dihedral Angle	ΔLength	δLength
One	17.97 cm	6.40°	6.40°	7.07 cm	2.17%
Two	10.56 cm	7.41°	7.41°	2.77 cm	1.85%
Three	0.00 cm	0.00°	0.00°	36 cm	24.00%
Four	10.56 cm	7.41°	7.41°	2.77 cm	1.85%
Five	17.07 cm	6.32°	6.32°	7.07 cm	2.17%
Six	16.75 cm	10.05°	6.45°	43.16 cm	10.79%
Seven	1.26 cm	12.66°	2.76°	89.09 cm	11.14%
Eight	31.02 cm	8.04°	8.04°	56.63 cm	14.16%
Floor	0.00 cm	N/A	N/A	N/A	N/A
Ceiling	3.00 cm	N/A	N/A	N/A	N/A
Mean	10.81 cm	7.29°	5.60°	30.57 cm	N/A

Table 7.14: Results for Scenario One: T-Shaped Room presenting the four error metrics: difference in position (ΔPosition), dihedral angle, weighted dihedral angle, difference in boundary length (Δlength), and relative error of the inferred boundaries length (δLength).

For the final two experiments for the third cuboid-shaped room, it is important to note that the recessed windows will not be detected, as they are only recessed by 48 cm from the wall. This is less than the minimum source-to-receiver distance, which means that a source and receiver pair cannot be suitably positioned such that a first-order reflection from the boundaries connected

to the windows can be detected. The consequence of this being that the first-order reflection for that boundary could originate from either the wall or the window, and as such the inferred boundary location will depend upon which of these reflections arrive at the receiver. In this test the wall will be considered as being the ground truth for the boundary, as it occupies the largest portion of Boundary One.

In Figure 7.35 and Table 7.15 the results for the first measurement set for the Third Cuboid Room can be seen. The results show that the main boundaries of the room have been correctly estimated with a maximum ΔPosition of 4 cm, and with boundary One being inferred at the wall, as opposed to at the window. As expected the recessed windows have not been individually inferred, and this result shows that for this example these additional features have not had a negative impact on the accuracy of the geometry inference method.

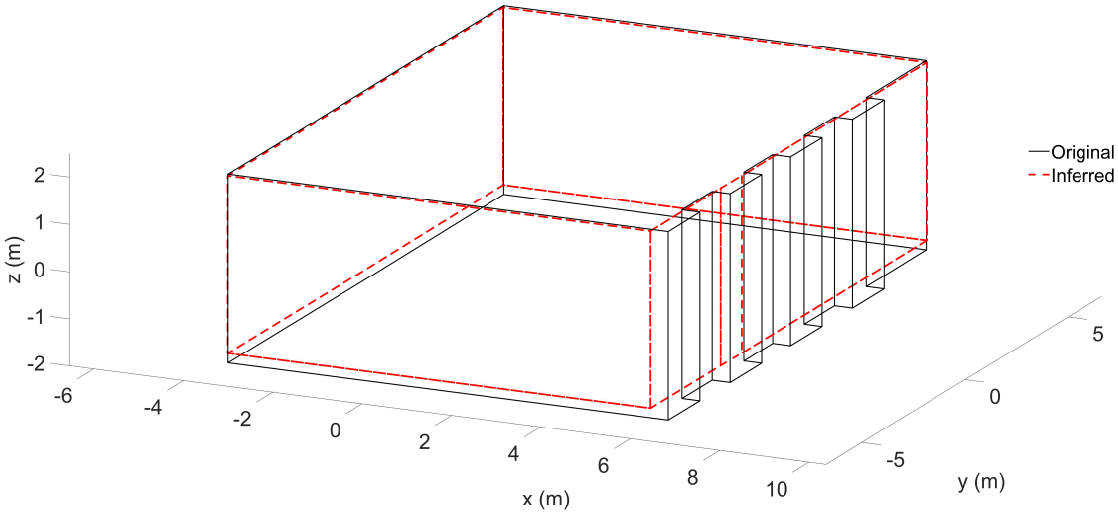

Figure 7.35: Inferred geometry (dashed red line) and desired geometry (solid line) for Scenario One Cuboid Room Three, measurement set one.

Boundary	**ΔPosition**	**Dihedral Angle**	**Weighted Dihedral Angle**	**ΔLength**	**δLength**
One	0.83 cm	0.19°	0.04°	1.00 cm	0.07%
Two	0.00 cm	0°	0°	2.00 cm	0.21%
Three	1.00 cm	0°	0°	1.00 cm	0.07%
Four	1.00 cm	0°	0°	1.00 cm	0.11%
Floor	1.00 cm	N/A	N/A	N/A	N/A
Ceiling	4.00 cm	N/A	N/A	N/A	N/A
Mean	1.31 cm	0.05°	0.01°	1.25 cm	N/A

Table 7.15: Results for Scenario One Cuboid Room Three, measurement set one, presenting the four error metrics: difference in position (ΔPosition), dihedral angle, weighted dihedral angle, difference in boundary length (Δlength), and relative error of the inferred boundaries length (δLength).

The results for measurement set two are presented in Figure 7.36 and Table 7.16, and show an increase in the error values for boundary Three, as a result of an additional angled boundary being inferred at the intersection point between boundaries Three and Four. In this case boundary One has been inferred at the window location which results in a 37 cm ΔPosition error, if the window position is assumed to be the correct boundary location there is a ΔPosition error of 11 cm. With the exception of the angled boundary, it can be seen that the general geometry of the room has been correctly estimated, and errors in the inferred lengths of the planes are as a result of the additional angled boundary and the estimation of Boundary One at the window location.

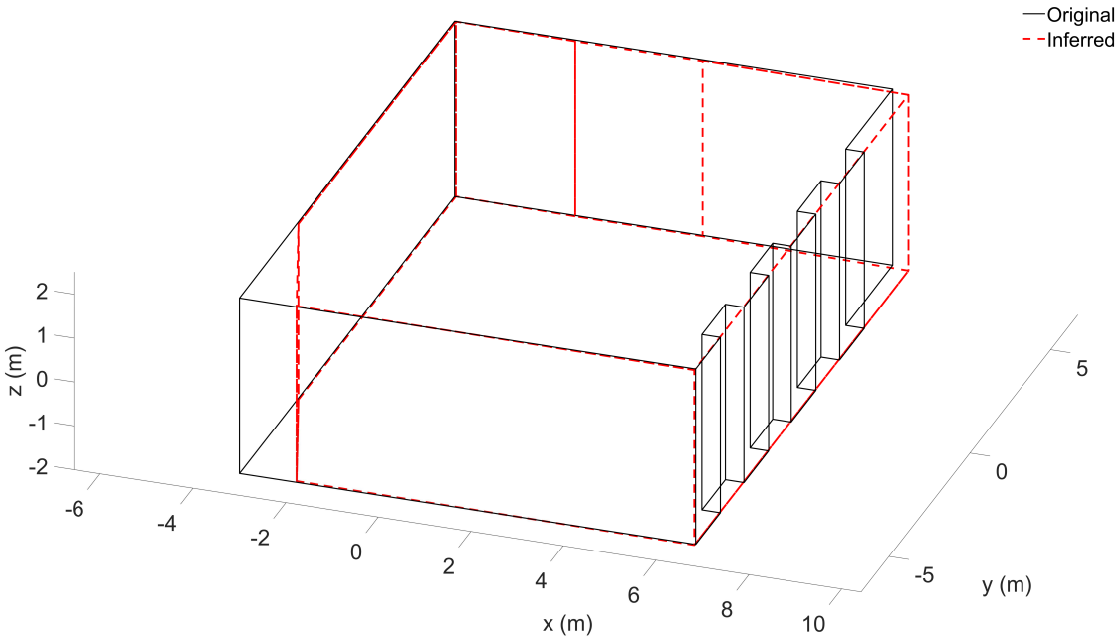

Figure 7.36: Inferred geometry (dashed red line) and desired geometry (solid line) for Scenario One Cuboid Room Three, measurement set two.

Boundary	**ΔPosition**	**Dihedral Angle**	**Weighted Dihedral Angle**	**ΔLength**	**δLength**
One	37.00 cm	0°	0°	9.00 cm	0.67%
Two	3.54 cm	0.75°	0.81°	36.14 cm	3.82%
Three	36.38 cm	9.31°	5.14°	15.52 cm	1.16%
Four	5.31 cm	0.60°	0.60°	82.95 cm	8.78%
Floor	2.00 cm	N/A	N/A	N/A	N/A
Ceiling	3.00 cm	N/A	N/A	N/A	N/A
Mean	14.53 cm	2.67°	1.64°	35.90 cm	N/A

Table 7.16: Results for Scenario One Cuboid Room Three, measurement set two, presenting the four error metrics: difference in position (ΔPosition), dihedral angle, weighted dihedral angle, difference in boundary length (Δlength), and relative error of the inferred boundaries length (δLength).

7.5.3 Test Case Two

As this scenario consists of two sets of 33 different measurement cases, the results are presented as the average boundary estimation errors across the two L-shaped rooms. The boundary positional error data produced has a non-parametric distribution, and as such statistical analysis of this data is performed using the non-parametric Kruskal-Wallis test which in MATLAB is the function *kruskalwallis* [142], and reported as (χ^2 = , p = , degrees of freedom =). Furthermore, the bootstrap process as defined in [162] is used to compute the 95% confidence interval for the mean values using the MATLAB implementation *bootstrapci* [163].

In Table 7.17 the results for the two sets of 33 measurement cases can be seen. There is a 7.41 cm difference between the mean boundary positional error across the two L-shaped rooms, with the second, larger, L-Shaped room having larger boundary positional errors. This increase in boundary position error is as a consequence of 11 cases for the second L-shaped room having an additional angled boundary being inferred, compared to five for the first. The first L-Shaped room has a larger number of angled boundaries inferred in the corners of the room, which results in a larger average dihedral angle and ΔLength. Furthermore, out of the 33 measurements for the first L-shaped room, three cases have additional inferred boundaries located outside of the inferred shape of the room, while only one case was observed for the second L-Shaped room, and as these do not align with an expected boundary location they do not directly impact the estimation errors for each boundary. These results highlight the variability in performance of the method with respect to differences in the SRIR, most likely as a result of overlapping reflections leading to less accurate estimates of DoA as was observed for the T-shaped room in scenario one. The minimum and maximum mean error for the measurement sets in L-shaped Room One are ΔPosition = [3.95 cm, 35.58 cm], Dihedral Angle = [2.24°, 11.22°], Weighted Dihedral Angle = [0.79°, 5.64°], and ΔLength = [6.48 cm, 110.98 cm], and for the second L-Shaped Room, ΔPosition = [4.22 cm, 32.81 cm], Dihedral Angle = [1.05°, 10.30°], Weighted Dihedral Angle = [1.06°, 4.21°], and ΔLength = [8.40 cm, 85.95 cm]. Comparing the variance in measurement accuracy between the two L-shaped rooms it can be seen that there is no significant difference for the ΔPosition (χ^2 = 0.0005, p = 0.98, degrees of freedom = 395), weighted dihedral angle (χ^2 = 2.59, p = 0.11, degrees of freedom = 395), and ΔLength (χ^2 = 0.35, p = 0.55, degrees of freedom = 395). However, there is a significant difference for the averaged dihedral angle (χ^2 = 10.25, p = 0.0014, degrees of freedom = 395), as a result of the additional boundaries inferred in the corners of the room for L-Shaped Room One. This suggests that while there are differences in the mean values between these two examples, the variability in performance

between the two sets are comparable. The best and worst case for these two L-Shaped room can be seen in Figures 7.37–7.38.

L-Shaped room	ΔPosition	Dihedral Angle	Weighted Dihedral Angle	ΔLength
One	11.52±0.09 cm	7.28°±0.05°	2.53°±0.01°	36.84±0.31 cm
Two	18.98±0.10 cm	3.69°±0.03°	2.59°±0.01°	5.63±0.36 cm

Table 7.17: Results for Scenario Two L-Shaped Rooms One and Two the results are presented as the mean of the four error metrics: difference in position (ΔPosition), dihedral angle, weighted dihedral angle, and difference in boundary length (Δlength).

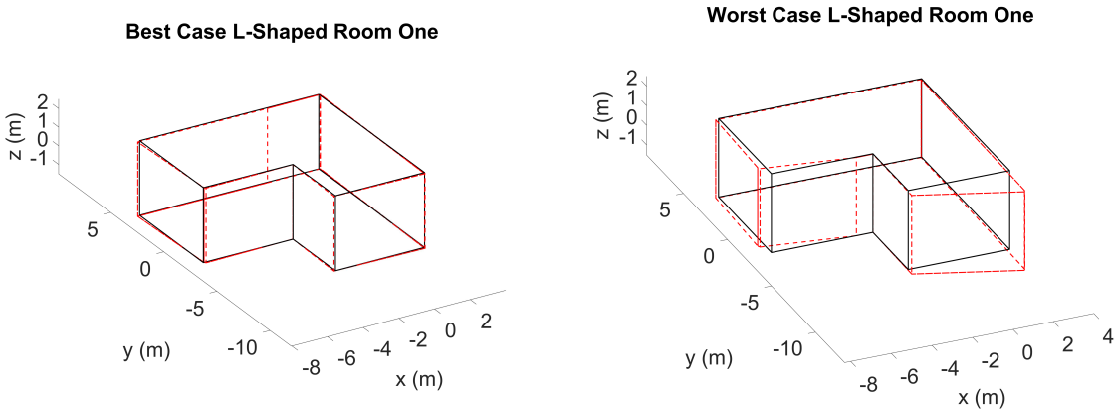

Figure 7.37: The best and worst cases for Scenario Two L-Shaped Room One. Inferred geometry (dashed red line) and desired geometry (solid line).

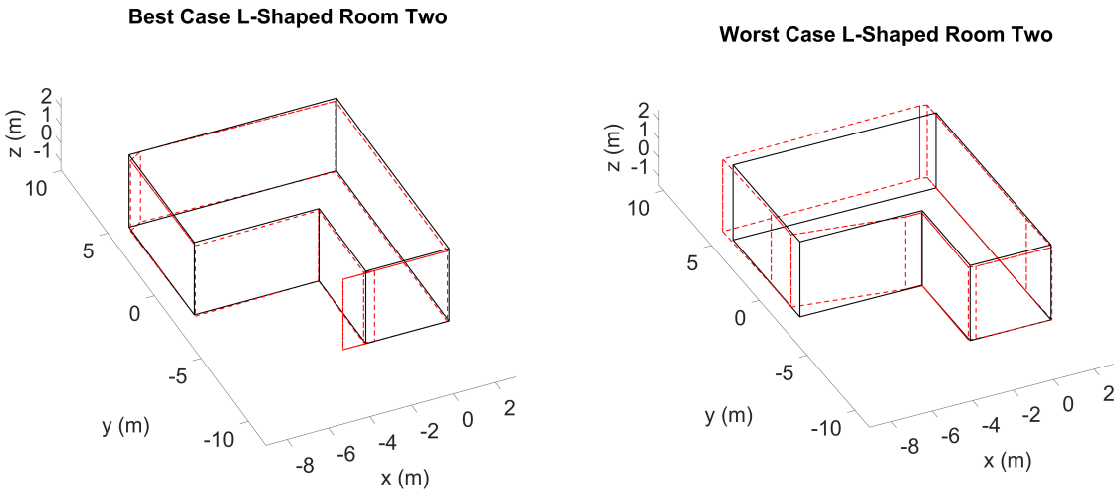

Figure 7.38: The best and worst cases for Scenario Two L-Shaped Room Two. Inferred geometry (dashed red line) and desired geometry (solid line).

7.5.4 Test Case Three

As can be seen in Figure 7.39 and the results in Table 7.18, the general shape of the room has been inferred with small dihedral angle values between the original and inferred boundaries, and only Boundary Two being inferred as two boundaries. However, there is a slight decrease in

accuracy for the boundary position estimation, and therefore the lengths of surrounding boundaries, when compared to the simulated cuboid-shaped rooms from Scenario One. These inaccuracies are likely due to either diffuse reflections, under- or over-estimation of the ToA for reflections in the measured impulse responses, or any inaccuracy in the estimated DoA for the reflections. These lead to incorrect estimation of the desired position for the image-source, which affects both the positioning of the boundary it infers, and any subsequent boundaries that are defined using this image-source. However, while there are larger positional errors for the boundaries, the mean accuracy for this measurement set, 14.5 cm, is comparable to the worst cases in Scenario One, which was 15.1 cm.

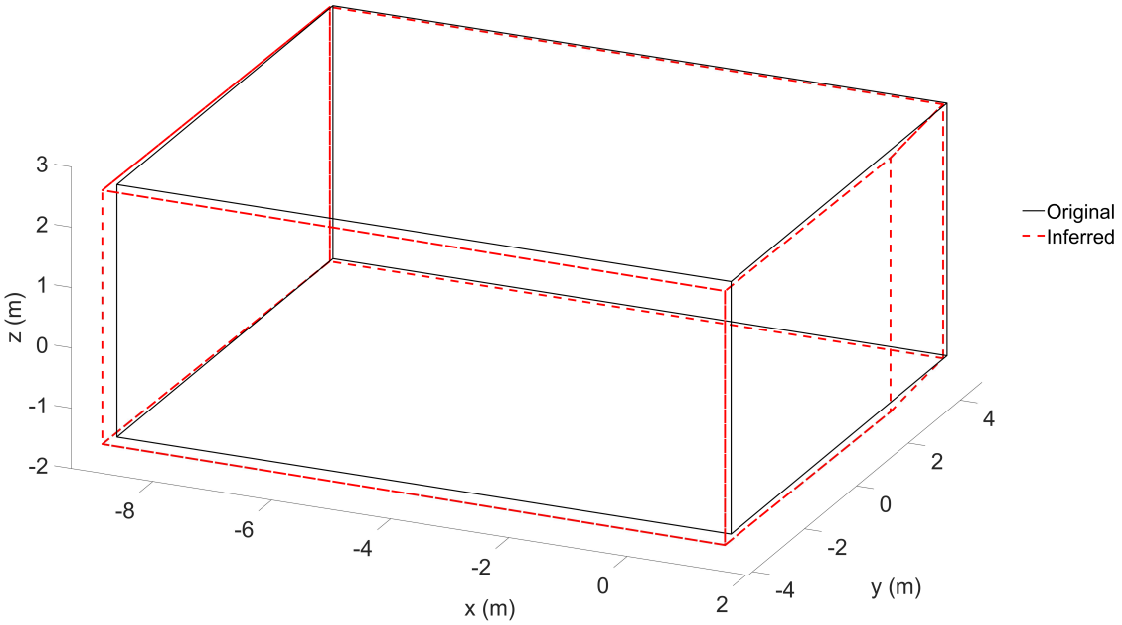

Figure 7.39: Inferred geometry (dashed red line) and desired geometry (solid line) for Scenario Three, measurement set one.

Boundary	ΔPosition	Dihedral Angle	Weighted Dihedral Angle	ΔLength	δLength
One	11.43 cm	0.54°	0.54°	18.65 cm	1.80%
Two	4.02 cm	2.28°	1.52°	34.63 cm	4.16%
Three	18.00 cm	0.00°	0.00°	1.40 cm	0.14%
Four	17.75 cm	0.60°	0.60°	24.05 cm	2.89%
Floor	13.00 cm	N/A	N/A	N/A	N/A
Ceiling	10.60 cm	N/A	N/A	N/A	N/A
Mean	12.47 cm	0.85°	0.66°	19.68 cm	N/A

Table 7.18: Results for Scenario Three, measurement set one, presenting the four error metrics: difference in position (ΔPosition), dihedral angle, weighted dihedral angle, difference in boundary length (Δlength), and relative error of the inferred boundaries length (δLength).

Measurement set two highlights the problems of having non-ideal SRIR measurements. The

first-order reflection for boundary Two in both SRIRs is not detected, as a result of the time-frame they are present in having a diffuseness estimation greater than 30% (the threshold used by the EDESAR method). The first-order reflection from boundary One has a 7° error in estimated elevation, producing a vertically angled plane that exceeds the defined threshold $\epsilon_{\tilde{\mathbf{n}}}$. Furthermore, the first-order reflection from the ceiling, while correctly detected, is not used to define the ceiling location. This is as a result of a false-positive detection within the noise component of the early part of the SRIR (see Figure 7.41), which has been inferred as a first-order reflection from the ceiling, producing a significant underestimation of position of the ceiling by 1.26 m. It is possible that this false-positive detection is as a result of noise produced by the extractor fans present in the room, which were directly above the receiver position in this measurement setup, as similar detections are present throughout the SRIRs.

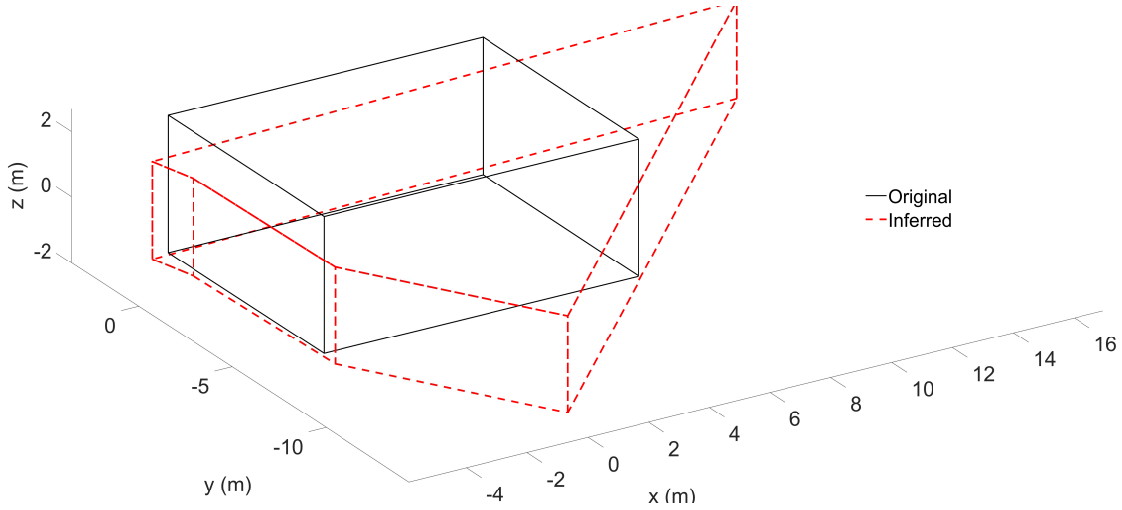

Figure 7.40: Inferred geometry (dashed red line) and desired geometry (solid line) for Scenario Three, measurement set two.

Boundary	**ΔPosition**	**Dihedral Angle**	**Weighted Dihedral Angle**	**ΔLength**	**δLength**
One	430.84 cm	59.32°	59.32°	299.12 cm	28.93%
Two	467.57 cm	43.76°	43.76°	1419.25 cm	170.58%
Three	25.57 cm	2.86°	2.86°	950.47 cm	91.92%
Four	25.57 cm	6.85°	3.20°	67.11 cm	8.07%
Floor	4.00 cm	N/A	N/A	N/A	N/A
Ceiling	1256.00 cm	N/A	N/A	N/A	N/A
Mean	367.76 cm	28.19°	27.28°	683.99 cm	N/A

Table 7.19: Results for Scenario Three, measurement set two, presenting the four error metrics: difference in position (ΔPosition), dihedral angle, weighted dihedral angle, difference in boundary length (Δlength), and relative error of the inferred boundaries length (δLength).

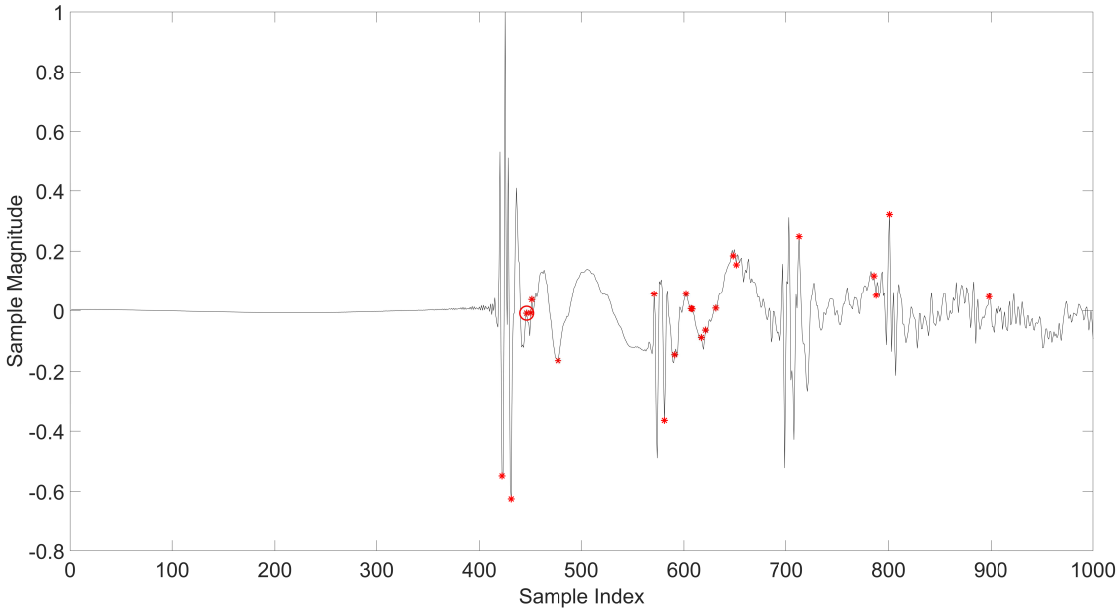

Figure 7.41: Spatial Room Impulse Response One used for Scenario Three, measurement set 2, where the red asterisks denote the detected reflection locations, and the red circle denoted the detection that has resulted in a 1.26 m underestimation of the ceiling position.

7.6 Discussion

Preliminary testing of the proposed method shows that when presented with ground-truth values of time- and direction-of-arrival for 311 reflections across two measurement positions for a cuboid-shaped room, the proposed geometry inference method is capable of producing an exact estimate of the room's shape. This result suggests that errors in estimation accuracy of the proposed method are more likely as a result of inaccuracies in the estimated time- and direction-of-arrival values, missed reflections, and false-positive detections. The results showed that when introducing normally distributed errors into the ToA an increase in the ΔPositional error was observed. Therefore, it would be expected that if the ToA values were biased, the inferred geometry would exhibit bias in boundary position. When considering errors in the DoA estimation process, typically larger errors were observed across all metrics when azimuth DoA errors existed, while elevation errors typically produced larger errors when a $\Delta\phi \geq 5^{\circ}$ for a boundary's first-order reflection across all measurement positions were observed, which resulted in the reflection being ignored by the proposed method. It is important to note that for one case when an elevation error resulted in both first-order reflections for a boundary being rejected, a more accurate estimate of the desired boundary's location was produced as a result of a second-order reflection. However, while this shows that it is possible to infer a boundary without a first-order reflection, the proposed method will more reliably produce an accurate estimate of

a boundary when first-order reflections are present. Finally, the results show that comparable estimates of the geometry were achievable for SNRs greater than or equal to 20 dB. However, for SNRs lower than 20 dB false-positive detections made by the EDESAR method within the noise floor resulted in additional unwanted boundaries being inferred. These results detail the importance of having accurate data and show the potential impact of erroneous estimates of ToA, DoA, missed/rejected first-order reflections, and false positives. However, it is important to note that in all cases the severity of these observed errors will depend on how the resulting inferred boundaries intersect with their neighbouring boundaries.

These results show that with the exception of Scenario Three Measurement Set Two, all convex rooms have mean boundary positional errors with a comparable level of accuracy to prior work in [6–9], which presented distance errors of between 4–7 cm (image-source reversion) [6], 4.21–9.05 cm (direct localisation) [9], 1.7–22.0 cm (direct localisation) [8], 4.9–24.5 cm (image-source reversion) [8], and 20.46 cm (image-source reversion) [7]. Comparing the accuracy of the proposed method when analysing non-convex room to the performance of the current state-of-the-art methods for convex rooms as presented in [7–9], shows that comparable accuracy is achieved for the L-shaped room in Scenario One and 44 of the cases in Scenario Two. Furthermore, the T-shaped room from Scenario One and mean performance for the L-Shaped rooms from Scenario Two are comparable to [7], and within 6.21 cm of the maximum average error in [8]. This decrease in performance compared to the convex cases presented in this chapter are likely as a result of the increased complexity of the room being inferred and consequently the reflection density of the resulting SRIR. The proposed method achieved a comparable level of accuracy to previous work, with a difference in mean positional error for a room of between 12.35–25.81 cm when comparing the maximum boundary positional error values to those presented in previous studies, using at most three measurement positions (48 spherical harmonic domain channels), compared to the 6–78 (maximum of 1056 channels) used in these previous studies.

From the results in Scenario Two it is evident that the performance of the proposed method varies between different measurement set-ups within the same room. These differences are mainly as a result of higher-order reflections that have not been assigned to their corresponding already inferred boundaries, resulting in angled boundaries, generally in the corners of the room, that impact the inferred shape. This suggests that future work on the proposed method should focus on finding a more robust means of retracing reflection paths through existing inferred boundaries, ways of validating inferred boundaries, or ways of invalidating them. To validate a bound-

ary additional constraints could be imposed, such as requiring a higher-order reflection to be assignable to each boundary as in [109]. However, in the case of non-convex rooms this would require the use of multiple receiver positions for each source to ensure that these higher-order reflections are detected. Conversely, different approaches to invalidating inferred boundary, other than the geometry validation process suggested in Section 7.3.2, could be considered, such as image-processing methods used for image analyses and segmentation, or manual deletion of any remaining incorrect boundaries as they are generally obvious even without *a priori* knowledge of the room's shape.

7.7 Conclusions

The proposed method for geometry inference removes the need for using between 6–78 measurement locations (up-to 1056 channels of audio) and the assumption of convexity for the measurement environment as made in previous studies. The proposed method is therefore more widely applicable in practice, where rooms come in different shapes, sizes, complexity, and it is often impracticable to use such large numbers of measurement positions. This is achieved by exploiting spatiotemporal information contained within SRIRs measured using a spherical microphone array, to define the location of image-sources that are used to infer the location of reflective boundaries. A geometry validation process is then performed to refine the number of inferred boundaries to ideally only those that define the original enclosed space.

Preliminary results showed that when presented with exact data for 311 reflections across two measurement positions, the proposed method produces an exact estimate of the room's shape, suggesting that estimation inaccuracies are more likely as a result of inaccuracies in the estimated time- and direction-of-arrival. Additionally, these results showed that errors in the estimated geometry were as a result of inaccurate estimates of ToA, DoA, missed/rejected first-order reflections, or false-positive detections. This shows that the accuracy of the method does depend on how accurately the reflection data is extracted from the SRIRs, however, the severity of these errors also depends upon how the resulting inferred boundaries intersect. The proposed method was then tested across three scenarios using both simulated and measured data. Scenario One tested the method's performance across different room shapes with randomly defined source and receiver positions which satisfied the requirement that a first-order refection should be assignable to each boundary and detectable in at least one SRIR measurement. This showed that the proposed method was able to infer the geometry for rooms of different shapes, sizes, and complexity. The second scenario compared the variability in performance between two sets

of 33 source/receiver combinations measured in two different L-shaped rooms. This evidences the extent to which the method varied dependant on source/receiver position and different sized rooms of the same shape. The final scenario considered two real-world measurements, one of which did not contain all of the required first-order reflections. This showed the impact that real-world conditions, such as noise and acoustic phenomena that cannot be modelled using geometric acoustic modelling methods as used in Scenarios One and Two, had on the geometry inference accuracy. Furthermore, this scenario showed the consequence of not having a first-order reflection attributable to each boundary. The results showed that, with the exception of measurement set two in Scenario Three, all convex-shaped rooms were estimated with accuracy comparable to the methods presented in [6–9], which reported average distance errors of 4–7 cm (image-source reversion) [6], 4.21–9.05 cm (direct localisation) [9], 1.7–22.0 cm (direct localisation) [8], 4.9–24.5 cm (image-source reversion) [8], and 20.46 cm [7] (image-source reversion). For the case of the non-convex shaped rooms the results were comparable to the image-source reversion techniques presented in [7, 8], with a difference in mean positional error for a room of between 12.35–25.81 cm when comparing the maximum boundary positional error values to those presented in previous studies. This shows that the proposed method is comparable to the performance of the current-state-of-the-art methods, using, in these scenarios, at most three measurement positions (48 spherical harmonic domain channels) compared to the 6–78 measurement positions (maximum of 1056 channels) used in these previous studies. The results in Scenario Two highlighted how varied the performance of this method can be, with a 31.63 cm difference between the best and worst performing measurement set. This result highlights the areas future work should be focused, considering ways of optimising the retracing of reflections through existing boundaries to better deal with higher-order reflections, approaches to validating inferred boundaries, or means of invalidating incorrectly assigned ones.

Chapter 8

Conclusions and Future Work

8.1 Thesis Summary

This thesis has presented novel approaches to direction-of-arrival estimation for reflections in binaural room impulse responses, a spatiotemporal decomposition based method for reflection detection, and a geometry inference method for both convex- and non-convex-shaped rooms. In this section, a summary of this thesis and its contributions are presented.

In Chapter 2, the fundamentals of sound propagation, acoustic reflection, room impulse responses, the image-source method, neural networks, beamformers, and spherical harmonics were introduced. Room impulse responses are the characteristic response of a room to an excitation from an impulse-like broadband signal. These room impulse responses consist of a superposition of the direct source-to-receiver sound and reflections, produced by interactions with the boundaries and surfaces present in the space. Therefore, these room impulse responses convey information about the acoustics of a given room and the locations of any reflective boundaries and surfaces within the room. The theory introduced in this chapter underpinned the concepts of work presented in subsequent chapters of this thesis.

Chapter 3 introduced the concept of time- and direction-of-arrival estimation for reflections present in a room impulse response, and presented and reviewed current state-of-the-art methods. These reflection analysis stages are generally a prerequisite for geometry inference, as the information extracted for each reflection is directly relatable to the boundaries present in the measurement environment. Various approaches for estimating the time-of-arrival for reflections

in impulse responses have been presented across the literature, and generally these methods are capable of producing accurate estimates of time-of-arrival for discrete early reflections, but degrade in performance as reflection density, and consequently, number of overlapping reflections, increases. Furthermore, as these methods are founded on a temporal decomposition of a room impulse response, they cannot disambiguate between reflections arriving simultaneously from different directions, and will detect them as a single arrival. It is therefore intuitive to consider the spatiotemporal decomposition of spatial room impulse responses to enable detection of overlapping and simultaneously arriving reflections. This concept forms the basis by which the novel reflection detection and analysis approach was developed in Chapter 6.

In Chapter 4, the concept of geometry inference was introduced, and a review of the current state-of-the-art methods presented. Geometry inference focuses on the problem of estimating the locations of reflective boundaries within an environment from reflections captured across a number RIRs. There are two main approaches to this in the literature, image-source reversion and direct localisation. Image-source reversion techniques exploit the properties of the image-source method to infer the locations of boundaries from a set of candidate image-sources defined by the time-of-arrival, or time- and direction-of-arrival, of reflections. Direct localisation techniques use some mathematical approach to directly relate the time-of-arrival to a boundary, using ellipses with axes defined by the time-of-arrival. Boundaries are then inferred by finding a line that is tangential to a set of ellipses defined by the same reflection across multiple receiver positions. Current state-of-the-art methods use between 6–78 (a maximum of 1056 channels) measurement positions spaced throughout a given environment, and require a first-order reflection for each boundary attributable to and detectable in all, or a subset, of these measurement positions. To this extent, these methods are only accurate when an assumption of convexity is valid, where this requirement of a first-order reflection from each boundary being detectable across measurement positions is easily met. It was therefore proposed that by using a compact microphone array capable of capturing both the time- and direction-of-arrival, boundary locations, and consequently the room's shape, can be inferred for both convex- and non-convex-shaped rooms, using only a sufficient number of measurement positions to ensure each boundary has a first-order reflection is detectable, in at least one measurement.

In Chapter 5, a method for direction-of-arrival estimation of reflections in binaural room impulse responses was presented using current state-of-the-art methodology based on binaural model fronted neural networks. This chapter aimed to establish whether a two-channel microphone array, capable of capturing three-dimensional spatial information, can produce direction-of-arrival

estimates that are accurate enough for use in geometry inference. The proposed method uses a binaural model to compute the interaural cross-correlation and interaural level difference between the signals measured at the left and right ear. These interaural cues were used to create a feature space for a segment of a binaural room impulse response that contains either the direct sound or a reflection. To disambiguate between front and rear hemispheres, where similarities in interaural cues exist, a second set of interaural cues, calculated from the same time-frame in a second binaural room impulse response measured with the head having been rotated $\pm 90^{\circ}$, was used. The combined feature spaces for these two measurements are analysed by a cascade-forward neural network and an estimate of the direction-of-arrival is produced. The results presented in this chapter showed that the proposed method performed comparably for binaural room impulse responses measured with two different binaural dummy heads and two different loudspeakers. However, there was a large reduction in accuracy when analysing reflections, with in the best case only 40.97% of the estimates being within $\pm 5^{\circ}$ of the expected direction-of-arrival, compared to 81.25% for the direct sound. This reduction in accuracy is likely as a result of multiple points of reflection on the boundary producing multiple, closely arriving, reflections at each ear, resulting in a blurring of the interaural cues due to these interfering signals, or as a result of lower signal-to-noise ratios observed for the reflections. Furthermore, even when accounting for measurement bias the direction-of-arrival for reflections was less accurately estimated, suggesting that the errors observed are more than just a product of system misalignments. It was suggested that this approach would not yield accurate enough estimates of direction-of-arrival for geometry inference, particularly as the results suggest that performance would likely further degrade in the presence of overlapping reflections, and so an alternative microphone array was proposed for the remainder of this thesis.

In Chapter 6, a spatiotemporal decomposition based reflection detection method applicable to spherical microphone arrays was proposed. This method performed spherical harmonic domain beamforming on short time-frames from a spatial room impulse response, generating a heat map of signal intensity over a grid of directions-of-arrival. Reflections are then detected by searching for regions of high-intensity within this heat map. To define the temporal region of the reflection, subsequent time-frames are then analysed to find the time-frame when the reflection is no longer present. Beamforming is once again performed on this temporal region, steered in the direction of the arriving signal, to detect the time-of-arrival for the reflections. Results presented in this chapter compared the accuracy of this method to implementations of two state-of-the-art reflection detection methods, the circular-variance local maxima, and dynamic time warping based matching pursuit methods. The results showed that the proposed method generally produced

more accurate estimates of time-of-arrival, with a minimum difference of 1.99 μs and a maximum of 33.52 μs. The main benefit of this approach is that simultaneously arriving reflections can be detected as individual discrete reflections, where existing methods would detect this as one arrival. The results, however, also showed that when analysing real-world spatial room impulse responses, the proposed method occasionally detected the same reflection multiple times. This was as a result of differences in the spatial width of a reflection within the heat map across multiple time-frames. While this presents a problem for general use, as these multiple detections have the same direction-of-arrival they will be assigned to the same boundary, and as such is of little consequence for geometry inference.

Chapter 7 presented an image-source reversion method, which was tested with both convex- and non-convex-shaped rooms. This method uses the time- and direction-of-arrival for reflections estimated using the method presented in Chapter 6 to compute the location of the image-sources that produce each reflection when specularity is assumed. From these image-source locations a set of candidate boundaries are produced by searching for the most-likely previous-sources that can define the reflection path. To validate candidate boundary locations, and remove incorrectly inferred boundaries, a three step geometric acoustic validation process was proposed. Preliminary testing showed that, when presented with ground-truth values of ToA and DoA an exact estimate of the room's geometry was achieved, and that errors in the estimated geometry were typically a result of inaccurate estimates of ToA and DoA, missed/rejected reflections, and false-positive detections within the noise floor. Furthermore, the severity of the estimation error depends upon how the resulting inferred boundaries intersect with their neighbouring boundaries. The results presented showed that the proposed method performed comparably to state-of-the-art methods when analysing cuboid-shaped rooms. The proposed method also performed comparably to these existing methods when analysing convex-shaped rooms that are not cuboidal and non-convex-shaped rooms. A difference in mean positional error, of the boundaries in a room, of between 12.35–25.81 cm when comparing the maximum boundary positional error values to those presented in previous studies, using at most three measurement positions (48 spherical harmonic domain channels), compared to the 6–78 (maximum of 1056 channels) used in these previous studies.

8.2 Restatement of Hypothesis

The hypothesis for this thesis, as introduced in Chapter 1 is as follows:

Given a compact microphone array and a sufficient number of spatial room impulse responses

to ensure a first-order reflection is detectable for each boundary, accurate boundary estimation, and consequently room shape estimation, can be achieved for both convex- and non-convex-shaped rooms.

The results presented in Chapter 7 clearly support this hypothesis, with a 2.94 cm difference in mean positional error between the best convex and non-convex case. However, the results showed that the accuracy of the method varied between measurement positions within the same room, mainly as a result of errors in time- and direction-of-arrival estimation, which resulted in angled boundaries being generated. While the variability was shown to be statistically similar between two differently sized rooms of the same shape, the results indicated that further development of the overall process is still required.

Comparing the results to those presented for current state-of-the-art methods shows that the proposed method for non-convex rooms performs comparably to convex cases presented in the literature, with for the best cases 2.61 cm difference in the mean boundary positional error for a room, and a difference between 12.35–25.81 cm for the worst cases. Furthermore, this comparable level of accuracy was achieved using at most three measurement positions (48 spherical harmonic domain channels), compared to the 6–78 (maximum of 1056 channels) used in these previous studies.

While the work in this thesis supports the hypothesis, further work is still required to reduce the variability in performance between different measurement positions within the same room. These improvements will produce more robust geometry inference methods, that are applicable in numerous fields, such as, speech recognition, sound source separation, dereverberation, audio forensics, and simultaneous localisation and mapping problems. In order to achieve this, the overall process needs to be refined and further validated using additional real-world non-convex environments. Based on these findings, areas of future development are outlined in what follows.

8.3 Novel Contributions

In addressing the hypothesis the following novel contributions to the field have been identified:

Application of a binaural model fronted neural network for direction-of-arrival estimation of reflections in binaural room impulse responses

The binaural model fronted neural network direction-of-arrival estimator presented in Chapter 5, is novel in its application, and, to the author's knowledge, is the first study to consider neural

networks for the estimation of direction-of-arrival for short time-frame of binaural room impulse responses, such as the reflections in binaural room impulse responses. Furthermore, to the author's knowledge, this is only the second ever study to consider the problem of direction-of-arrival estimation of reflections in binaural room impulse responses. The results showed the potential of using such a method, and through analysis of these results future research areas have been suggested.

A novel spatiotemporal decomposition reflection detection method, capable of detecting simultaneously arriving reflections, from different directions, as individual discrete events
The spatiotemporal decomposition method presented in Chapter 6 is a novel approach to reflection detection. The tests presented showed that the proposed method produces more accurate estimates of time-of-arrival for reflections compared to two state-of-the-art methods, the circular-variance local maxima, and dynamic time warping based matching pursuit methods. Furthermore, to the authors knowledge, the proposed method is the first to consider the detection of simultaneously arriving reflections.

A novel boundary estimation, room shape inference, and boundary validation method, applicable to both convex- and non-convex shaped rooms
The novel geometry inference method presented in Chapter 7, is, to the author's knowledge, the first geometry inference method to consider non-cuboid-shaped rooms, and in particular non-convex-shaped rooms. This is validated by presenting tests across rooms of various shapes, size, and complexity, showing comparable performance to existing methods that only consider cuboid-shaped rooms. This method also presented a novel room shape inference and boundary validation process, addressing the problem of incorrectly inferred additional boundaries, which have been commonly ignored in previous studies. Furthermore, a study of performance variability over different source and receiver positions is presented.

Model Validation
Objective analysis of the results are presented, to validate the model across different shaped rooms. This expands on the cases that geometry inference is applicable to, and suggests areas of further research in the field.

8.4 Future Work

Binaural Direction-of-Arrival Estimation for Reflections in Binaural Room Impulse Responses

From the results presented in Chapter 5, it is evident that further work is required to improve the robustness of this method when analysing reflections. The results suggest that the reduction in accuracy for the direction-of-arrival estimation of reflections could be due to multiple reflection points on the boundary producing closely arriving reflections in the binaural room impulse response which blur the interaural cues. Therefore, given that the direct sound's direction-of-arrival is estimated accurately, it is possible that the accuracy of this model could be improved further by training the neural network with a dataset of reflections in addition to the head related impulse responses. Furthermore, it is likely that the performance of this method will degrade as a result of overlapping reflections. Therefore, future work could also look to expand the multi-conditional training set to also include cases with overlapping reflections, as opposed to just varying signal-to-noise ratios. Future work should also expand on the model to include estimation of elevation direction-of-arrival as well.

Spatiotemporal Decomposition Reflection Detection Methods

The key issues presented in Chapter 6 and 7 for the proposed spatiotemporal decomposition reflection detection method were, the identification and detection of interfering noise as reflection information, and detecting the same reflection multiple times in real-world measurement cases. Future work in this area could consider the use of subspace based beamformers, which decompose the recorded signal into a desired signal and noise subspace, to remove the noise component of the spatial room impulse response. Furthermore, the process by which the same reflection is detected across adjacent time-frames needs to be refined to account for changes in the spatial width of the reflection. This could be achieved by comparing the estimated direction-of-arrival between detections across time-frames, as opposed to the spatial region occupied by the reflection.

Geometry Inference Methods for Convex- and Non-Convex-Shaped Rooms

From the results presented in Chapter 7, it is evident that the main drawback of the proposed geometry inference is the variability in performance between measurement positions. The main cause of this variability is due to additional angled boundaries being inferred in addition to the actual boundary position. This is as a result of an incorrectly assigned previous-source for a higher-order reflection from a given boundary. There are several approaches that could be explored to solve this problem. Firstly, as these angled boundaries are usually adjacent to the correctly inferred boundary, these incorrect boundaries could be manually removed. A computational solution could consider different means of retracing reflections through the environment. This could be achieved by using the image-source method to backtrace an image-source through

already inferred boundaries that intersect the path from image-source-to-receiver, repeating the process until the reflection is retraced back to the source location or a maximum reflection order is reached - at which point a new boundary is defined. This can allow reflection paths to be traced for cases when the correct previous-source belongs to a reflection that was not detected within the spatial room impulse response, and could also improve robustness to time- and direction-of-arrival estimation errors. Alternatively, image-processing techniques could potentially be explored to define a way of invalidating outlier inferred boundary locations.

Additional Real-World Validation

The geometry inference method presented in this thesis has predominantly been validated using CATT-Acoustic simulated spatial room impulse responses, with only one real-world example presented. Future work should further validate the method in real-world scenarios, particularly for measurements taken in more complex convex- and non-convex-shaped rooms. In addition to this, to the author's knowledge, no existing methods for geometry inference have considered the implications that additional reflective surfaces of finite length (chairs, tables, etc.) have on the ability to infer the shape of the room from a set of candidate boundaries. Therefore, future studies should explore the implications on geometry inference of such surfaces, as they can potentially produce additional candidate inferred boundaries, which would increase the complexity of the room shape inference process. Furthermore, testing with different compact microphone arrays could be considered, as the methods presented in this thesis can be applied to other compact arrays for which beamforming is applicable.

8.5 Closing Remarks

This thesis has presented work which expands the applications of geometry inference methods to consider convex- and non-convex-shaped rooms. Geometry inference has potential applications in various aspects of acoustics and signal processing research, where normally *a priori* knowledge of a room's boundary locations would be required, which is not possible when implemented within consumer technology. In acoustics consultancy, geometry inference can be used as a means of deriving key reflection in a given environment, providing data that can be used when acoustically treating the room. The geometric model of a room can be used to simulate the acoustic conditions, and consequently the SRIRs, for different source and receiver positions within the environment. Geometry inference in this context can be used to generate a room model, which subsequently can be used to generate additional SRIRs throughout the environment. This has potential applications in interactive media such as video games where

Spatial Room Impulse Responses (SRIRs) can be used to produce a more realistic rendering of an acoustic scene, producing an immersive experience for the player. In smart home-devices, knowledge of the surrounding environment, and therefore geometry inference, can be used as a means of enhancing speech recognition through source separation and dereverberation as seen in [1–3]. Furthermore, geometry inference can be applied to robotics as a means of providing real-time information about a robot's surrounding environment and its current and previous position [4]. Finally, in the context of virtual and augmented reality, geometry inference can be used to track a user's position within an environment or produce more robust methods for spatial audio rendering by evaluating a user's loudspeaker setup and listening environment, which subsequently can be accounted for when rendering a virtual auditory environment [5], so producing an ideally more immersive user experience. From these applications, it is evident that removing *a priori* knowledge of an environment, it is of paramount importance to arrive at a method for geometry inference that is universally robust to rooms of different shape, size, complexity, and measurement conditions.

It is important to note that while a spherical microphone arrays was adopted in this thesis, any compact microphone array, for which beamforming can produce sufficiently accurate estimates of DoA, can be used. However, further work is still required to decrease variance in accuracy as a result of different source and receiver positions within the same room. As such the challenge of producing a universally robust geometry inference method has yet to be met, and further testing in more complex convex and non-convex real-world rooms is still required.

Part IV

Appendices

Appendix A

Spatial Room Impulse Responses

This appendix contains the Spatial Room Impulse Response (SRIR) produced for each example case used in Chapter 7, the red asterisk on the plots indicate the locations of a detected candidate reflection.

Scenario One

In Figure A.1 the SRIR and detected reflections for Scenario One Cuboid Room One can be seen. From the results it can be seen that the main reflections in the SRIRs have been detected, however, there are a few obvious false-positive detections in both SRIRs in areas where the signal magnitude approaches zero.

In Figures A.2 the SRIR and detected reflections for Scenario One Cuboid Room Two can be seen. From the results it can be seen that the main reflections in the SRIRs have been detected, however, as with the first cuboid room, there are a obvious false-positive detections in both SRIRs.

In Figures A.3 the SRIR and detected reflections for Scenario One Octagonal Room can be seen. From the results it can be seen that the main reflections in the SRIRs have been detected, however, as with the two cuboid rooms, there are a obvious false-positive detections in both SRIRs.

In Figures A.4 the SRIR and detected reflections for Scenario One L-Shaped Room can be

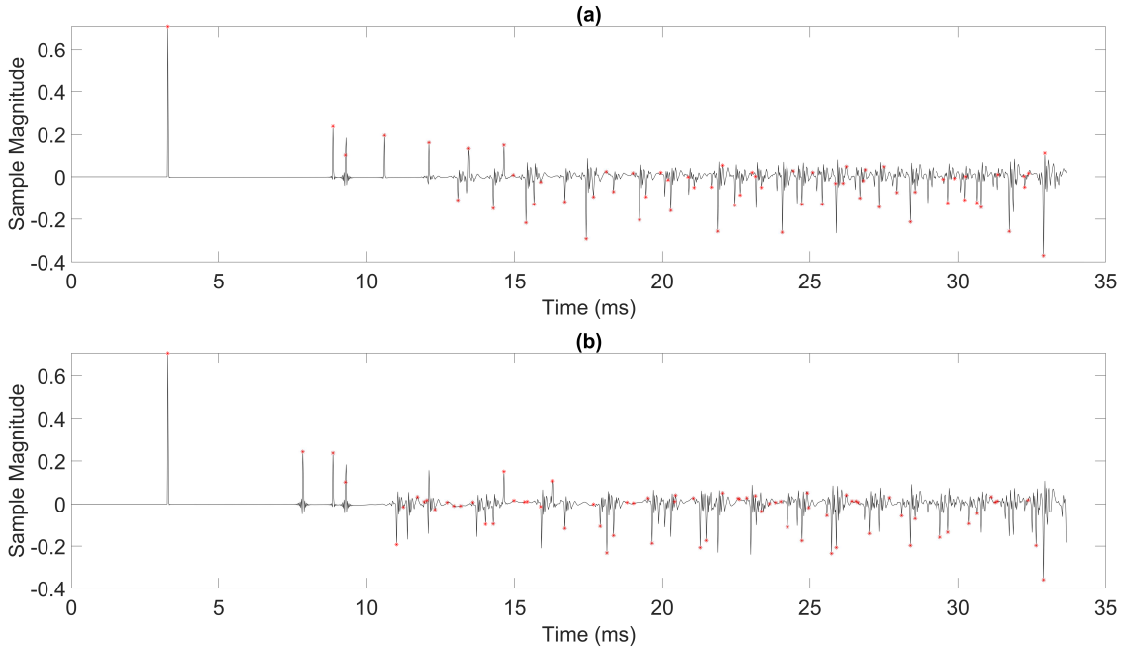

Figure A.1: The omnidirectional channel for the CATT-Acoustics simulated SRIRs for Scenario One Cuboid-Shaped Room One - the red asterisks indicate the locations where the EDESAR method has detected a reflection.

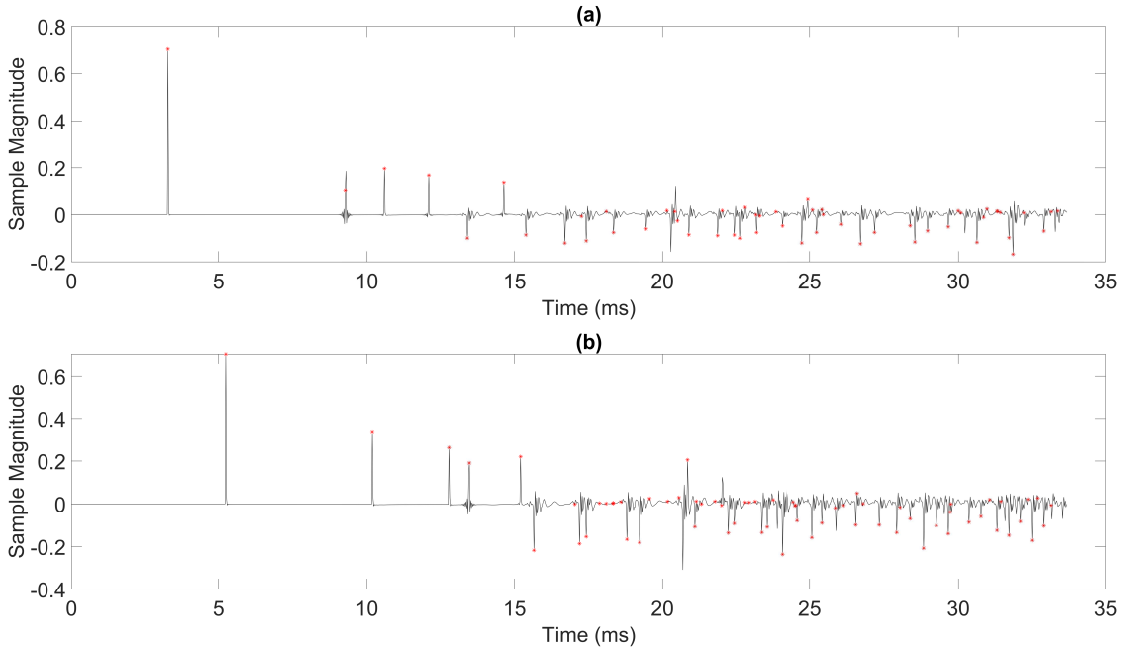

Figure A.2: The omnidirectional channel for the CATT-Acoustics simulated SRIRs for Scenario One Cuboid-Shaped Room Two - the red asterisks indicate the locations where the EDESAR method has detected a reflection.

seen. From the results it can be seen that the main reflections in the SRIRs have been detected, however, as with the previous cases, there are a obvious false-positive detections in both SRIRs, with more present in the first SRIR.

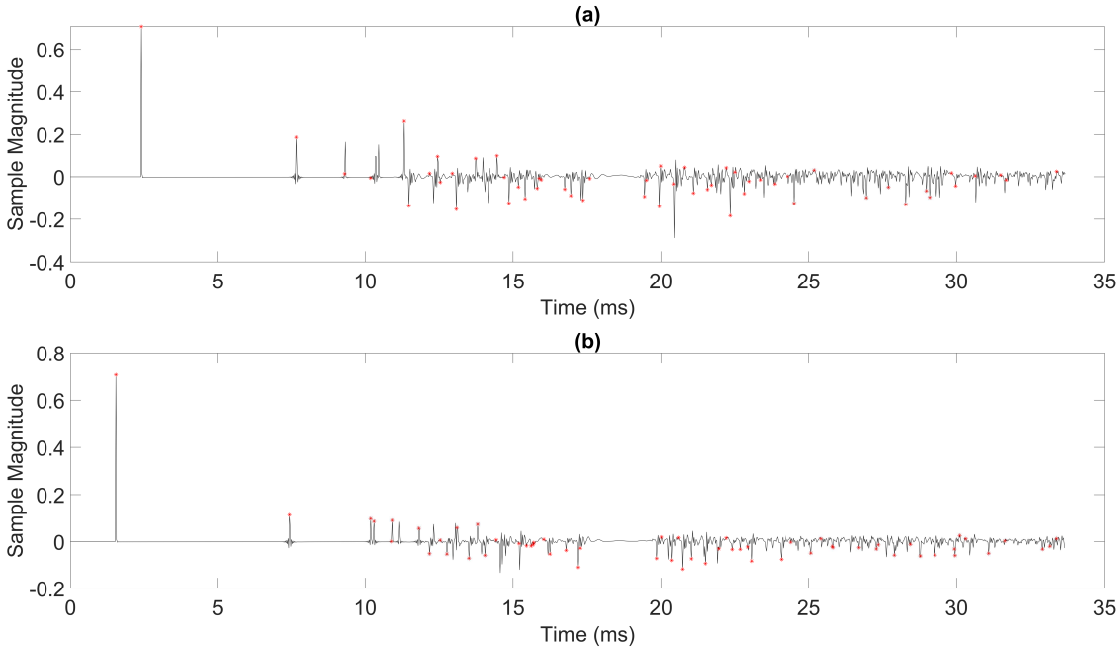

Figure A.3: The omnidirectional channel for the CATT-Acoustics simulated SRIRs for Scenario One Octagonal-Shaped Room - the red asterisks indicate the locations where the EDESAR method has detected a reflection.

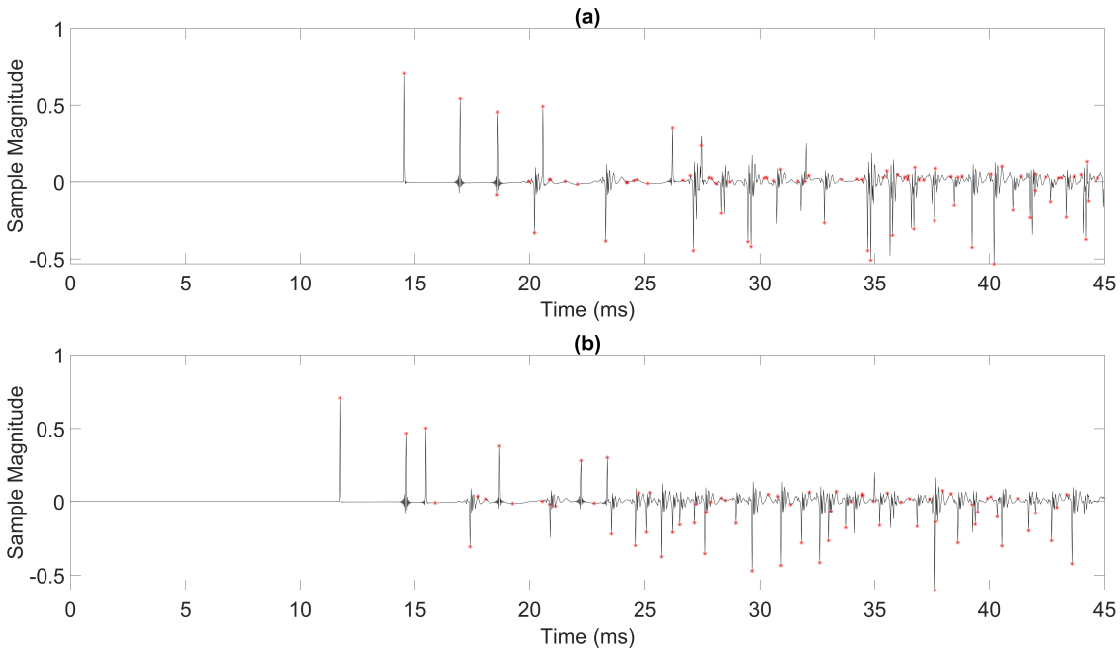

Figure A.4: The omnidirectional channel for the CATT-Acoustics simulated SRIRs for Scenario One L-Shaped Room - the red asterisks indicate the locations where the EDESAR method has detected a reflection.

In Figures A.5 the SRIR and detected reflections for Scenario One T-Shaped Room can be seen. From the results it can be seen that the main reflections in the SRIRs have been detected, however, as with the previous cases, there are a obvious false-positive detections in both SRIRs, with more present in the first and second SRIR. Furthermore, it can be seen that for the third

SRIR, which was positioned in the alcove of the T-Shaped Room, there are fewer more sparsely distributed reflections.

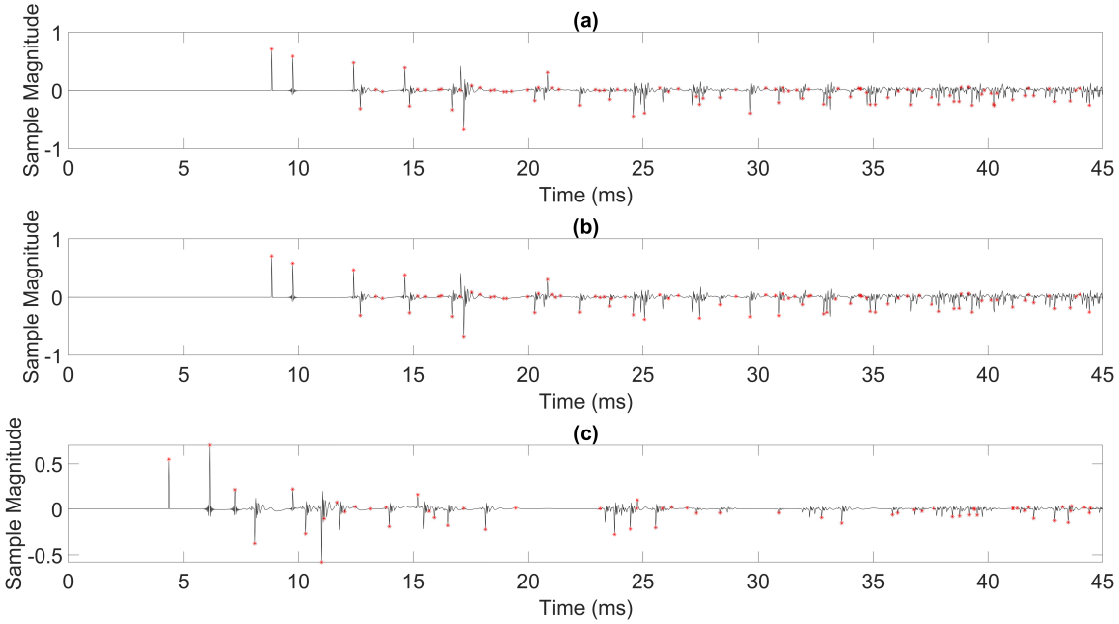

Figure A.5: The omnidirectional channel for the CATT-Acoustics simulated SRIRs for Scenario One T-Shaped Room - the red asterisks indicate the locations where the EDESAR method has detected a reflection.

In Figures A.6 the SRIR and detected reflections for Scenario One Cuboid Room Three, measurement set one, can be seen. From the results it can be seen that the main reflections in the SRIRs have been detected, however, there are fewer obvious false-positive detections in both SRIRs, when compared with the previous cases.

In Figures A.7 the SRIR and detected reflections for Scenario One Cuboid Room Three, measurement set one, can be seen. From the results it can be seen that the main reflections in the SRIRs have been detected, however, there are again fewer obvious false-positive detections in both SRIRs, when compared with the previous cases.

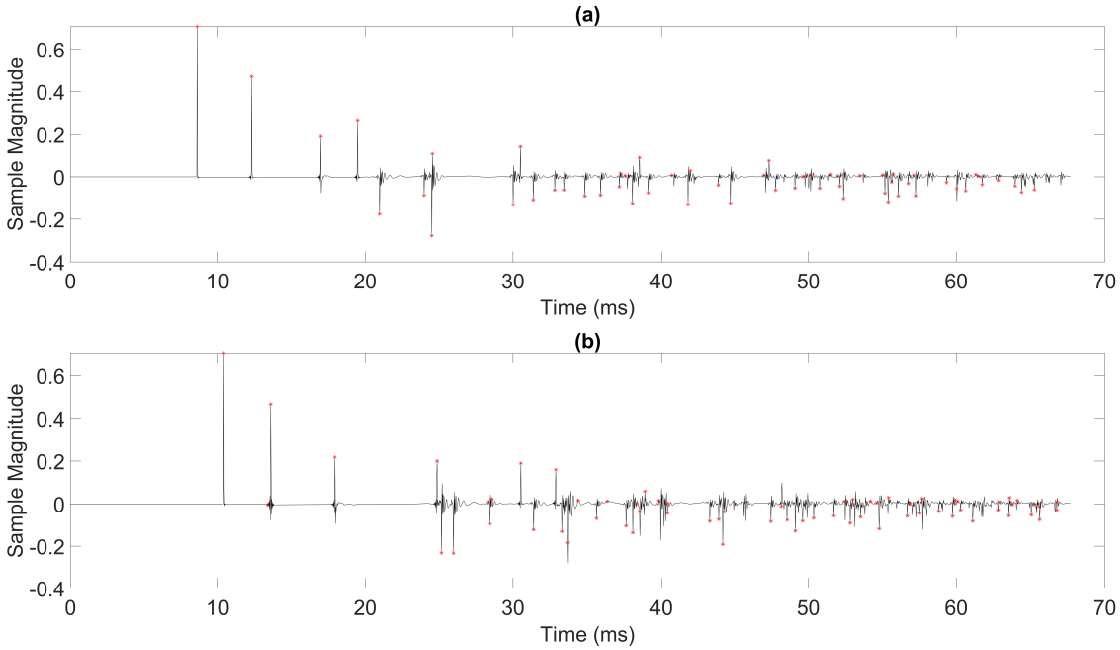

Figure A.6: The omnidirectional channel for the CATT-Acoustics simulated SRIRs for Scenario One Cuboid Room Three Measurement Set One - the red asterisks indicate the locations where the EDESAR method has detected a reflection.

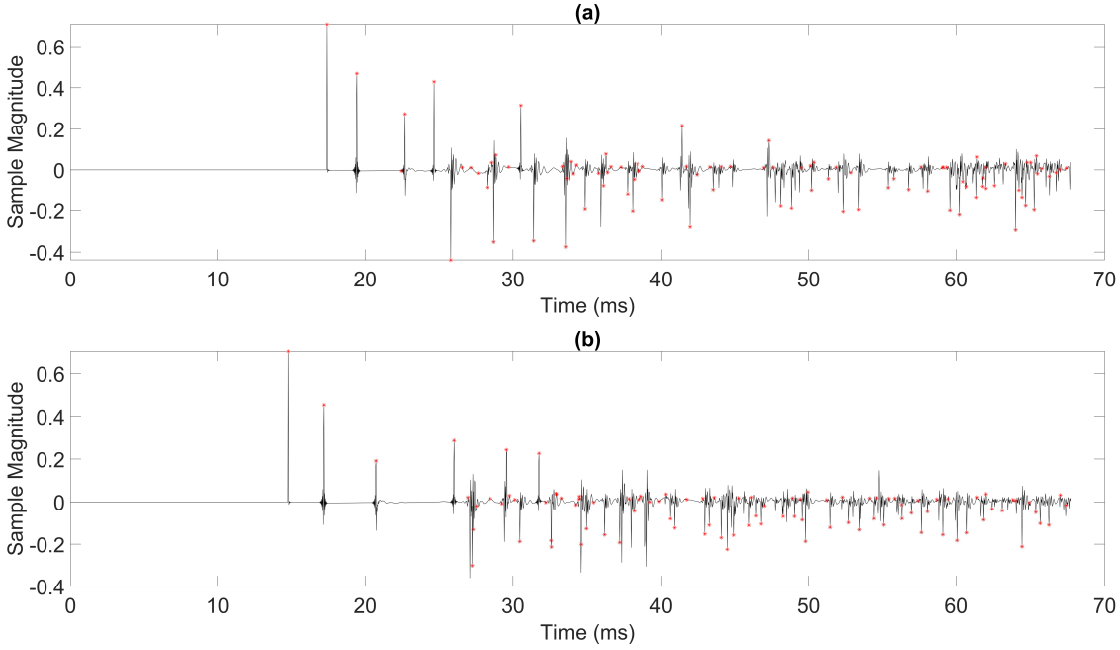

Figure A.7: The omnidirectional channel for the CATT-Acoustics simulated SRIRs for Scenario One Cuboid Room Three Measurement Set Two - the red asterisks indicate the locations where the EDESAR method has detected a reflection.

Scenario Three

In Figure A.8 the SRIR and detected reflections for Scenario Three, measurement set one, can

be seen. These results have been discussed in detail in Chapter 6.

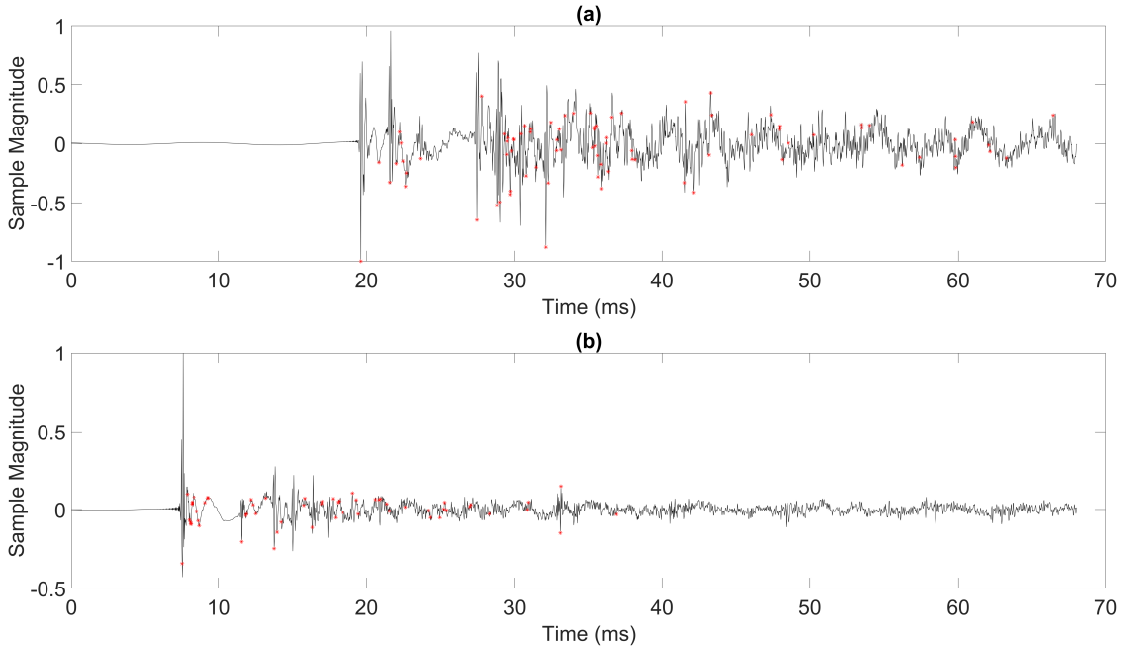

Figure A.8: The omnidirectional channel for the real-world measured SRIRs for Scenario Three, Measurement Set One - the red asterisks indicate the locations where the EDESAR method has detected a reflection. These SRIR were measured using an EigenMike EM32 spherical microphone array, Genelec 8030 loudspeaker, and the exponential sine-sweep method.

In Figure A.9 the SRIR and detected reflections for Scenario Three, measurement set one, can be seen. From the results it is evident that there are fewer correct detections, mainly as a result of a noisier signal. As was discussed in Chapter 6, it can be seen that there have been numerous cases of false-positive detections after the arrival of a reflection at the microphone array.

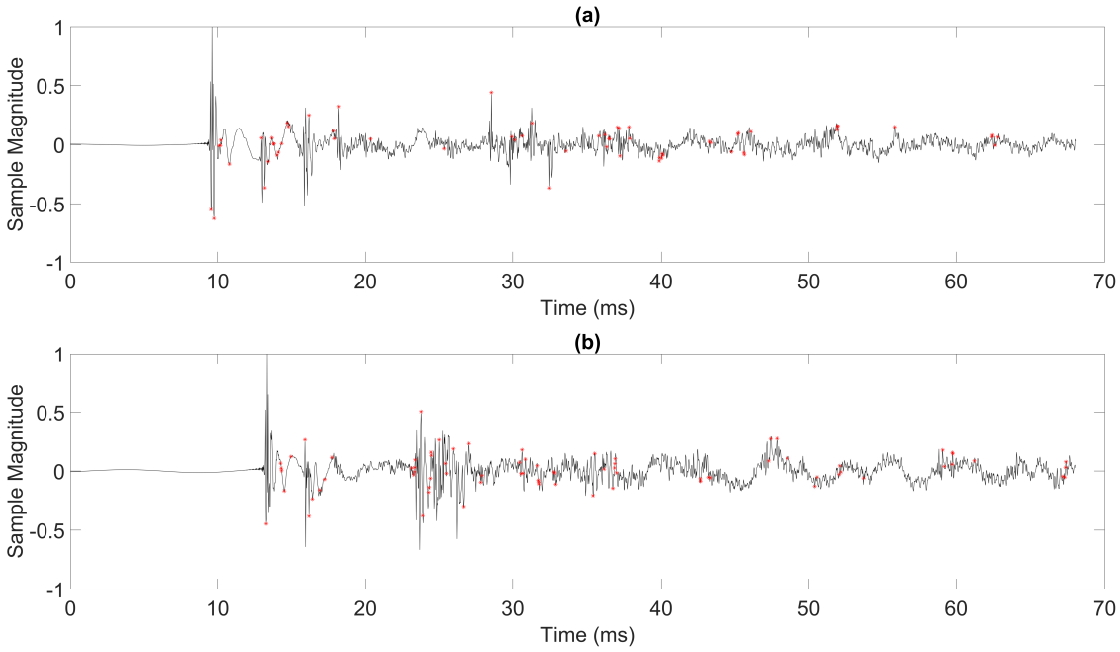

Figure A.9: The omnidirectional channel for the real-world measured SRIRs for Scenario Three, Measurement Set Two - the red asterisks indicate the locations where the EDESAR method has detected a reflection. These SRIR were measured using an EigenMike EM32 spherical microphone array, Genelec 8030 loudspeaker, and the exponential sine-sweep method.

Appendix B

List of Acronyms

ADAM Adaptive Moment

ARC Acoustic Reflection Cartographer

BRIR Binaural Room Impulse Response

C-DYPSA Clustered - Dynamic Phase-Slope Algorithm

COMEDIE Covariance Matrix Eigenvalue Diffuseness Estimation

CVLM Circular-Variance Local-Maxima

DNN Deep Neural Network

DoA direction-of-arrival

DTW Dynamic Time Warping

EB-ESPRIT Eigenbeam - Estimation of Signal Parameters via Rotational Invariance Techniques

EB-MUSIC Eigenbeam-Multiple Signal Classification

EDESAR Eigenbeam Detection and Evaluation of Simultaneously Arriving Reflections

EDM Euclidian Distance Matrix

ERB Equivalent Rectangular Bandwidth

ESPRIT Estimation of Signal parameters by Rotational Invariance Techniques

ETSAC Ellipsoid Tangent Sample Consensus

FFT Fast Fourier Transform

GMM Gaussian Mixture Model

HRIR Head Related Impulse Response

IACC Interaural Cross-Correlation

ILD Interaural Level Difference

ISDAR-LIB Image-Source Direction and Ranging-Loudspeaker-Image Bisection

ITD Interaural Time Difference

KEMAR Knowles' Electronic Manakin for Acoustic Research

MCT multi-conditional training

MLP Multilayer Perceptron

MUSIC Multiple Signal Classification

MVDR Minimum-Variance Distortionless Response

NN Neural Network

PWD Plane-Wave Decomposition

RIR Room Impulse Response

RMS root mean square

SIRR Spatial Impulse Response Rendering

SNR signal-to-noise ratio

SOM Self-Organising Map

SRIR Spatial Room Impulse Response

STFT Short Time Fourier Transform

ToA time-of-arrival

XWT Cross Wavelet Transform

Appendix C

List of Symbols

f - Frequency in Hz

c - Speed of sound

λ - wave length

p - change in pressure

ρ - instantaneous pressure

ρ_0 - Mediums resting pressure

p_{ref} - Reference pressure

x, y, z - Cartesian coordinate for x- y-, and z-axes

p_r - Right going wave pressure component

p_l - Left going wave pressure component

c - Speed of Sound

v - Particle Velocity

Z_0 - Acoustic impedance of a medium

ω - Angular frequency

k - wave number

i - $\sqrt{-1}$

$\widehat{p}$ - Pressure amplitude

p_{air} - Air pressure

ξ - Air density

T - Temperature centigrade

κ - Adiabatic Exponent

θ_i - Angle of Incidence

θ_r - Angle of Reflection

p_i - Incident Pressure

v_i - Incident Particle Velocity

p_r - Reflected Pressure

v_R - Reflected Particle Velocity

R_f - Reflection factor

Z - Acoustic impedance of a boundary

α - Absorption coefficient or x-axis directional cosine

$\mathbf{h}(t)$ - A single channel room impulse response

t - Time

δ - Dirac delta

τ - Time-of-arrival or time delay

a - Amplitude of an arriving reflection

$\mathbf{r}(t)$ - Time varying residual noise component

$\mathbf{b}$ - Point on a boundary

$\mathbf{n}$ - Unit normal for a boundary

$\mathbf{s}$ - Cartesian coordinates for the source position

$\widetilde{\mathbf{s}}$ - Cartesian coordinates for image-source locations

$\mathbf{m}$ - Cartesian coordinates for the microphone

$\widetilde{\mathbf{b}}$ - Estimated point on a boundary

$\widetilde{\mathbf{n}}$ - Estimated unit normal for a boundary

$\mathbf{X}$ - Output from a microphone array

$\mathbf{U}$ - Matrix containing the X-, Y-, and Z-channel of a B-Format signal

$\mathbf{I}$ - Instantaneous intensity or Identity matrix

$\mathbf{g}_{q,n,m}$ - Weighted spherical harmonic transform vector

$\widehat{\mathbf{X}}$ - Fourier transformed output of a microphone array

$\mathbf{w}$ - vector of weights or the omnidirectional channel of a b-format recording

$\widetilde{\mathbf{x}}$ - Steered output of a microphone array

θ - Azimuth direction-of-arrival

ϕ - Elevation direction-of-arrival

Ψ - Concatenation of azimuth and elevation direction-of-arrival

$\mathbf{y}$ - Vector of spherical harmonics

$\mathbf{Y}$ - Spherical harmonic function

$\mathbf{H}(t)$ - Multi-channel spatial room impulse response

$\mathbf{R}(t)$ - Spatially-white time-variant residual noise matrix

P_n^m - Associated Legendre polynomial of order n and degree m.

$\mathbf{\Delta}(\theta, \phi)$ - Generalised array response (spatial filter) matrix

$\mathfrak{J}$ - Spherical Bessel function

$\mathfrak{H}$ - Spherical Hankel function

$\mathbf{x}$ - Signal vector

$\Re$ - Real-component

$\hat{\mathbf{u}}$ - Unit vector pointing in the direction of the sound source

$\widetilde{d}$ - Estimated number of sources

E - Statistical expectation

$\mathbf{\Omega}$ - Generalised eigenvectors

Λ - Diagonal matrix of eigenvalues

$R_{\mathbf{XX}}$ - Covariance matrix for signal $\mathbf{X}$

$\mathbf{e}$ - Eigenvectors

$\bar{\lambda}$ - Eigenvalues

$\mathbf{p}_{\mathbf{MUSIC}}$ - MUSIC spectrum.

$\boldsymbol{E}_n$ - Noise subspace

$\mathbf{E_s}$ - Signal subspace

$\widetilde{\Delta}$ - Array displacement vector

Λ_s and Λ_n - The signal and noise diagonal subspace eigenvalues

$\mathbf{D}_0$ - EB-ESPRIT auxiliary matrix.

ζ - Directional intensity map

$\mathbf{W_x}$ - Continuous wavelet transform of signal $\mathbf{x}$

$\mathbf{x}_l$ - Left channel of a binaural signal

$\mathbf{x}_r$ - Right channel of a binaural signal

$\psi_0(\frac{n'-n}{s})$ - Translated wavelet transform

ψ_0 - Wavelet function

FFT - Fast-Fourier Transform

IFFT - Inverse Fast-Fourier Transform

H - Heaviside step function

ω_0 - The dimensionless oscillating period of the wavelet

s_0 - Smallest resolvable scale

f_λ - Fourier Period

$\mathbf{c}$ - Cross-correlation function

t_f - Time-frame

v_{t_f} - Circular Variance

s_{t_f} - Average cosine of azimuth direction-of-arrival

c_{t_f} - Average sine of azimuth direction-of-arrival

$\mathbf{W}_{\mathbf{x_{l,r}}}$ - Cross-wavelet transform of signal $\mathbf{x}_l$ and $\mathbf{x}_r$

$\mathbf{T}$ - Radon transformed image

$\mathbf{R}$ - Interpolated image

$\mathbf{g}(n)$ - Centre of gravity of a signal

$\mathbf{d}(n)$ - Phase-slope of the signal

μ_{local} - Average magnitude of a signal

$T\mu_{local}$ - The averaging time

ϵ - A threshold parameter (subscript denotes which variable it is a threshold for)

$\mathbf{r}_h$ - Residual room impulse response (matching pursuit)

$\mathbf{ds}$ - The direct sound

$\mathbf{0}$ - A vector of zeros

idx - Index

$\widehat{\mathbf{w}}_h$ - Warp vector for a reflection

$\widehat{\mathbf{w}}_a$ - Warp vector for the direct sound

$\gamma_{\mathbf{ds}_k}$ - Scaling value

υ - Error metric representing the difference between the warped direct sound and a candidate reflection

$\widetilde{\mathbf{D}}$ - Euclidean distance matrix

$\mathbf{l}$ - Vector containing the dimensions of a cuboid room

T_{60} - Reverb time

V - Volume of a room

S - Surface Area of a room

$\hat{\mathbf{t}}$ - Vector containing the time-of-arrival for a reflection at each microphone in an array

$\mathbf{t}$ - Vector containing the true time-of-arrival for a reflection at each microphone in an array

Σ_k - Time-of-arrival error covariance matrix.

D^2 - Mahalanobis distance between two points

Σ_x and Σ_y - Covariance matrix for x or y boundary

$\mathbf{n}1$ - Boundary normal vector for boundary 1

d - Distance in meters

$\mathbf{O}$ - Ellipsoid parameter matrix

$\mathbf{B}$ - Matrix containing the four corners of a boundary

$\widetilde{\mathbf{B}}$ - Matrix containing the four corners of an inferred boundary

$\mathbf{widehatT}$ - Translation matrix

$\widehat{\mathbf{R}}$ - Rotation matrix

$\widehat{\mathbf{S}}$ - Scaling matrix

$\mathbf{r}_I$ - Point of reflection on a boundary

$\widetilde{l}$ - Estimated distance between two boundaries

$\mathbf{g}_m$ - Room transfer function

$\widehat{\mathbf{X}}$ - Matrix containing the Gammatone filtered version of signal $\mathbf{x}$

$\widetilde{\mathbf{X}}$ - Cochleagram output for filtered signal $\widehat{\mathbf{X}}$

θ_{rotation} - Azimuth rotation of the binaural dummy head microphone

$\widetilde{\mathbf{x}}_0$ - Feature vector fed to the NN

μ - Mean of a vector

σ - Standard deviation of a vector.

$\mathbf{b}$ - Vector containing the bias values for each neuron in a layer of a NN

$\widetilde{\mathbf{x}}_i$ - Output activation level for each neuron in hidden layer i.

$\mathcal{P}(x|y)$ - The probability of solution x given data y

$\widehat{\mathbf{w}}$ - MVDR beamforming weights

$\mathbf{\Lambda}$ - The directional spectrum

$\widehat{\mathbf{\Lambda}}$ - The greyscale image of the directional spectrum

Δl - The difference in propagation distance

$\Delta\widetilde{\mathbf{n}}$ - Deviation from a zero z-axis coefficient for a boundary normal vector

$\Delta\angle$ - Difference in x and y directional cosines for the reflection produced from the possible previous-source to boundary and the image-source to receiver

β - y-axis directional cosine

$\widetilde{\alpha}$ - x-axis directional cosine for a ray reflecting from the point of incidence on the boundary for a line going from previous-source-to-boundary

$\widetilde{\beta}$ - y-axis directional cosine for a ray reflecting from the point of incidence on the boundary for a line going from previous-source-to-boundary

$\mathbf{p}_r$ - Point of rotation

$\bar{\mathbf{W}}$ - $[3 \times 2]$ matrix containing two points that are orthogonal to the boundary normal

$x(\mathbf{B}_1, \mathbf{B}_2)$, $y(\mathbf{B}_1, \mathbf{B}_2)$, and $z(\mathbf{B}_1, \mathbf{B}_2)$ - The x, y, and z points of intersection between boundaries $\mathbf{B}_1$ and $\mathbf{B}_2$

T_{r_o} - Time of arrival for a defined reflection order

r_o - Reflection order

Appendix D

Accompanying Material

Chapter 5:

These python scripts require the following Python libraries to be installed: Numpy V.1.17.0 [165], SciPy V.0.18.1 [166], and Tensorflow V.0.12.1 [132].

The python code was tested using Python 3.2.5, using an Anaconda Python environment.

The MATLAB code requires the freely available Malcolm Slaney's Auditory Toolbox [127] to be downloaded and placed in same folder (as these files are copyrighted, and not licenced for distribution, they are not included as part of the accompanying material).

The MATLAB code was tested using MATLAB R2018a.

The contents of folder titled Supporting_Material_Chapter_5 is as follows:

Binaural_Model

Cochleagram - This folder contains the cochleagram function [128], and corresponding licence file.

runAnalysis.m - This MATLAB script analyses the provided dataset and produces the feature vector used in testing. Users can change the variables head ('KEMAR' or 'KU100'), signalType ('directSound' or 'reflection'), and speaker ('EquatorD5' or 'Genelec8030'). **Run this script to generate the resulting normalised feature vector**

BinauralModelCochlea.m - This MATLAB function analyses a given binaural signal and outputs the interaural cross-correlation, interaural level difference, interaural time difference, the cochlea output for the left and right channel and the centre frequencies of the gammatone filter band. This function requires the following toolboxes to work: Malcolm Slaney's Auditory Toolbox [127]. This function is called by *generateFeatureVector.m*.

generateFeatureVector.m - This MATLAB function generates a feature vector from an input binaural signal x, and a version of the signal captured after the binaural dummy head has been rotated by either +90° or -90° degree (variables xPos90 and xNeg90 respectively). This function is called by the *generateTestData.m*.

generateTestData.m - This MATLAB function analyses the included binaural dataset, it takes the input variables: head - the binaural dummy head used for the measurements either 'KEMAR' or 'KU100', speaker - the speaker used for the measurements either 'EquatorD5' or 'Genelec8030', and signalType - the type of signal being analysed either 'directSound' or 'reflection'. This function is called by the *runAnalysis.m* script.

Audio Files

- 144 direct sound components captured with the KEMAR 45BC binaural dummy head microphone and the Equator D5 speaker.
- 144 reflected components captured with the KEMAR 45BC binaural dummy head microphone and the Equator D5 speaker.
- 144 direct sound components captured with the KU100 binaural dummy head microphone and the Equator D5 speaker.
- 144 reflected components captured with the KU100 binaural dummy head microphone and the Equator D5 speaker.
- 144 direct sound components captured with the KEMAR 45BC binaural dummy head microphone and the Genelec 8030 speaker.
- 144 reflected components captured with the KEMAR 45BC binaural dummy head micro-

phone and the Genelec 8030 speaker.

- 144 direct sound components captured with the KU100 binaural dummy head microphone and the Genelec 8030 speaker.

- 144 reflected components captured with the KU100 binaural dummy head microphone and the Genelec 8030 speaker.

Generate_Training_Data

Cochleagram - This folder contains the cochleagram function [128], and corresponding licence file.

BinauralModelRun.m - This MATLAB script is used to run the binaural model and generate the feature spaces for the KEMAR SADIE Database [123] - Requires the HRIR from the SADIE database to run. **Run this script to generate compute the binaural model output for each HRIR in the SADIE dataset.**

GenerateHeadRotation_SWN.m - MATLAB Script used to generate the training data matrix.

binaryClassifierAzOnly - MATLAB function that generate a binary classifier for a set of azimuth directions-of-arrival.

KEMARSADIETarget.mat - List of all the directions-of-arrival contained within the KEMAR SADIE database [123]

spatialWhiteNoise.wav - The spatially white noise generated by convolving white Gaussian noise with all HRIRs in the KEMAR SADIE database, and taking the mean.

KEMARSADIE_BinauralModelOut_CochleaModel_1sOverlap.mat - Interaural-cross correlation, interaural level difference, and interaural time difference for the HRIRs within the KEMAR SADIE database.

NN_Training_Scripts

NNTrainRun.py - Python script used to run the training procedure for the NN. **Run this script in a python command line environment, such as Anaconda, to train the NN.**

CFFNNRunTrained.py - Python script used to test the neural network once trained. This script is called by *NNTrainRun.py*.

CFFNNRunTraining.py - Python script used to test the neural network during training. This script is called by *NNTrainRun.py*.

CFFNNRunTrain.py - Python script that performs a single training iteration. This script is called by *NNTrainRun.py*.

initCFFNN_2.py - This Python Script is used to initialises the cascade-forward neural network.

Test_Data - This folder contains the test data for the KEMAR Equator Dataset.

Training_Data - This folder contains the training data used to train the neural network.

Trained_NN_Scripts

AnalyseDoA.py - Python script which can be run to test the pre-trained neural network using the pre-generated test data. Upon running the script the user will be prompted to select different test data options. The variable *DoA* contains the estimated directions-of-arrival made by the neural network, and the variable *yDiff* contains the difference between the estimated and expected directions-of-arrival. **Run this script in a python command line environment, such as Anaconda, to produce estimates of DoA for an input feature matrix.**

DirectionAnalysis.py - Python script containing a set of functions used to define and run the pre-trained neural network. Called by *AnalyseDoA.py*.

noLayers.txt - A text file containing the number of layers used when training the neural network - one in this case.

neg90 - This folder contains the Gaussian normalisation parameters stored as text files and the weights and biases for the trained neural network - these are all for the -90° rotation neural network.

pos90 - This folder contains the Gaussian normalisation parameters stored as text files and the weights and biases for the trained neural network - these are all for the +90° rotation neural network.

testData - This folder contains pre-generated test data for the different binaural dummy head microphones, speaker, and signal type combinations.

Thesis_Data_Analyses

runAnalysisOfData.m - This MATLAB script is used to analyse the direction-of-arrival estimates produced by the neural networks as used in this thesis. Generates a table where column one is the number of exact estimations, column two is the number of estimations within $\pm 1^{\circ}$, column three is the number of estimations within $\pm 5^{\circ}$, column four is the percentage of front/back confusions, column five is the mean relative error, and column six is the RMS error. **Run this script to generate the results presented in Section 5.6 of this thesis.**

frontBackCheck.m - MATLAB function used to find front/back confusions within the estimated directions-of-arrival. This function is called by *runAnalysisOfData.m*.

The '.mat' files containing: the estimated directions-of-arrival for each test case, the difference between the estimated and expected direction-of-arrival, and the expected direction-of-arrival.

Chapter 6:

The contents of folder titled Supporting_Material_Chapter_6 is as follows:

Folder - Scenario One

CVLM - Folder containing the code for the implementation of the circular variance local maxima method. *ReflectionDetection.m* is the main function.

DTW - Folder containing the code for the implementation of the dynamic time warping matching pursuit method. *DTWReflectionDetection.m* is the main function (currently set up to work with the random pulse only as the DS is hard coded)

EDESAR - Contains the function *EDESAR.m* which is used to detect reflections using the proposed Eigenbeam Detection and Evaluation of Simultaneously Arriving Reflections (EDESAR) method, and the MVDR and Steered response power map functions.

External_Code - This folder contains the MATLAB functions and licences for external coded needed to run the EDESAR method.

createDummySRIR.m - This function generates a train of random pulses.

EDESAR_Test_PhD_Random.m - MATLAB script that generates a random train of pulses and analyses them using the EDESAR method. **Run this script to randomly generate a train of pulses and analyse it using the proposed EDESAR method.**

plotComparisonResults_Random.m - Script used to generate the results presented in Chapter Six - Scenario One. **Run this script to generate the figures and results presented in Section 6.5.1**

pulse.wav - Audio file containing the pulse used to generate the random train of pulses.

Workspace_RandomIR1.mat - .mat filed containing the EDESAR results for the random train of pulses.

Workspace_RandomIR1_CV.mat - .mat filed containing the CVLM results for the random train of pulses.

Workspace_RandomIR1_DTW.mat - .mat filed containing the DTW results for the random train of pulses.

reflectionDetectionCVLM.m - MATLAB script that runs the CVLM reflection detection method on the random train of pulses.

reflectionDetectionDTW.m - MATLAB script that runs the DTW reflection detection method on the random train of pulses.

Folder - Scenario Two

CVLM - Folder containing the code for the implementation of the circular variance local maxima method. *ReflectionDetection.m* is the main function.

DTW - Folder containing the code for the implementation of the dynamic time warping matching pursuit method. *DTWReflectionDetection.m* is the main function.

EDESAR - Contains the function *EDESAR.m* which is used to detect reflections using the proposed Eigenbeam Detection and Evaluation of Simultaneously Arriving Reflections (EDESAR) method, and the MVDR and Steered response power map functions.

External_Code - This folder contains the MATLAB functions and licences for external coded needed to run the EDESAR method.

SRIR - Folder containing the simulated SRIR

EDESAR_Test_PhD_Simulated.m - MATLAB script that analyses the simulated SRIR using the EDESAR method. **Run this script to estimate the ToA for reflections present in the simulated SRIR using the proposed EDESAR method.**

plotComparisonResults_Simulated.m - Script used to generate the results presented in Chapter Six - Scenario Two. **Run this script to generate the figures and results presented in Section 6.5.2**

Workspace_Simulated_EDESAR.mat - .mat filed containing the EDESAR results for the simulated SRIR.

Workspace_Simulated_CV.mat - .mat filed containing the CVLM results for the simulated SRIR.

Workspace_Simulated_DTW.mat - .mat filed containing the DTW results for the simulated SRIR.

reflectionDetection_Simulated_CVLM.m - MATLAB script that runs the CVLM reflection detection method on the simulated SRIR. **Run this script to estimate the ToA for reflections present in the simulated SRIR using the CVLM method.**

reflectionDetection_Simulated_DTW.m - MATLAB script that runs the DTW reflection detection method on the simulated SRIR. **Run this script to estimate the ToA for reflections present in the simulated SRIR using the DTW Matching Pursuit method.**

Folder - Scenario Three

CVLM - Folder containing the code for the implementation of the circular variance local maxima method. *ReflectionDetection.m* is the main function.

DTW - Folder containing the code for the implementation of the dynamic time warping matching pursuit method. *DTWReflectionDetection.m* is the main function.

EDESAR - Contains the function *EDESAR.m* which is used to detect reflections using the pro-

posed Eigenbeam Detection and Evaluation of Simultaneously Arriving Reflections (EDESAR) method, and the MVDR and Steered response power map functions.

External_Code - This folder contains the MATLAB functions and licences for external coded needed to run the EDESAR method.

SRIR - Folder containing the simulated SRIR

EDESAR_Test_PhD_Real.m - MATLAB script that analyses the simulated SRIR using the EDESAR method. **Run this script to estimate the ToA for reflections present in the real-world SRIR using the EDESAR method.**

plotComparisonResults_Real.m - Script used to generate the results presented in Chapter Six - Scenario Three.

Workspace_Simulated_EDESAR.mat - .mat filed containing the EDESAR results for the real-world SRIR.

Workspace_Real_CV.mat - .mat filed containing the CVLM results for the real-world SRIR.

Workspace_Real_DTW.mat - .mat filed containing the DTW results for the real-world SRIR.

reflectionDetection_Real_CVLM.m - MATLAB script that runs the CVLM reflection detection method on the real-world SRIR. **Run this script to estimate the ToA for reflections present in the real-world SRIR using the CVLM method.**

reflectionDetection_Real_DTW.m - MATLAB script that runs the DTW reflection detection method on the real-world SRIR. **Run this script to estimate the ToA for reflections present in the real-world SRIR using the DTW Matching Pursuit method.**

Chapter 7:

The contents of folder titled Supporting_Material_Chapter_7 is as follows:

Folder - Scenario One

Additional_Functions - Folder containing MATLAB functions used to plot the data.

Geometry_Inference - MATLAB folder containing all the functions used to infer the geometry of the room, contents are as follows:

- *computeDirectionalCosines.m* - Computes the direction cosines for a vector defining a ray.

- *constrainPlane.m* - Constrain the boundary based on the points of intersection with neighbouring non-parallel boundaries.

- *constrainPlane2.m* - Constrain the boundary based on the points of intersection with neighbouring non-parallel boundaries (path from image-source-to-receiver must intersect with boundary).

- *constrainRoomFromPlanes.m* - Function used to refine the candidate boundaries to ideally those of the desired. Calls the functions: *generateInferredPlanes.m*, *generateInferredPlanes2.m*, *removePlanes_InteriorPathwayInvalidation.m*, *removePlanes_lineOfSight.m*, *removePlanes_notConnected.m*, and *removePlanes_source2ReceiverPathInvalidation.m*.

- *findPreviousSoruce.m* - Function used to find the most likely candidate previous-source for each image-source.

- *generateFloorCeiling.m* - Function that generates the inferred boundaries for the floor and ceiling based on the corners of all of the inferred boundaries.

- *generateInferredPlanes.m* - Computes the points of intersections between boundaries using *planePlaneIntersection.m*, and then constrains each boundary using *constrainPlane.m*.

- *generateInferredPlanes2.m* - Computes the points of intersections between boundaries using *planePlaneIntersection.m*, and then constrains each boundary using *constrainPlane2.m*.

- *generatePlane.m* - Generate a candidate boundary using a point-on-the-boundary and the boundary's normal vector.

- *generateUnconstrainedPlane.m* - Generate a boundary for each image-source and previous-source pair, removing cases when the same boundary is inferred multiple times and cases when the boundary is inferred too close to the source/receiver.

- *inferGeometry.m* - Function that calls the boundary refinement and *removePlanes_Coincident.m* functions. This function also plots all inferred unconstrained boundaries on figure(100)

- *inferImageSource.m* - Infer the Cartesian coordinates for each image-source based on the time- and direction-of-arrival for each reflection.

- *linePlaneIntersection.m* - Find the point of intersection between a line and a boundary, where an infinite boundary and finite length line are assumed.

- *linePlaneIntersection_Constrained.m* - Find the point of intersection between a line and a boundary, where a finite length boundary and finite length line are assumed.

- *linePlaneIntersection_Constrained_rayNotBound.m* - Find the point of intersection between a line and a boundary, where a finite length boundary and infinite length line are assumed.

- *newDirCos.m* - Generate a set of directional cosines for a ray being reflected off a boundary, as defined in [153].

- *planePlaneIntersection.m* - Compute the point of intersection between two boundaries.

- *receiverIntersection.m* - Compute the point of intersection between a line and the receiver location.

- *removePlanes_Coincident.m* - Remove boundaries that are coincident - leaving one remaining.

- *removePlanes_InteriorPathwayInvalidation.m* - Check reflection paths between image-source-to-receiver searching for reflection paths from boundary-to-receiver that are occluded by another boundary.

- *removePlanes_lineOfSight.m* - Check that each boundary is in line-of-sight with the corresponding receiver, removing any that are not.

- *removePlanes_notConnected.m* - Check that all boundaries are connected to at least two other boundaries, removing any that are not.

- *removePlanes_source2ReceiverPathInvalidation.m* - Check that the path between the source-to-receiver is not occluded by any boundaries.

- *runGeometryInference.m* - This is the main function that calls the entire geometry inference process.

EDESAR - Contains the function *EDESAR.m* which is used to detect reflections using the proposed Eigenbeam Detection and Evaluation of Simultaneously Arriving Reflections (EDESAR) method, and the MVDR and Steered response power map functions.

External_Code - This folder contains the MATLAB functions and licences for external coded needed to run the EDESAR method.

SRIR - Folder congaing all the simulated SRIR used for Scenario One.

HigherOrderAmbi_testCase.m - MATLAB scripts used to run the test cases used in this thesis (replace testCase with Cuboid/LShaped/etc.) **Run these scripts to infer the geometry of the room for each test case presented in 7.5.1.**

Folder - Scenario Two

Additional_Functions - Folder containing MATLAB functions used to plot the data.

Geometry_Inference - MATLAB folder containing all the functions used to infer the geometry of the room, contents are as previously discussed.

EDESAR - Contains the function *EDESAR.m* which is used to detect reflections using the proposed Eigenbeam Detection and Evaluation of Simultaneously Arriving Reflections (EDESAR) method.

External_Code - This folder contains the MATLAB functions and licences for external coded needed to run the EDESAR method.

Analysed_Reflection_Data(2) - Folders containing all the results from the EDESAR method for the SRIRs used for Scenario Two.

L_Shaped_Multiple_SR_Combinations.m - Runs the geometry inference process for all 33 source

combinations for L-Shaped Room One Scenario Two as used in this thesis - these combinations are chosen as they ensure the first-order reflection constrain is met, alternate combinations will not necessarily produce valid geometry inference as a result of this requirement not being met. **Run this scripts to infer the geometry of the room for the first L-Shaped Room presented in 7.5.2.**

L_Shaped_Multiple_SR_Combinations2.m - Runs the geometry inference process for all 33 source combinations for L-Shaped Room Two Scenario Two as used in this thesis - these combinations are chosen as they ensure the first-order reflection constrain is met, alternate combinations will not necessarily produce valid geometry inference as a result of this requirement not being met. **Run this scripts to infer the geometry of the room for the second L-Shaped Room presented in 7.5.2.**

boundaryCombinationsLShaped.mat - .mat file containing the inferred boundaries belong to each desired boundary for the first L-Shaped room.

boundaryCombinationsLShaped2.mat - .mat file containing the inferred boundaries belong to each desired boundary for the second L-Shaped room.

Folder - Scenario Three

Additional_Functions - Folder containing MATLAB functions used to plot the data.

Geometry_Inference - MATLAB folder containing all the functions used to infer the geometry of the room, contents are as previously discussed.

EDESAR - Contains the function *EDESAR.m* which is used to detect reflections using the proposed Eigenbeam Detection and Evaluation of Simultaneously Arriving Reflections (EDESAR) method.

External_Code - This folder contains the MATLAB functions and licences for external coded needed to run the EDESAR method.

SRIR - Folder congaing all the simulated SRIR used for Scenario Three.

RealWorld_GeometryInference_Set1.m - This script runs the geometry inference process for Scenario Three, measurement set one. **Run this scripts to infer the geometry of the first real-**

world cuboid-shaped room presented in 7.5.3.

RealWorld_GeometryInference_Set2.m - This script runs the geometry inference process for Scenario Three, measurement set two. **Run this scripts to infer the geometry of the second real-world cuboid-shaped room presented in 7.5.3.**

Bibliography

[1] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012. [Online]. Available: https://ieeexplore.ieee.org/document/6296524. [Accessed: Sept. 10, 2019]

[2] A. Asaei, M. Golbabaee, H. Bourlard, and V. Cevher, "Structured sparsity models for reverberant speech separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 3, pp. 620–633, 2014. [Online]. Available: https://ieeexplore.ieee.org/document/6698345. [Accessed: Sept. 10, 2019]

[3] E. A. Habets and J. Benesty, "A two-stage beamforming approach for noise reduction and dereverberation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 5, pp. 945–958, 2013. [Online]. Available: https://ieeexplore.ieee.org/document/6409417. [Accessed: Sept. 10, 2019]

[4] I. Dokmanic, L. Daudet, and M. Vetterli, "From acoustic room reconstruction to slam," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings*, vol. 2016-May, pp. 6345–6349, 2016. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7472898. [Accessed: Sept. 10, 2019]

[5] A. Canclini, D. Marković, F. Antonacci, A. Sarti, and S. Tubaro, "A room-compensated virtual surround system exploiting early reflections in a reverberant room," *European Signal Processing Conference*, no. Eusipco, pp. 1029–1033, 2012. [Online]. Available: https://www.eurasip.org/Proceedings/Eusipco/Eusipco2012/Conference/papers/1569580563.pdf

[6] I. Dokmanic, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli, "Acoustic echoes reveal room shape," in *Proceedings of the National Academy of Sciences 2013*, vol.

110, pp. 12 186–12 191. [Online]. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.1221464110 [Accessed: May. 31, 2019].

[7] S. Tervo and T. Tossavainen, "3D Room Geometry Estimation from Measured Impulse Responses," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2012*, pp. 513–516. [Online]. Available: https://ieeexplore.ieee.org/document/6287929 [Accessed: May. 31, 2019].

[8] L. Remaggi, P. J. B. Jackson, P. Coleman, and W. Wang, "Acoustic Reflector Localization : Novel Image Source Reversion and Direct Localization Methods," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 2, pp. 296–309, 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7762761/. [Accessed: May. 31, 2019]

[9] Y. E. Baba, A. Walther, and E. A. Habets, "3D room geometry inference based on room impulse response stacks," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 26, no. 5, pp. 857–872, 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8214218. [Accessed: May. 31, 2019]

[10] G.R.A.S, (n.d.), "Head & Torso Simulators," G.R.A.S, http://www.gras.dk/. [Online]. Available: http://www.gras.dk/products/head-torso-simulators-kemar.html [Accessed: Sept. 21, 2016].

[11] Neumann, (n.d.), "Dummy Head KU100," Neumann, https://www.neumann.com/. [Online]. Available: https://www.neumann.com/?lang=en&id=current_microphones&cid=ku100_description [Accessed: Sept. 21, 2016].

[12] H. Sun, E. Mabande, K. Kowalczyk, and W. Kellermann, "Localization of distinct reflections in rooms using spherical microphone array eigenbeam processing," *The Journal of the Acoustical Society of America*, vol. 131, no. 4, pp. 2828–2840, 2012. [Online]. Available: https://asa.scitation.org/doi/abs/10.1121/1.3688476?journalCode=jas. [Accessed: July 22 2019]

[13] MH Acoustics, (2013), "EigenStudio." [Online]. Available: https://mhacoustics.com/products [Accessed: May. 31, 2019].

[14] S. Tervo, J. Pätynen, and T. Lokki, "Acoustic reflection path tracing using a highly directional loudspeaker," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 245–248, 2009. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/5346530. [Accessed: May. 31, 2019]

[15] I. Kelly and F. Boland, "Detecting Arrivals in Room Impulse Responses with Dynamic Time Warping," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 7, pp. 1139–1147, 2013, [Accessed: May. 31, 2019].

[16] CATT, 2016, "CATT-Acoustic," CATT, http://www.catt.se/. [Online]. Available: http://www.catt.se/ [Accessed: May. 31, 2019].

[17] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, p. 943, 1979, [Accessed: May. 31, 2019].

[18] H. Kuttruff, *Room Acoustics*. London: Applied Science Publishers LTD, 1973.

[19] M. Vorländer, *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality.* Berlin: Springer-Verlag Berlin and Heidelberg GmbH & Co. K, 2008.

[20] F. Everest, Alton, *The Master Handbook of Acoustics*, 4th ed. NewYork: McGraw Hill, 2001.

[21] D. Howard and J. Angus, *Acoustics and Psychoacoustics*, 4th ed. United Kingdom: Elsevier Science, 2009.

[22] A. Kohlrausch, J. Braasch, D. Kolosssa, and J. Blauert, "An Introduction to Binaural Processing," in *The Technology of Binaural Listening*, J. Blauert, Ed. Springer-Verlag Berlin and Heidelberg, 2013, ch. 1, pp. 1–32.

[23] T. M. Mitchell, *Machine Learning*, 1st ed. New York: New York; London : McGraw-Hill, 1997.

[24] C. Stergiou and D. Siganos, "Neural Networks," Imperial College London. [Online]. Available: https://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html [Accessed: Oct. 11, 2016].

[25] B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988. [Online]. Available: https://ieeexplore.ieee.org/document/665. [Accessed: July. 19, 2019]

[26] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, "3D Source localization in the spherical harmonic domain using a pseudointensity vector," in *18th European Signal Processing Conference 2010*, no. 5, Aalborg, Denmark, pp. 442–446. [Online]. Available: http://ieeexplore.ieee.org/document/7096575/ [Accessed: Sept. 26, 2016].

[27] H. Teutsch, *Modal Array Signal Processing: Principles and Applications of Acoustic Wavefield Decomposition*. Berlin: New York: Springer, 2007.

[28] B. Rafaely, "Plane-wave Decomposition of the pressure on a sphere by spherical convolution," *ISVR Techical Memorandum*, vol. 910, 2003. [Online]. Available: http://eprints.soton.ac.uk/46555/1/Pub9273.pdf. [Accessed: July. 17, 2019]

[29] D. P. Jarrett, E. A. Habets, and P. A. Naylor, *Theory and Applications of Spherical Microphone Array Processing*. Springer International Publishing, 2017.

[30] A. Politis, 2016, "Spherical Array Porcessing Toolbox." [Online]. Available: https://github.com/polarch/Spherical-Array-Processing [Accessed: May. 31, 2019].

[31] P. M. Morse and K. U. Ingard, *Theoretical Acoustics*. New York: McGraw Hill, 1968.

[32] N. Ma, G. J. Brown, and T. May, "https://arxiv.org/pdf/1904.03001.pdf," in *Interspeech 2015*, pp. 1–5. [Online]. Available: http://orbit.dtu.dk/fedora/objects/orbit:140619/datastreams/file_111863222/content [Accessed: May. 31, 2019].

[33] H. E. de Bree, P. Leussink, T. Korthorst, H. Jansen, T. S. Lammerink, and M. Elwenspoek, "The μ-flown: a novel device for measuring acoustic flows," *Sensors and Actuators A: Physical*, vol. 54, no. 1, pp. 552 – 557, 1996. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0924424797800131. [Accessed: Jan. 09, 2020]

[34] J. Merimaa and V. Pulkki, "Spatial impulse response rendering I: Analysis and synthesis," *AES: Journal of the Audio Engineering Society*, vol. 53, no. 12, pp. 1115–1127, 2005. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=13401. [Accessed: May. 31, 2019]

[35] S. Shelley, D. Murphy, and A. Chadwick, "B-Format Acoustic Impulse Response Measurement and Analysis In the Forest at Koli National Park, Finland," in *Proceedings of the 16th international conference on Digital Audio Effects (DAFx-13) 2013*, Maynooth, Ireland, pp. 1–5. [Online]. Available: http://eprints.whiterose.ac.uk/78001/1/56.dafx2013_submission_26.pdf [Accessed: May. 31, 2019].

[36] G. Kearney, "Auditory Scene Synthesis using Virtual Acoustic Recording and Reproduction," PhD Thesis, Department of Electronics and Electrical Engineering, Trinity College Dublin, Dublin, 2010.

[37] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal of the*

Audio Engineering Society, vol. 55, no. 6, pp. 503–516, 2007. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=14170. [Accessed: Sept. 11 2019]

[38] A. H. Moore, C. Evers, and P. A. Naylor, "2D Direction of Arrival Estimation of Multiple Moving Sources using a Spherical Microphone Array," in *24th European Signal Processing Conference 2016*, pp. 1217–1221. [Online]. Available: https://ieeexplore.ieee.org/document/7760442 [Accessed: July 23 2019].

[39] A. H. Moore, C. Evers, and P. A. Naylor, "Direction of Arrival Estimation in the Spherical Harmonic Domain Using Subspace Pseudointensity Vectors," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 25, no. 1, pp. 178–192, 2017. [Online]. Available: https://ieeexplore.ieee.org/document/7575731. [Accessed: Jan. 09, 2020]

[40] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1143830. [Accessed: Oct. 04, 2016]

[41] MathWorks, (2019), "cov." [Online]. Available: https://www.mathworks.com/help/matlab/ref/cov.html [Accessed: Jan. 09, 2020].

[42] R. Roy and T. Kailath, "ESPRIT – Estimation of Signal Parameters via Rotational Invariance Techniques," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 7, pp. 984–995, 1989. [Online]. Available: http://ieeexplore.ieee.org/document/32276/. [Accessed: Sept. 29, 2016]

[43] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 387–392, 1985. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1164557. [Accessed: July 23 2019]

[44] MathWorks, (n.d.), "Dimensions of signal subspace," MathWorks, http://uk.mathworks.com/. [Online]. Available: https://uk.mathworks.com/help/phased/ref/aictest.html [Accessed: Oct. 03, 2016].

[45] J. Dmochowski, J. Benesty, and S. Affes, "Broadband Music: Opportunities and Challenges for Multiple Source Localization," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics 2007*, no. 2, New Paltz, NY, pp. 18–21.

[Online]. Available: http://www.wirelesslab.ca/File/pdf_files/conf/C105.pdf [Accessed: Oct. 04, 2016].

[46] B. Rafaely, Y. Peled, M. Agmon, D. Khaykin, and E. Fisher, *Spherical Microphone Array Beamforming*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 281–305, [Accessed: May. 31, 2019]. [Online]. Available: https://doi.org/10.1007/978-3-642-11130-3_11

[47] B. P. Ng, M. H. Er, and C. Kot, "Array gain/phase calibration techniques for adaptive beamforming and direction finding," *IEE Proceedings: Radar, Sonar and Navigation*, vol. 141, no. 1, pp. 25–29, 1994. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=300390. [Accessed: Jan. 12, 2020]

[48] Z. Ye and C. Liu, "Non-sensitive adaptive beamforming against mutual coupling," *IET Signal Processing*, vol. 3, no. 1, pp. 1—-6, 2007. [Online]. Available: https://ieeexplore.ieee.org/document/4745839. [Accessed: Jan. 12, 2020]

[49] B. Liao, S.-C. Chan, and K.-M. Tsui, "Recursive steering vector estimation and adaptive beamforming under uncertainties," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 49, no. 1, pp. 489–501, 2013. [Online]. Available: https://ieeexplore.ieee.org/document/6404116. [Accessed: Jan. 12, 2020]

[50] D. Khaykin and B. Rafaely, "Acoustic analysis by spherical microphone array processing of room impulse responses," *The Journal of the Acoustical Society of America*, vol. 132, no. 1, 2012. [Online]. Available: https://asa.scitation.org/doi/abs/10.1121/1.4726012?journalCode=jas. [Accessed: May. 31, 2019]

[51] N. Huleihel and B. Rafaely, "Spherical array processing for acoustic analysis using room impulse responses and time-domain smoothing," *The Journal of the Acoustical Society of America*, vol. 133, no. June 2013, pp. 3995–4007, 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23742353. [Accessed: May. 31, 2019]

[52] A. L. Swindlehurst, B. Ottersten, R. Roy, and T. Kailath, "Multiple Invariance ESPRIT," *IEEE Transactions on Signal Processing*, vol. 40, no. 4, 1992. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=127959. [Accessed: Oct. 03, 2016]

[53] H. Sun, H. Teutsch, E. Mabande, and W. Kellermann, "Robust localization of multiple sources in reverberant environments using EB-ESPRIT with spherical microphone arrays," in *ICASSP, IEEE International Conference on Acoustics, Speech*

and Signal Processing, Proceedings 2011, no. 4, pp. 117–120. [Online]. Available: https://ieeexplore.ieee.org/document/5946342 [Accessed: May. 31, 2019].

[54] H. Teutsch and W. Kellermann, "Detection and localization of multiple wideband acoustic sources based on wavefield decomposition using spherical apertures," in *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP). 2008*, no. 3, Las Vegas, NV, pp. 5276–5279. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4518850 [Accessed: May. 31, 2019].

[55] C.-I. C. Nilsen, I. Hafizovic, and S. Holm, "Robust 3D Sound Source Localization Using Spherical Microphone Arrays," *AES Convetion 134*, 2013. [Online]. Available: https://secure.aes.org/forum/pubs/conventions/?elib=16804. [Accessed: July 22 2019]

[56] B. Rafaely, "Phase-Mode versus Delay-and-Sum Spherical," *IEEE Signal Processing Letters*, vol. 12, no. 10, pp. 713–716, 2005. [Online]. Available: https://ieeexplore.ieee.org/document/1510666. [Accessed: July 23 2019]

[57] B. Rafaely, "Plane-wave decomposition of the sound field on a sphere by spherical convolution," *The Journal of the Acoustical Society of America*, vol. 116, no. 4, pp. 2149–2157, 2004. [Online]. Available: https://asa.scitation.org/doi/10.1121/1.1792643

[58] D. P. Jarrett, "Spherical Microphone Array Porcessing For Acoustic Parameter Estimation and Signal Enhancement," Ph.D. dissertation, Imperial College London, 2013.

[59] M. Park and B. Rafaely, "Sound-field analysis by plane-wave decomposition using spherical microphone array," *The Journal of the Acoustical Society of America*, vol. 118, no. 5, p. 3094, 2005. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/118/5/10.1121/1.2063108. [Accessed: Jan. 30, 2020]

[60] S. Delikaris-Manias, D. Pavlidiy, V. Pulkki, and A. Mouchtaris, "3D localization of multiple audio sources utilizing 2D DOA histograms," *European Signal Processing Conference*, vol. 2016-Novem, no. 1, pp. 1473–1477, 2016. [Online]. Available: https://ieeexplore.ieee.org/document/7760493. [Accessed: July 23 2019]

[61] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Adaptive Antennas for Wireless Communications*, vol. 57, no. 8, pp. 1408–1418, 1969. [Online]. Available: https://ieeexplore.ieee.org/document/1449208. [Accessed: May. 31, 2019]

[62] MathWorks, (2019), "mvdrweights," Massachusetts. [Online]. Available: https://www.mathworks.com/help/phased/ref/mvdrweights.html [Accessed: Jan. 18, 2020].

[63] D. P. Jarrett, E. A. Habets, and P. A. Naylor, "Spherical harmonic domain noise reduction using an MVDR beamformer and DOA-based second-order statistics estimation," *IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings (ICASSP)*, pp. 654–658, 2013. [Online]. Available: https://ieeexplore.ieee.org/document/6637729. [Accessed: May. 31, 2019]

[64] S. Vesa and T. Lokki, "Segmentation and Analysis of Early Reflections From a Binaural Room Impulse Response." *Technical Reports in Media Technology, Helsinki University of Technology*, pp. 1–10, 2009. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.183.3778&rep=rep1&type=pdf. [Accessed: Oct. 04, 2016]

[65] C. Torrence and G. P. Compo, "A Practical Guide to Wavelet Analysis," *Bulletin of the American Meteorological Society*, vol. 79, no. 1, pp. 61–78, 1998. [Online]. Available: http://journals.ametsoc.org/doi/pdf/10.1175/1520-0477%281998%29079%3C0061%3AAPGTWA%3E2.0.Co%3B2. [Accessed: Oct. 07, 2016]

[66] K. D. Donohue, "Audio Array Toolbox," University of Kentucky. [Online]. Available: http://vis.uky.edu/distributed-audio-lab/about/ [Accessed: Oct. 10, 2016].

[67] B. Gardner and K. Martin, (2000), "HRTF Measurements of a KEMAR Dummy-Head Microphone," MIT Media Lab Machine Listening Group. [Online]. Available: http://sound.media.mit.edu/resources/KEMAR.html [Accessed: Oct. 10, 2016].

[68] V. R. Algazi, R. O. Duda, and D. M. Thompson, (2015), "The CIPIC HRTF Database," CIPIC Interface Laboratory. [Online]. Available: https://www.ece.ucdavis.edu/cipic/spatial-sound/hrtf-data/ [Accessed: Oct. 10, 2016].

[69] H. Abdi, D. Valentin, and B. Edelman, *Neural Networks*. California: SAGE Publications, Inc., 1999.

[70] C. Neti, E. D. Young, and M. H. Schneider, "Neural network models of sound localization based on directional filtering by the pinna." *Journal of the Acoustical Society of America*, vol. 92, no. December 1992, pp. 3140–3156, 1992. [Online]. Available: https://asa.scitation.org/doi/abs/10.1121/1.404210. [Accessed: Oct. 10, 2016]

[71] B. P. Yuhas, "Automated Sound Localization Through Adaptation," in *IJCNN., International Joint Conference on Neural Networks. (Volume 2) 1992*, Baltimore, MD, pp. II–907 – II–912. [Online]. Available: http://ieeexplore.ieee.org/document/226872/ [Accessed: Oct. 10, 2016].

[72] S. Jha and T. Durrani, "Direction of arrival estimation using artificial neural networks," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 21, no. 5, pp. 1192–1201, 1991. [Online]. Available: http://ieeexplore.ieee.org/document/120069/. [Accessed: Oct. 10, 2016]

[73] D. Goryn and M. Kaveh, "Neural Networks for Narrowband and Wideband Direction Finding," in *International Coference on Acoustics, Speech and Signal Processing (ICASSP-88) 1988*, New York, USA, pp. 2164–2167. [Online]. Available: http://ieeexplore.ieee.org/document/197061/ [Accessed: Oct. 10, 2016].

[74] C. Neti, M. H. Schneider, and E. D. Young, "Maximally Fault Tolerant Neural Networks," *IEEE Transactions on Neural Networks*, vol. 3, no. 1, pp. 14–23, 1992. [Online]. Available: http://ieeexplore.ieee.org/document/105414/. [Accessed: Oct. 10, 2016]

[75] R. Rastogi, R. Gupta, and R. Kumaresan, "Array Signal Processing with Interconnected Neuron-like Elements," in *IEEE International Conference on Acoustics Speech and Signal Processing 1987*, Dallas, Texas, pp. 2328–2331. [Online]. Available: http://ieeexplore.ieee.org/document/1169330/ [Accessed: Oct. 10, 2016].

[76] K. Palomäki, V. Pulkki, and M. Karjalainen, "Neural network approach to analyze spatial sound," in *AES 16th International Conference: Spatial Sound Reproduction 1999*, Rovaniemi, Finland, pp. 233–245. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=8032 [Accessed: Oct. 10, 2016].

[77] S. K. Jha and T. Durrani, "Bearing estimation using Neural Optimisation methods," in *International Conference on Acoustics, Speech and Signal Processing, 1990. ICASSP-90*, Albuquerque, New Mexico, pp. 129–133. [Online]. Available: http://ieeexplore.ieee.org/document/115984/ [Accessed: Oct. 10, 2016].

[78] H. O'Dwyer, E. Bates, and F. M. Boland, "A machine learning approach to detecting sound-source elevation in adverse environments," in *Audio Engineering Society Convention 144 2018*, [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=19485 [Accessed: Jan. 7, 2020].

[79] MathWorks, (2016), "fitnet." [Online]. Available: http://uk.mathworks.com/help/nnet/ref/fitnet.html [Accessed: Oct. 24, 2016].

[80] R. Raul, *Neural Networks A Systematic Introduction*. Berlin: Springer-Verlag

Berlin Heidelberg, 1996, [Accessed: May. 31, 2019]. [Online]. Available: http://page.mi.fu-berlin.de/rojas/neural/neuron.pdf

[81] D. J. C. MacKay, "Bayesian Interpolation," *Neural Computation*, vol. 4, no. 3, pp. 415–447, 1992. [Online]. Available: http://www.mitpressjournals.org/doi/abs/10.1162/neco.1992.4.3.415. [Accessed: Dec. 1, 2017]

[82] D. W. Marquardt, "An Algorithm for Least-Squares Estimation of Nonlinear Parameters," *Journal of the Society for Industrial and Applied Mathematics*, vol. 11, no. 2, pp. 431–441, 1963. [Online]. Available: https://www.jstor.org/stable/2098941?seq=8#metadata_info_tab_contents. [Accessed: July. 17, 2019]

[83] Møller Martin Fodslette, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Networks*, vol. 6, no. 4, pp. 525–533, 1993, [Accessed: July. 17, 2019].

[84] T. May, S. Van De Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 1, pp. 1–13, 2011. [Online]. Available: https://ieeexplore.ieee.org/document/5406118. [Accessed: May. 31, 2019]

[85] J. Woodruff and D. Wang, "Binaural localization of multiple sources in reverberant and noisy environments." *IEEE Transactions on Acoustic Speech and Signal Processing*, vol. 20, no. July, pp. 1503–1512, 2012. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6129395. [Accessed: Dec. 14, 2016]

[86] T. May, N. Ma, and G. J. Brown, "Robust localisation of multiple speakers exploiting head movements and multi-conditional training of binaural cues," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings*, vol. 2015-August, pp. 2679–2683, 2015, [Accessed: May. 31, 2019].

[87] N. Ma, (n.d.), "An Efficient Implementation of Gammatone Filters." [Online]. Available: http://staffwww.dcs.shef.ac.uk/people/N.Ma/resources/gammatone/ [Accessed: July 23 2019].

[88] J. O. Smith, (2011), "The Bark Frequency Scale." [Online]. Available: https://ccrma.stanford.edu/~jos/sasp/Bark_Frequency_Scale.html [Accessed: July 23 2019].

[89] J. Backman and M. Karjalainen, "Modelling of human directional and spatial hearing using neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing 1993*, vol. 1, pp. I–125–I–128. [Online]. Avail-

able: http://www.scopus.com/inward/record.url?eid=2-s2.0-0027228819&partnerID=40&md5=a56358b5e131e8a58ce164449e92e0f1 [Accessed: May. 31, 2019].

[90] L. A. Jeffress, "A Place Theory of Sound Localization," *Journal of Comparative and Physiological Psychology*, vol. 41, no. 1, pp. 35–39, 1948. [Online]. Available: https://psycnet.apa.org/record/1948-02487-001. [Accessed: May. 31, 2019]

[91] S. A. Shamma, N. Shen, and P. Gopalaswamy, "Stereausis: Binaural processing without neural delays," *Journal of the Acoustical Society of America*, vol. 86, no. 3, pp. 989–1006, 1989. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/86/3/10.1121/1.398734. [Accessed: Oct. 12, 2016]

[92] R. S. Chadwick, W. J. Wilbur, K. A. Morrish, and J. Rinzel, "A biophysical model of cochlear processing: Intensity dependence of pure tone responses," *The Journal of the Acoustical Society of America*, vol. 80, no. 1, pp. 133–145, 1986. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/3745659. [Accessed: Oct. 12. 2016]

[93] B. C. J. Moore, R. W. Peters, and B. R. Glasberg, "Auditory filter shapes at low center frequencies," *Journal of the Acoustical Society of Americal*, vol. 88, no. 1, pp. 132–140, 1990. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/88/1/10.1121/1.399960. [Accessed: Oct. 11, 2016]

[94] C.-H. Jeong, J. Brunskog, and F. Jacobsen, "Room acoustic transition time based on reflection overlap." *The Journal of the Acoustical Society of America*, vol. 127, no. 5, pp. 2733–2736, 2010. [Online]. Available: http://orbit.dtu.dk/files/4466140/Jeong.pdf. [Accessed: Oct. 20, 2016]

[95] F. Meyer, "Topographic Distance and Watershed Lines," *Mathematical Morphology and its Applications to Signal Processing*, vol. 38, pp. 113–125, 1994. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0165168494900604?via%3Dihub. [Accessed: July 23 2019]

[96] Y. E. Baba, A. Walther, and A. P. Habets, "Time of Arrival Disambiguation using the Linear Radon Transform," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2017*, pp. 106–110. [Online]. Available: https://ieeexplore.ieee.org/document/7952127/ [Accessed: May. 31, 2019].

[97] J. Radon, "On the Determination of Functions from their Integral Values along Certain Manifolds," *IEEE TRANSACTIONS ON MEDICAL IMAGING*, vol. 11, no. 4,

pp. 285–286, 1986. [Online]. Available: https://ieeexplore.ieee.org/document/4307775. [Accessed: July 23 2019]

[98] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 34–43, 2007. [Online]. Available: https://ieeexplore.ieee.org/document/4032783. [Accessed: May. 31, 2019]

[99] L. Remaggi, "Acoustic Reflector Localisation for Blind Source Separation and Spatial Audio," PhD Thesis, University of Surrey, 2017.

[100] M. Kuster, "Reliability of estimating the room volume from a single room impulse response," *The Journal of the Acoustical Society of America*, vol. 124, no. 2, p. 982, 2008. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/124/2/10.1121/1.2940585. [Accessed: May. 31, 2019]

[101] G. Defrance, L. Daudet, and J.-D. Polack, "Detecting arrivals within room impulse responses using matching pursuit," in *Proceedings of the 11th International Conference on Digital Audio Effects (DAFx-08) 2008*, pp. 1–4. [Online]. Available: http://legacy.spa.aalto.fi/dafx08/papers/dafx08_51.pdf [Accessed: May. 31, 2019].

[102] H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978. [Online]. Available: https://ieeexplore.ieee.org/document/1163055. [Accessed: July 23 2019]

[103] G. Kearney, C. Masterson, S. Adams, and F. Boland, "Dynamic Time Warping for Acoustic Response Interpolation : Possibilities and Limitations," in *17th European Signal Processing Conference (EUSIPCO) 2009*, Glasgow, Scotland, pp. 705–709. [Online]. Available: http://www.eurasip.org/Proceedings/Eusipco/Eusipco2009/contents/papers/1569193060.pdf [Accessed: May. 31, 2019].

[104] J. E. Gentle, *Matrix Algebra: Theory, Computations, and Applications in Statistics*. Springer Science+Business Media, 2007.

[105] W. S. Torgerson, "Multidimensional Scaling: I. Theory and Method," *Psychometrika*, vol. 17, no. 4, pp. 401–419, 1952. [Online]. Available: https://link.springer.com/article/10.1007/BF02288916. [Accessed: Aug. 7, 2019]

[106] D. Arteaga, D. Gracia-Garzon, T. Mateos, and J. Usher, "Scene inference

from audio," in *AES Convention 134 2013*, pp. 1–10. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=16794 [Accessed: May. 31, 2019].

[107] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by Simulated Annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983. [Online]. Available: https://pdfs.semanticscholar.org/e893/4a942f06ee91940ab57732953ec6a24b3f00.pdf. [Accessed: Sept. 11 2019]

[108] L. Chambers, *Practical Handbook of Genetic Algorithms*. Chapman and Hall/CRC, 1995.

[109] F. Ribeiro, D. Florêncio, D. Ba, and C. Zhang, "Geometrically constrained room modeling with compact microphone arrays," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 5, pp. 1421–1432, 2012, [Accessed: May. 31, 2019].

[110] S. Tervo and T. Tossavainen, "3D Room Geometry Estimation from Measured Impulse Responses," in *IEEE International Conference on Acoustics Speech and Signal Processing 2012*, pp. 513–516. [Online]. Available: https://ieeexplore.ieee.org/document/6287929 [Accessed: May. 31, 2019].

[111] K. Levenberg, "A Method for the Solution of Certain Non-Linear Problems in Least Squares," *Quarterly of Applied Mathematics*, vol. 1944, no. 2, pp. 164–168, 1944. [Online]. Available: https://www.ams.org/journals/qam/1944-02-02/S0033-569X-1944-10666-0/. [Accessed: Aug. 7, 2019]

[112] P. C. Mahalanobis, "On the generalized distance in statistics," *Proceedings of the National Institute of Sciences of India*, vol. 2, pp. 49–55, 1936.

[113] E. Weisstein, (2019), "Dihedral Angle." [Online]. Available: http://mathworld.wolfram.com/DihedralAngle.html [Accessed: May. 31, 2019].

[114] F. Antonacci, A. Sarti, and S. Tubaro, "Geometric reconstruction of the environment from its response to multiple acoustic emissions," *Acoustics Speech and Signal ...*, no. d, pp. 2822–2825, 2010. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5496186. [Accessed: Aug. 7, 2019]

[115] F. Antonacci, J. Filos, M. R. P. Thomas, E. a. P. Habets, A. Sarti, P. a. Naylor, and S. Tubaro, "Inference of room geometry from acoustic impulse responses," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 10, pp.

2683–2695, 2012. [Online]. Available: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6255766. [Accessed: May. 31, 2019]

[116] E. Nastasia, F. Antonacci, A. Sarti, and S. Tubaro, "Localization of planar acoustic reflectors through emission of controlled stimuli," *European Signal Processing Conference*, no. Eusipco, pp. 156–160, 2011. [Online]. Available: https://ieeexplore.ieee.org/document/7074175. [Accessed: May. 31, 2019]

[117] J. Filos, A. Canclini, F. Antonacci, A. Sarti, and P. A. Naylor, "Localization of Planar Acoustic Reflectors from the Combination of Linear Estimates," *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, no. Eusipco, pp. 1019–1023, 2012. [Online]. Available: https://ieeexplore.ieee.org/document/6334299. [Accessed: May. 31, 2019]

[118] M. Kuster, D. de Vries, E. M. Hulsebos, and A. Gisolf, "Acoustic imaging in enclosed spaces: Analysis of room geometry modifications on the impulse response," *The Journal of the Acoustical Society of America*, vol. 116, no. 4, p. 2126, 2004. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/116/4/10.1121/1.1785591. [Accessed: Oct. 17, 2016]

[119] L. Zamaninezhad, P. Annibale, and R. Rabenstein, "Localization of environmental reflectors from a single measured transfer function," *ISCCSP 6th International Symposium on Communications, Control and Signal Processing, Proceedings*, pp. 157–160, 2014. [Online]. Available: https://ieeexplore.ieee.org/document/6877839. [Accessed: May. 31, 2019]

[120] J. Filos, A. P. Habets, and P. A. Naylor, "A Two-Step Approach to Blindly Infer Room Geometries," *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2010. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.707.7889. [Accessed: Aug. 7, 2019]

[121] Y. El Baba, A. Walther, and E. A. Habets, "Reflector localization based on multiple reflection points," *European Signal Processing Conference*, vol. 2016-Novem, pp. 1458–1462, 2016, [Accessed: May. 31, 2019].

[122] *Acoustics – Measurements of room acoustic parameters Part 1: Performance Spaces (ISO 3382-1:2009)*, British Standards Institute Std., 2009.

[123] G. Kearney, (2016), "SADIE Binaural Measurements," Spatial Audio for Domestic

Interactive Entertainment. [Online]. Available: https://www.york.ac.uk/sadie-project/database_old.html [Accessed: May. 31, 2019].

[124] V. Pulkki, M. Karjalainen, and J. Huopaniemi, "Analyzing Virtual Sound Source Attributes Using Binaural Auditory Model," *J. Audio Eng. Soc*, vol. 47, no. 4, pp. 203 – 217, 1999. [Online]. Available: http://lib.tkk.fi/Diss/2001/isbn9512255324/article5.pdf. [Accessed: Oct. 11, 2016]

[125] J. Woodruff and D. Wang, "Sequential organization of speech in reverberant environments by integrating monaural grouping and binaural localization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 7, pp. 1856–1866, 2010, [Accessed: May. 31, 2019].

[126] J. C. Middlebrooks and D. M. Green, "Sound Localization By Human Listeners," *Annual Review of Pyschology*, vol. 42, no. February 1991, pp. 135–159, 1991. [Online]. Available: http://www.annualreviews.org/doi/abs/10.1146/annurev.ps.42.020191.001031. [Accessed: May. 31, 2019]

[127] M. Slaney, (1998), "Auditory Toolbox," Palo Alto, CA. [Online]. Available: https://engineering.purdue.edu/~malcolm/interval/1998-010/ [Accessed: May. 31, 2019].

[128] B. Gao, (2014), "Cochleagram and IS-NMF2D for Blind Source Separation." [Online]. Available: http://uk.mathworks.com/matlabcentral/fileexchange/48622-cochleagram-and-is-nmf2d-for-blind-source-separation?focused=3855900&tab=function [Accessed: May. 31, 2019].

[129] B. Rafaely and A. Avni, "Interaural cross correlation in a sound field represented by spherical harmonics," *The Journal of the Acoustical Society of America*, vol. 127, no. 2, p. 823, 2010. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/127/2/10.1121/1.3278605. [Accessed: July. 04, 2019]

[130] E. Weisstein, (2019), "Cross-Correlation." [Online]. Available: http://mathworld.wolfram.com/Cross-Correlation.html [Accessed: July. 04, 2019].

[131] M. S. Shanker, M. Y. Hu, and M. S. Hung, "Effect of data standardization on neural network training," *Omega*, vol. 24, no. 4, pp. 385–397, 1996, [Accessed: Jan. 15, 2020].

[132] Google, 2019, "TensorFlow," Google, https://www.tensorflow.org/. [Online]. Available: https://www.tensorflow.org/ [Accessed: Oct. 26, 2016].

[133] MathWorks, (2019), "cascadeforwardnet." [Online]. Available: https://uk.mathworks.com/help/deeplearning/ref/cascadeforwardnet.html

[134] S. E. Fahlman and C. Lebiere, "The Cascade-Correlation Learning Architecture," *Advances in neural information processing systems 2*, pp. 524–532, 1990, [Accessed: May. 31, 2019].

[135] Y. A. LeCun, L. Bottou, G. B. Orr, and K. R. Müller, "Efficient BackProp," in *Neural Networks: Tricks of the Trade*. Springer Berlin Heidelberg, 1998, vol. Lecture Notes, ch. 1, pp. 9–50, [Accessed: May. 31, 2019].

[136] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations 2015*, pp. 1–15. [Online]. Available: http://arxiv.org/abs/1412.6980 [Accessed: May. 31, 2019].

[137] Google, (2019), "tf.nn.softmax_cross_entropy_with_logits." [Online]. Available: https://www.tensorflow.org/api_docs/python/tf/nn/softmax_cross_entropy_with_logits [Accessed: July. 05, 2019].

[138] Equator Audio, 2016, "Equator D5 Coaxial Loudpseakers," Equator Audio, http://www.equatoraudio.com. [Online]. Available: http://www.equatoraudio.com/New-Improved-D5-Studio-Monitors-Pair-p/d5.htm [Accessed: Oct. 25, 2016].

[139] Genelec, (n.d.).

[140] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," *Proc. AES 108th conv, Paris, France*, 2000. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=10211. [Accessed: May. 31, 2019]

[141] F. Stevens and D. Murphy, "Spatial impulse response measurement in an urban environment," in *Presented at the 55th AES International Conference on Spatial Audio 2014*, Helsinki, Finland, pp. 1–8. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=17355 [Accessed: June. 18, 2019].

[142] MathWorks, 2006, "kurkalwallis." [Online]. Available: https://uk.mathworks.com/help/stats/kruskalwallis.html [Accessed: Aug. 28, 2019].

[143] B. Rafaely, B. Weiss, and E. Bachmat, "Spatial aliasing in spherical microphone arrays," *IEEE Transactions on Signal Processing*, vol. 55, no. 3, pp. 1003–1010, 2007, [Accessed: May. 31, 2019].

[144] M. Acoustics, (n.d.), "Eigenmike Microphone," MH Acoustics,

http://www.mhacoustics.com/. [Online]. Available: http://www.mhacoustics.com/products [Accessed: Sept. 21, 2016].

[145] N. Epain and C. T. Jin, "Spherical Harmonic Signal Covariance and Sound Field Diffuseness," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1796–1807, 2016. [Online]. Available: https://ieeexplore.ieee.org/document/7501622. [Accessed: May. 31, 2019]

[146] B. N. Gover, J. G. Ryan, and M. R. Stinson, "Microphone array measurement system for analysis of directional and spatial variations of sound fields," *The Journal of the Acoustical Society of America*, vol. 112, no. 5, pp. 1980–1991, 2002. [Online]. Available: https://asa.scitation.org/doi/10.1121/1.1508782. [Accessed: Jan. 15, 2020]

[147] A. Politis, (2015), "getSH," Espoo, Finland. [Online]. Available: https://github.com/polarch/Spherical-Harmonic-Transform/blob/master/getSH.m [Accessed: May. 31, 2019].

[148] MathWorks, "mat2gray.m." [Online]. Available: https://uk.mathworks.com/help/images/ref/mat2gray.html [Accessed: Sept. 11 2019].

[149] MathWorks, (2017), "Watershed." [Online]. Available: https://uk.mathworks.com/help/images/ref/watershed.html [Accessed: May. 31, 2019].

[150] S. Eddins, 2002, "The Watershed Transform: Strategies for Image Segmentation." [Online]. Available: https://uk.mathworks.com/company/newsletters/articles/the-watershed-transform-strategies-for-image-segmentation.html [Accessed: May. 31, 2019].

[151] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*. Berlin, Heidelberg: MIT Press, 2009, pp. 603–611.

[152] CATT-Acoustics, *CATT-A V9.1 User's Manual*. Gothenburg Sweden: CATT-Acoustics, 2016.

[153] A. Krokstad, S. Strøm, and S. Sørsdal, "Calculating the Acoustical Room Response by The Use of Ray Tracing Technique," *Journal of Sound and Vibration*, vol. 8, pp. 118–125, 1968, [Accessed: May. 31, 2019].

[154] R. Stafford, 2016, "How can I plot a 3D plane knowing its center point coordinates and its Normal." [Online]. Available: https://uk.mathworks.com/matlabcentral/answers/

291485-how-can-i-plot-a-3d-plane-knowing-its-center-point-coordinates-and-its-normal [Accessed: Aug. 28, 2019].

[155] E. Weisstein, (2019), "Plane-Plane Intersection." [Online]. Available: http://mathworld.wolfram.com/Plane-PlaneIntersection.html [Accessed: July. 02, 2019].

[156] N. Khaeld, 2011, "plane_intersect." [Online]. Available: https://uk.mathworks.com/matlabcentral/fileexchange/17618-plane-intersection [Accessed: June. 18, 2019].

[157] E. Weisstein, (2019), "Line-Plane Intersection." [Online]. Available: http://mathworld.wolfram.com/Line-PlaneIntersection.html [Accessed: July. 01, 2019].

[158] A. Southern, "shoebox_response.m," Espoo, Finland.

[159] MH Acoustics, (2016), "Eigenbeam Datasheet," p. 12. [Online]. Available: https://mhacoustics.com/sites/default/files/EigenbeamDatasheet_R01A.pdf [Accessed: May. 31, 2019].

[160] E. Sengpiel, (2014), "Speed of sound in Humid Air." [Online]. Available: http://www.sengpielaudio.com/calculator-airpressure.htm [Accessed: Jan. 20, 2020].

[161] G. Naylor and J. H. Rindel, "Predicting room acoustical behaviour with the ODEON computer model," *The Journal of the Acoustical Society of America*, vol. 92, no. 4, p. 2346, 1992. [Online]. Available: https://asa.scitation.org/doi/10.1121/1.404931. [Accessed: May. 31, 2019]

[162] B. Efron, "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, vol. 7, no. 1, 1979. [Online]. Available: https://projecteuclid.org/euclid.aos/1176344552. [Accessed: Aug. 28, 2019]

[163] MathWorks, (2006), "bootstrapci." [Online]. Available: https://uk.mathworks.com/help/stats/bootci.html [Accessed: Aug. 28, 2019].

[164] H. Malik, "Acoustic environment identification and its applications to audio forensics," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 11, pp. 1827–1837, 2013. [Online]. Available: https://ieeexplore.ieee.org/document/6595031. [Accessed: Sept. 10, 2019]

[165] NumFocus, (2019), "NumPy." [Online]. Available: http://www.numpy.org/ [Accessed: Sept. 10, 2019].

[166] SciPy, (2019), “SciPy.” [Online]. Available: https://www.scipy.org/ [Accessed: Sept. 10, 2019].