Improvements in the Perceived Quality of Streaming and Binaural Rendering of Ambisonics

Tomasz Rudzki

Doctor of Philosophy

University of York

Physics, Engineering and Technology

May 2023

Abstract

With the increasing popularity of spatial audio content streaming and interactive binaural audio rendering, it is pertinent to study the quality of the critical components of such systems. This includes low-bitrate compression of Ambisonic scenes and binaural rendering schemes. This thesis presents a group of perceptual experiments focusing on these two elements of the Ambisonic delivery chain.

The first group of experiments focused on the quality of low-bitrate compression of Ambisonics. The first study evaluated the perceived timbral quality degradation introduced by the Opus audio codec at different bitrate settings and Ambisonic orders. This experiment was conducted using multi-loudspeaker reproduction as well as binaural rendering. The second study has been dedicated to auditory localisation performance in bitrate-compressed Ambisonic scenes reproduced over loudspeakers and binaurally using generic and individually measured HRTF sets. Finally, the third study extended the evaluated set of codec parameters by testing different channel mappings and various audio stimuli contexts. This study was conducted in VR thanks to a purposely developed listening test framework. The comprehensive evaluation of the Opus codec led to a set of recommendations regarding optimal codec parameters.

The second group of experiments focused on the evaluation of different methods for binaural rendering of Ambisonics. The first study in this group focused on the implementation of the established methods for designing Ambisonic-to-binaural filters and subsequent objective and subjective evaluations of these. The second study explored the concept of hybrid binaural rendering combining anechoic filters with reverberant ones. Finally, addressing the problem of non-individual HRTFs used for spatial audio rendering, an XR-based method for acquiring individual HRTFs using a single loudspeaker has been proposed.

The conducted perceptual evaluations identified key areas where the Ambisonic delivery chain could be improved to provide a more satisfactory user experience.

Contents

List of Figures

List of Tables

Acknowledgements

This research was made possible through the funding provided by Google and the School of Physics, Engineering and Technology, University of York.

I am deeply grateful to Prof. Gavin Kearney, who trusted me with this project, guided me through the whole journey and has always been exceptionally supportive and patient. I am also grateful to Prof. Helena Daffern for providing support and invaluable feedback on my work.

I am deeply thankful to Dr Jan Skoglund for his support.

I would like to thank AudioLab's research associates Jess Stubbs and Benjamin Lee for helping me carry out this research.

I would like to thank my colleagues at the AudioLab, who were always there for me.

I would like to thank my family for their endless support, especially my wife Nhung.

This thesis is dedicated to my long-time mentor, Dr Tadeusz Fidecki, who inspired me to pursue the audio research path until I became an "audio" doctor myself.

Declaration

I declare that this thesis is a presentation of original work and I am the sole author. All sources are acknowledged as references. This work has not previously been presented for an award at this, or any other, University.

In addition, I declare that parts of this research have been presented in the following publications:

- T. Rudzki, D. Murphy, and G. Kearney. A DAW-Based Interactive Tool for Perceptual Spatial Audio Evaluation. In *Audio Engineering Society Convention 145*. Audio Engineering Society, 2018
- T. Rudzki, I. Gomez-Lanzaco, P. Hening, J. Skoglund, T. McKenzie, J. Stubbs, D. Murphy, and G. Kearney. Perceptual Evaluation of Bitrate Compressed Ambisonic Scenes in Loudspeaker Based Reproduction. In *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society, 2019b
- T. Rudzki, I. Gomez-Lanzaco, J. Stubbs, J. Skoglund, D. T. Murphy, and G. Kearney. Auditory Localization in Low-Bitrate Compressed Ambisonic Scenes. *Applied Sciences*, 9(13):2618, 2019c
- T. Rudzki, C. Earnshaw, D. Murphy, and G. Kearney. SALTE Pt. 2: On the Design of the SALTE Audio Rendering Engine for Spatial Audio Listening Tests in VR. In *Audio Engineering Society Convention 147*. Audio Engineering Society, 2019a
- T. Rudzki, D. Murphy, and G. Kearney. XR-based HRTF Measurements. In *Audio Engineering Society Conference: AES 2022 International Audio for Virtual and Augmented Reality Conference*. Audio Engineering Society, 2022
- T. Rudzki, D. Murphy, and G. Kearney. User Preference Evaluation of Direct-to-Reverberant Ratio of Virtual Ambisonic Listening Spaces. In *Audio Engineering Society Conference: 2023 AES International Conference on Spatial and Immersive Audio*. Audio Engineering Society, 2023
- B. Lee, T. Rudzki, J. Skoglund, and G. Kearney. Context-Based Evaluation of the Opus Audio Codec for Spatial Audio Content in Virtual Reality. *Journal of the Audio Engineering Society*, 71(4):145–154, 2023

Chapter 1

Introduction

Spatial audio, also referred to as *immersive audio*, is gaining a substantial presence in the landscape of digital media content and services. It is used for music, film, gaming, podcasting, and virtual and augmented reality applications. With the rise of 5G mobile networks, spatial audio is expected to become an essential element of communication services in the future.

Spatial audio systems are expected to give the sensation of being in another acoustical space or to render virtual sound sources realistically or plausibly in the user's acoustic environment. To create such a convincing auditory experience, the sound reaching the ears needs to be crafted in a very specific way, which is in accordance with how the brain interprets the sounds of the real world. This can be achieved under certain technical conditions, including audio signal representation, perceptual coding, and rendering methods. Nevertheless, providing such conditions for a regular user is non-trivial and requires a balance between practicality and employed resources. This research focuses on finding this balance and proposing recommendations and methods aimed at making high-quality spatial audio within the reach of a regular user.

Over the time this research was carried out, the consumer electronics industry made a massive leap in the adoption of spatial audio. Interactive binaural audio rendering has been widely adopted to deliver immersive audio over headphones, making spatial audio more accessible. This includes the introduction of wireless headphones and earbuds with built-in sensors. There has also been a shift in the adoption of immersive audio formats. In a recent survey by Production Expert[1], one-third of post-production and music mixers responded that they work in the Dolby Atmos format, which is a three-dimensional extension of the established surround sound formats. According to professionals whose comments have been published alongside the survey results, the increasing adoption of Atmos, especially for mixing music content, is driven by a mixture of creative, business, and marketing goals.

In addition to commercial formats, a royalty-free audio technique known as Ambisonics (Gerzon, 1973; Zotter and Frank, 2019) is constantly growing its user base organically. The adoption of Ambisonics is mainly driven by sound artists,

[1]Atmos survey – `https://www.pro-tools-expert.com/production-expert-1/who-is-mixing-in-atmos-we-have-the-survey-results/`

academic researchers and XR developers. It is used to deliver immersive audio assets in games (Deleflie and Goodwin, 2007) as well as spatial music, e.g. Ambisonic Music Library[2]. Ambisonics has also become a standard for 360° video productions. Some binaural renderers use it as an input format or an intermediate sound bed, which is then decoded to the left and right ear headphone signals, e.g. Google Resonance Audio[3]. The wide range of Ambisonic production software tools, often free and open source, makes the ecosystem accessible. There is also a wide range of Ambisonic microphones available on the market.

Streaming of Ambisonic audio is usually accompanied by 180° or 360° video streams and delivered using platforms like YouTube VR[4] or HOAST[5] which can also be used as an open source library to deploy custom content. Currently, YouTube VR supports 1st-order Ambisonic audio and non-diegetic stereo tracks for its immersive audio and video streaming, while HOAST supports Ambisonic audio up to 4th-order alongside 360° videos (Deppisch et al., 2020). Moreover, Ambisonics is at the centre of the Immersive Audio Model and Formats specification[6] developed by the Alliance for Open Media[7].

The established audio quality of Ambisonic streaming is not optimal. The community of spatial audio creators has raised their concerns that streaming services like YouTube do not support higher-order Ambisonics[8], consequently degrading the audio experience. Ambisonic streaming requires efficient bitrate compression algorithms. In recent years, advancements have been proposed in the field of rendering Ambisonic signals for headphone playback, which have not been comprehensively evaluated against the established systems. As the individual ear shape plays a role in the perception of sound, it is desirable to research the binaural reproduction of Ambisonics using individual binaural filters. With the rising popularity of immersive audio, it is pertinent to research these elements.

The Ambisonic delivery chain generally consists of multichannel signals carrying audio scene representation that are compressed using perceptual coding to be streamed over the network. On the receiving end, the signals are then rendered typically to headphones. Figure 1.1 shows a simple schematic of the Ambisonic delivery chain. The following technical aspects may affect the perceived quality of Ambisonic audio:

- Ambisonic signal truncation order;
- Low-bitrate compression of Ambisonic signals;
- Filter design methods used for the binaural rendering of Ambisonics;
- Anechoic or reverberant binaural rendering;

[2]Ambisonic Music Library – `http://ambisonicmusiclibrary.com/`

[3]Resonance Audio – `https://resonance-audio.github.io/resonance-audio/`

[4]YouTube VR – `https://vr.youtube.com/`

[5]HOAST – `https://hoast.iem.at/`

[6]IAMF – `https://aomediacodec.github.io/iamf/`

[7]Alliance for Open Media – `https://aomedia.org/`

[8]Spatial Audio in VR/AR/MR Facebook group discussion - `https://www.facebook.com/groups/SpatialAudioVRARMR/permalink/2959851487491492/`

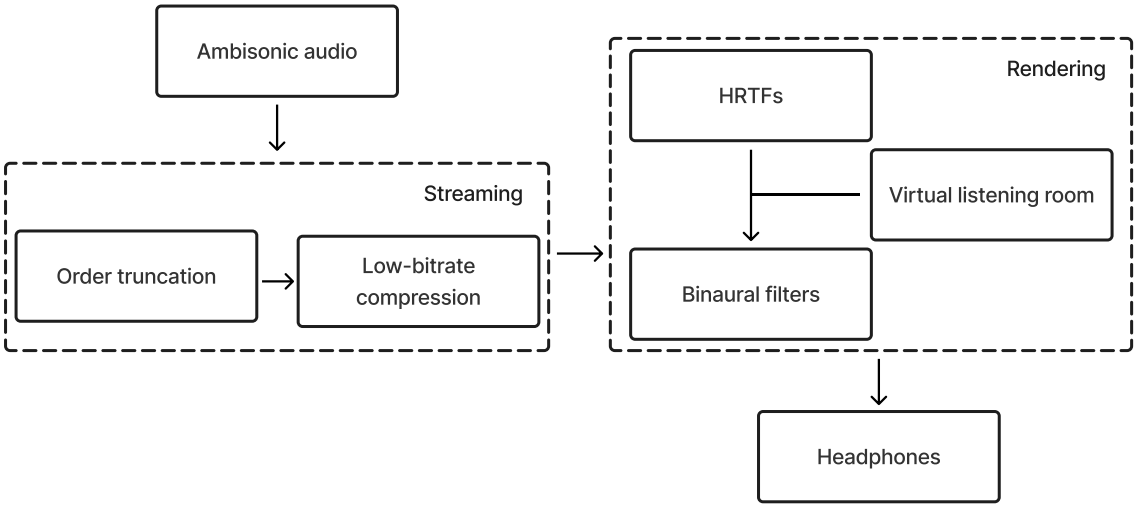


Figure 1.1: Ambisonic delivery chain.

- Generic or individual binaural filters (HRTFs).

1.1 Statement of Hypothesis

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

Streaming and rendering of Ambisonics can be improved through perceptual evaluation and optimisation of the Ambisonic delivery chain.

1.2 Novel Contributions

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

- **Perceptual evaluation of low-bitrate compression of Ambisonics.** A group of perceptual experiments focusing on the quality of low-bitrate compression of Ambisonics and different spatial audio reproduction methods. As a result, optimal codec parameters were obtained for streaming Ambisonic audio under different rendering and context conditions.
- **Evaluation of methods for the binaural rendering of Ambisonics.** The results of the experiment have pointed towards the optimal binaural filter design method for rendering Ambisonic scenes of different orders.
- **User-preference evaluation of virtual Ambisonic listening spaces.** This experiment evaluated a hybrid approach to rendering Ambisonics which combines anechoic and reverberant filters. The results showed that such a

method could provide an alternative to the standard anechoic rendering. Based on user responses, preferred direct-to-reverberant sound ratios were established.

- **XR-based system for individual HRTF measurements.**
  Finally, to address the problem of non-individual HRTFs used for spatial audio rendering, an XR-based method to acquire individual HRTFs using a single loudspeaker has been proposed.

1.3 Thesis Structure

This thesis is structured as follows. First, a review of research on the perception of sound is presented in Chapter 2. This chapter discusses the basic physics of sound, the human auditory system and spatial hearing. Chapter 3 discusses sound reproduction methods for spatial audio, established formats and perceptual coding schemes, followed by a review of the literature on perceptual evaluation of spatial audio systems. This includes established listening test paradigms, problems associated with subjective assessment of spatial audio quality and how different perceptual attributes can be systematised in this context.

Chapter 4 presents a study focused on the evaluation of perceived timbral quality degradation introduced by the Opus audio codec at different bitrate settings and Ambisonic orders. This experiment was conducted using multi-loudspeaker reproduction as well as binaural rendering using generic and individually measured HRTF sets. Chapter 5 presents a study focused on auditory localisation performance within the presented bitrate-compressed Ambisonic scenes. The impact of the employed reproduction method on the collected responses was also investigated, as the scenes were reproduced over loudspeakers and binaurally using generic and individually measured HRTF sets. Chapter 6 extends the evaluated set of codec parameters by testing different channel mappings and various audio stimuli contexts. This study also moves the user interface into VR thanks to a purposely developed listening test framework.

Chapter 7 focuses on the implementation of established as well as alternative methods for the binaural rendering of Ambisonics. The chapter also presents subsequent objective and subjective evaluations of these. Chapter 8 presents an experiment exploring user preferences of the direct-to-reverberant sound ratio (DRR) of virtual Ambisonic listening spaces in relation to different types of reverberation and different Ambisonic audio content. Chapter 9 discusses a Head Related Transfer Function (HRTF) measurement system that uses minimal hardware configuration.

Chapter 10 concludes this thesis. A summary of the findings is presented. Appendix A and Appendix B present listening test tools developed as part of the research presented in this thesis.

Chapter 2

Auditory Perception

Auditory perception is at the heart of this dissertation. The experiments presented in the subsequent chapters are based on controlled presentations of auditory stimuli rendered using loudspeakers or headphones. This chapter discusses the relevant principles of sound and its perception. Section 2.1 discusses the physics of sound propagation. It is followed by Section 2.2, which introduces basic information on the human auditory system. Section 2.3 further extends this chapter by reviewing the literature on spatial hearing and introducing the psychoacoustic concepts referred to in this dissertation.

2.1 The Physics of Sound

2.1.1 Sound Propagation

Sound in a physical context can be broadly defined as a mechanical disturbance of an elastic medium. However, a more accurate description of sound focuses on the propagation of this disturbance in a medium, which typically consists of air, other gas, solid or liquid. Howard and Angus (2013) propose a simple one-dimensional model of a sound-propagating medium using an analogy to golf balls connected by springs. The golf balls represent masses of molecules, whereas the connecting springs represent the forces between them. Once the first golf ball is pushed, the adjacent spring is compressed and pushes the neighbouring golf ball, as illustrated in Figure 2.1. This movement corresponds to the way sound propagates in an unconstrained medium and is known as the *longitudinal wave*. The region where molecules are pushed together is called *compression*, where they are being pulled apart - *rarefaction*. Compression and rarefaction are observed in the air as a momentary increase and decrease in atmospheric pressure.

Speed of Sound

The speed of sound propagation depends on the physical properties of the medium. It is higher for solids and liquids than for gases. In the case of air and other gases, it is strongly affected by the mass of its molecules and the absolute temperature.

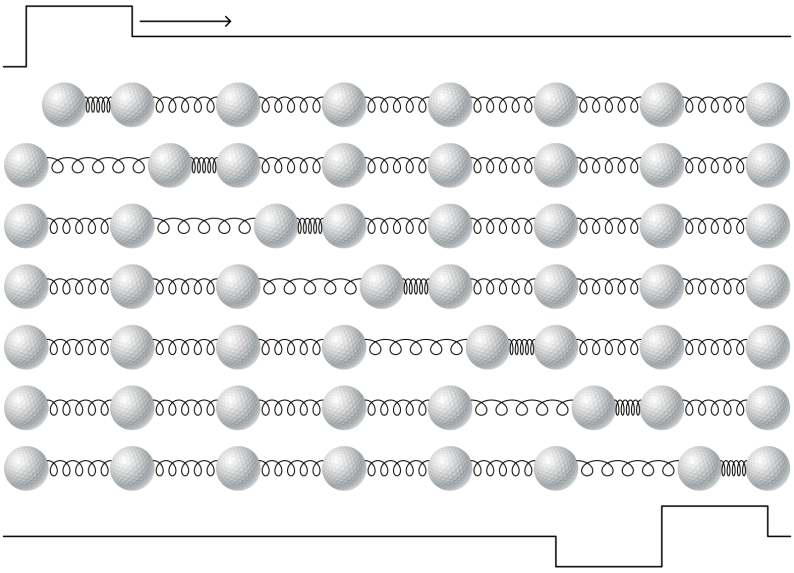


Figure 2.1: Ball-and-spring model of a propagating sound pulse. Adapted from Howard and Angus (2013).

Equation 2.1 can approximate the speed of sound c in a normal temperature range in dry air. In this equation, T_C is the temperature of air expressed in °C. The mass of molecules may slightly increase in humid air, increasing the speed of sound. For example, at a temperature of 20°C, 50% humidity and 100 kPa atmospheric pressure, the speed of sound is equal to 344 $\mathrm{m\,s^{-1}}$.

$$c = 331.3 + (0.59 \times T_C) \tag{2.1}$$

Wavelength and Frequency

If the medium is periodically excited, the distance between the regions of the same pressure is known as *wavelength*, denoted using λ. Equation 2.2 shows the relation between the speed of sound and the period T, which is the time taken by the full compression–rarefaction cycle for a single point in space. The wavelength is inversely proportional to the rate of pressure variation, known as *frequency*, denoted as f in Equation 2.3. Figure 2.2 shows a sine wave propagating in a material.

$$T = \frac{\lambda}{c} \tag{2.2}$$

$$f = \frac{1}{T} \tag{2.3}$$

Sound Pressure

The amplitude of sound can be described using either *pressure* or *particle velocity* component. Because human ears are sensitive to pressure and it is easier to measure,

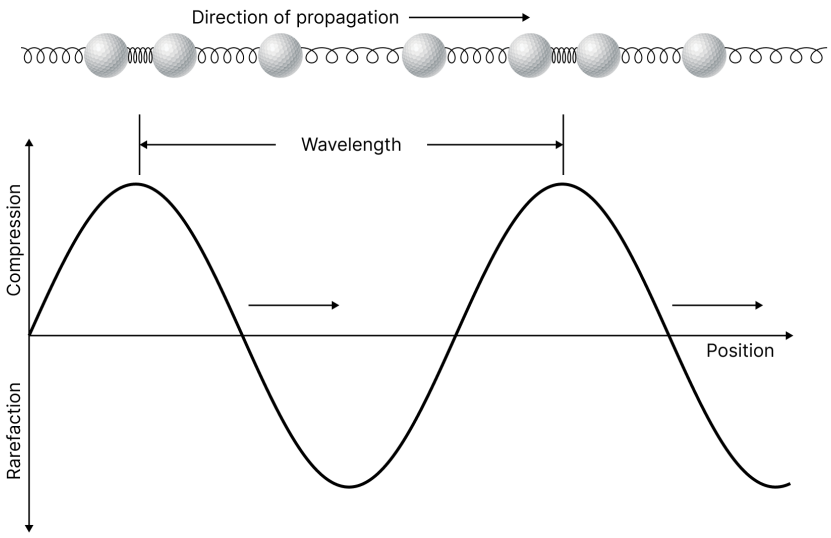


Figure 2.2: Sine wave propagating in a material and its wavelength. Adapted from Howard and Angus (2013).

it is used more commonly in psychoacoustics and audio engineering. The root mean square (RMS) value of the atmospheric pressure deviation at a particular point is therefore called the *sound pressure*. Assuming a point source, the sound pressure p changes proportionally to the inverse distance r between the source and the point in space.

$$p \propto \frac{1}{r} \tag{2.4}$$

2.2 Human Auditory System

The human ear is composed of three main structures: the outer ear, the middle ear, and the inner ear, as seen in Figure 2.3. The most external part of the outer ear is the pinna, which has a unique shape that interacts with the sound reaching the ears. The outer ear also consists of the ear canal, a narrow tube leading to the eardrum. The middle ear is an air-filled chamber located behind the eardrum. It contains three tiny bones called malleus, incus, and stapes that transmit sound vibrations from the eardrum to the inner ear. Ossicles play an essential role by matching the impedance difference between the two mediums: air in the ear canal and fluid that fills the inner ear (Van Opstal, 2016). The inner ear is a complex system of chambers and canals deep within the skull. It contains the cochlea, which is responsible for generating impulses sent to the brain in response to vibrations. The human auditory system relies on the analysis of these impulses.

Excitation of the hearing mechanism at specific amplitudes and frequencies results in the perception of an *auditory event*. The sound pressure range perceived by humans is extremely wide. It extends from about 10^{-5} Pa to 10^{2} Pa and the frequency range of perceived sounds extends from 20 Hz to 20 kHz. As human perception is

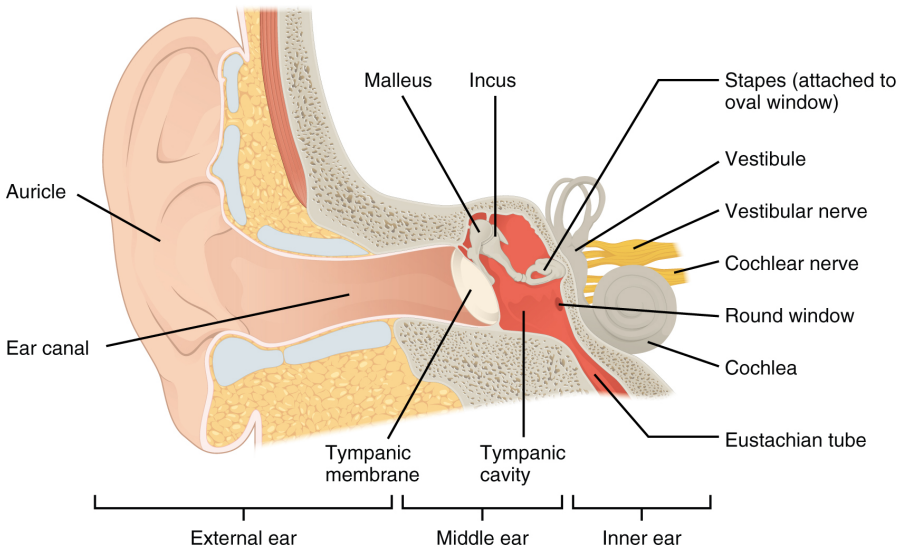


Figure 2.3: The anatomy of the human ear. Reproduced from Biga et al. (2020) under CC license.

logarithmic, i.e. a linear change in physical stimulation causes a logarithmic change in perception (Warren, 1981), it is common to express sound pressure using the *decibel scale*. The following equation defines the relation between sound pressure level (SPL) and sound pressure p:

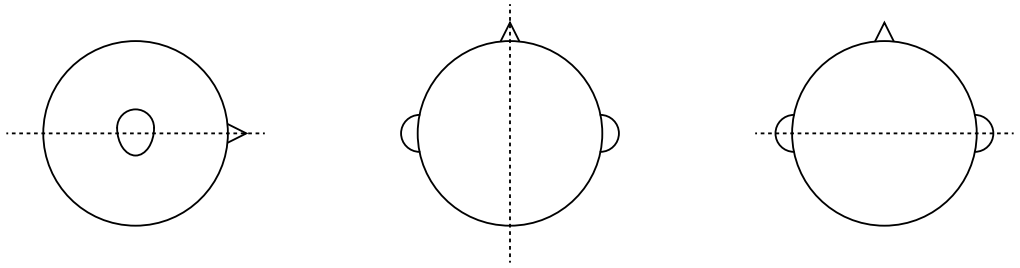
$$SPL = 20 \log_{10} \frac{p}{p_0}, \tag{2.5}$$

where p_0 is the minimum sound pressure perceived by humans equal to 20×10^{-6}Pa at a frequency of 1 kHz.

Due to the information processing capabilities of the brain, physical bodies that emit sound waves can be perceived as *auditory sources*, exhibiting the following subjective characteristics: *loudness*, *pitch*, and *timbre*. Loudness is associated primarily with the amplitude of the sound wave, pitch with its fundamental frequency, and timbre with its spectral shape. All these quantities are also affected by the temporal properties of the sound, e.g. duration. For more detailed information on the human auditory system and psychoacoustics, the reader is referred to the textbook by Zwicker and Fastl (2013).

2.3 Spatial Hearing

The auditory system is not only capable of recognising different sound sources but is also able to localise their origin in space. Interpretation and exploitation of spatial paths between the sound source and the ears can be referred to as spatial hearing. Identification of the approximate position of the sound source is possible through the set of attributes of the ear signal known as *auditory localisation cues*, which depend on the direction of incidence and distance of the source. This section provides a brief review of existing research on auditory localisation cues.



(a) Horizontal plane (b) Median plane (c) Frontal plane

Figure 2.4: The three principal planes.

2.3.1 Spatial Coordinates

To discuss the properties of spatial hearing, it is necessary to establish a standard convention to describe the position of the sound source in relation to the listener. This includes the use of three anatomical planes, commonly known as the horizontal, median, and frontal planes. Figure 2.4 shows the location of these planes. It is also common in the literature to refer to these planes as axial, sagittal, and coronal anatomical planes. Moreover, to describe the exact position of the sound source, it is typical to use the spherical coordinate system depicted in Figure 2.5. The azimuth angle describes the horizontal direction relative to the front of the listener, while the elevation angle describes the vertical direction in relation to the horizontal plane. The system used throughout this thesis uses positive azimuth angles for source positions on the left side of the listener and positive elevation angles for positions above the horizontal plane. However, the reader might encounter alternative systems used within the audio production and research fields. The most common example of such a difference is using an azimuth angle with values increasing towards the right side of the listener, as adopted by most spatial audio plugin developers, e.g. Kronlachner (2014a). The origin of the spatial coordinate system used in audio research is typically located at the intersection of the median plane and the interaural axis that runs across both ears.

2.3.2 Localisation cues

The basic property of spatial hearing is the ability to derive information about the sound source based on the differences in left and right ear signals. Venturi conducted the first known research on spatial hearing in the late 18th century (Wade and Deutsch, 2008). He showed that people can point the direction of the incoming flute sound. He associated this ability with differences in sound intensity between the ears. Lord Rayleigh (Rayleigh, 1907) conducted further research on spatial hearing and proposed the *duplex theory of localisation*, which specifies that the auditory localisation of low- and high-frequency sounds is based on the respective phase and intensity difference between the ear signals. The literature describes these differences as the *interaural localisation cues* or simply the *binaural cues*. Figure 2.6 shows a spherical model of the head in a horizontal plane and approximated paths of sound reaching both ears. Localisation cues which exist in single ear signals are called

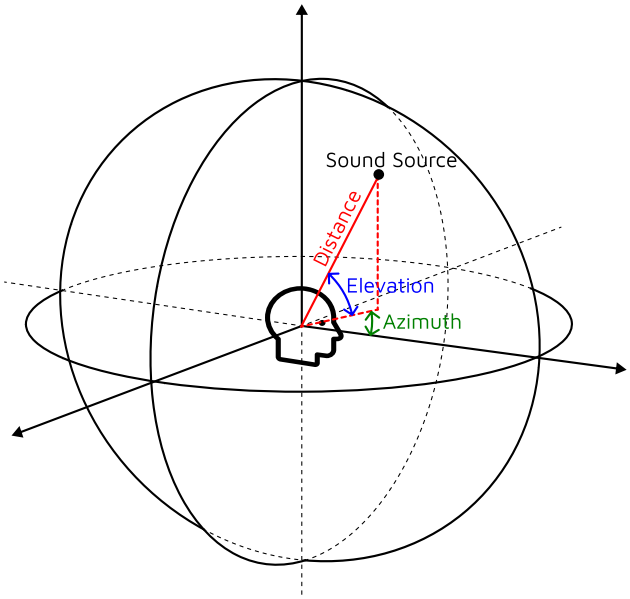


Figure 2.5: Spherical coordinate system.

monoaural signals. Localisation also depends on the distance between the listener and the sound source. Sound sources within ca. 1 m distance from the listener's head are considered to be located in the *near field*, while sources further away are considered to be in the *far field*.

Interaural Time Difference

The difference in the arrival time of a sound wave in both ears is known as the Interaural Time Difference (ITD). It increases with a displacement of the sound source from the median plane. For a medium-size head, the ITD ranges from -650 to 650 µs. It is the predominant localisation cue at frequencies below ca. 1.4 kHz, for which the auditory system is sensitive to phase difference (Mills, 1958). ITD can also be discriminated for high-frequency sounds based on their envelope, that is, temporal changes in their amplitude (Henning, 1974). However, according to Yost (2017), the envelope dependency does not contribute to the auditory localisation of real sound sources.

ITD in the horizontal plane can be approximated based on the frequency-dependent model of sound travelling around a rigid sphere (Kuhn, 1977) using the following equation:

$$ITD = \frac{ar}{c} \sin\theta, \tag{2.6}$$

where θ is the angle between the median axis and incidence direction of the sound source in a horizontal plane expressed in radians, a is a non-dimensional parameter ($a = 3$ at frequencies below 500 Hz ($\theta < 90°$) and is gradually decreasing to $a = 2$ at frequencies above 2000 Hz ($\theta < 60°$)), r is the approximate radius of the head and c is the speed of sound.

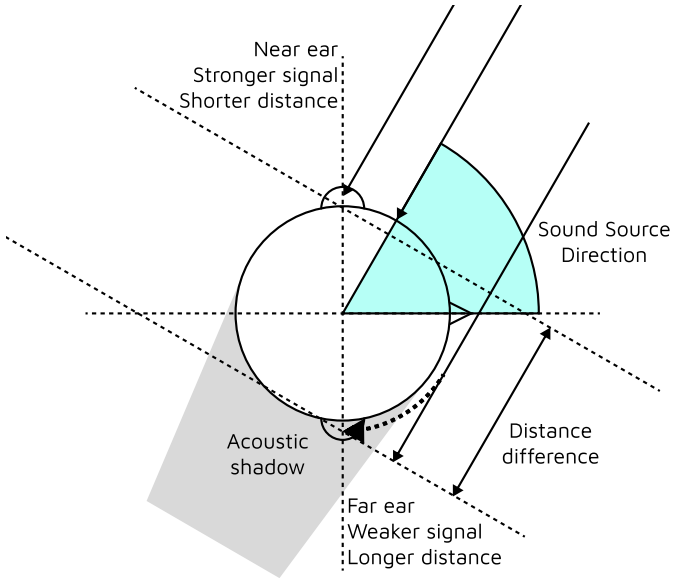


Figure 2.6: Binaural localisation cues caused by the time-of-arrival and frequency-dependent sound intensity difference between the ear signals.

Interaural Level Difference

Due to the acoustic shadowing effect of the listener's head placed in a path of an incident sound, the signals in both ears exhibit the Interaural Level Difference (ILD) (Blauert and Allen, 1997), which increases with the angle between the median plane and the source incidence. Far-field ILD is relatively negligible at low frequencies and it increases at mid and high frequencies to a maximum value of ca. 20 dB at 10 kHz. Therefore ILD dominates horizontal sound localisation at frequencies above 1.5 kHz, although human sensitivity to ILD is frequency-independent (Salminen, 2015). A substantial increase in ILD can be observed for both low and high frequencies when the source is located in the near field, as shown by Brungart and Rabinowitz (1999); Shinn-Cunningham et al. (2000).

Spectral Cues

It was observed in the 19th century that the directional sensation of sound is affected by the orientation of the sound source relative to the pinna (Thompson, 1879). Later, Batteau (1967) suggested that the pinna uniquely transforms the incoming sound according to each direction of arrival. Further studies revealed that the combined acoustic effects of the outer ear, head, and shoulders produce spectral changes at mid and high frequencies. Spectral cues are critical for vertical sound source localisation (Gardner and Gardner, 1973) and for distinguishing the direction of sounds within the *cone of confusion* (Wallach, 1939), i.e. located at the same angular distance from the interaural axis.

Head Movements

Head movements can reduce localisation confusion by changing interaural and monoaural cues, consequently delivering additional information to the auditory system. The importance of head movements in auditory localisation has initially been investigated by Young (1931), who reported that limiting head movements results in less accurate localisation. Further studies found that head movements can reduce the front-back confusion error if the duration of the sound is long enough, i.e. 600–800 ms (Łętowski and Łętowski, 2012). However, when the stimulus is short and is presented during rapid head movement, the auditory localisation can be strongly distorted (Cooper et al., 2008).

Vision and Memory Cues

When the auditory and visual cues conflict with each other and the sound source is in the person's field of view, its position is usually determined by the visual cue. This phenomenon is known as the *ventriloquism effect*. It is an example of the *capture effect* (Ghirardelli and Scharine, 2009), where information received through one sensory channel can be affected by information received in another. The listener's anticipation and memory can strongly affect auditory localisation as well. Sounds familiar to them are localised more precisely (Łętowski and Łętowski, 2012).

2.3.3 Distance Perception

The perception of sound source distance depends on a combination of cues depending on the actual distance between the listener and the source and the environment in which they are located. In the free far field, i.e. with only the direct sound path present, the sound source distance can be perceived based on the sound wave pressure changing proportionally with the inverse distance from the point source. Another cue comes from the fact that air attenuation varies with frequency. High frequencies are attenuated more for a source located further from the listener than for a nearby source. If the source is located in a region proximate to the listener, the distant-dependent ILD change serves as a distance cue (Shinn-Cunningham et al., 2000).

In environments with sound-reflecting surfaces, e.g. rooms, signals reaching the ears consist not only of direct sound but also sound reflected from the surfaces. The ratio of direct sound energy to reflected sound energy (DRR) decreases with increasing distance between the source and the listener. Bronkhorst and Houtgast (1999) showed that the perceived distance depends on this ratio. The sound spectrum of the ear signals also changes as a result of reverberation, potentially contributing to more accurate distance perception of sound sources familiar to the listener (Zahorik, 2002a).

2.3.4 Headphone Listening

As the sound reproduction discussed in this dissertation is performed primarily with headphones, it is pertinent to introduce the following two concepts explicitly related to this type of listening.

Externalisation

Sounds originating in the real world tend to be perceived as being *externalised*, that is, located a certain distance from the listener's head. However, the mismatch in ear signals between the natural and headphone listening conditions often causes a perception of sound somewhere within the listener's head. Hartmann and Wittenberg (1996) reported that externalisation depends on the accuracy of ITD at frequencies below 1 kHz and ILD at all frequencies equally under anechoic conditions, as well as the retention of the correct spectrum in each ear. According to Kulkarni and Colburn (1998), a certain level of spectral smoothing can be introduced to ear signals without degrading externalisation. However, a later study by Hassager et al. (2016) showed that the removal of spectral detail from direct sound under reverberant conditions affects perceived externalisation.

As reported by Begault et al. (2001), externalisation depends on the presence of reverberation. Another study by Catic et al. (2013, 2015) showed that temporal fluctuations in ILD and interaural coherence (IC) observed in reverberant environments are essential cues for the perception of externalisation, especially for frontal sound sources. Brimijoin et al. (2013) showed that rotational head movements affect externalisation.

Binaural Lateralisation

When the dichotic signals are provided to both ears through headphones, a lateral displacement of the phantom sound source occurs along the interaural axis. The relationship between time and level differences in ear signals and the perceived shift of the phantom sound source inside the listener's head is called *lateralisation* (Blauert and Allen, 1997). Early experiments by Jeffress and Taylor (1961) show that listeners can identify lateralised sounds perceived within the head as coming from external visual targets placed in the free space around them. The relationship between the lateral displacement of the auditory event is a function of both direction and distance. If the localisation judgment space is limited to the anterior half of the horizontal plane and the distance from the listener is fixed, then the lateral displacement will directly correspond to the direction of the phantom sound source. Figure 2.7 shows this relationship. Lateralisation is essential when considering not fully externalised binaural stimuli, e.g. due to spectral mismatch.

2.3.5 Localisation Performance

The performance of human auditory localisation varies depending on the direction of incidence of the sound wave and the distance, level, and frequency characteristics of

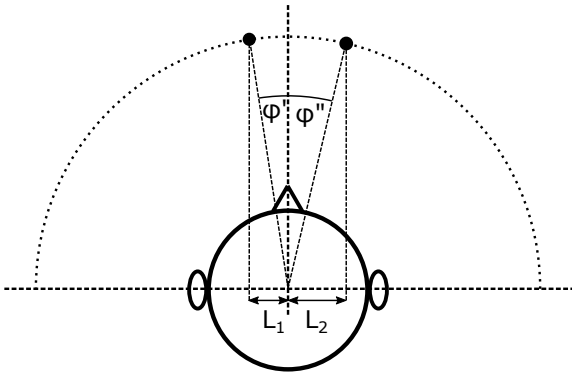


Figure 2.7: A relationship between lateralization and perceived lateral angle of the phantom sound source at a fixed distance.

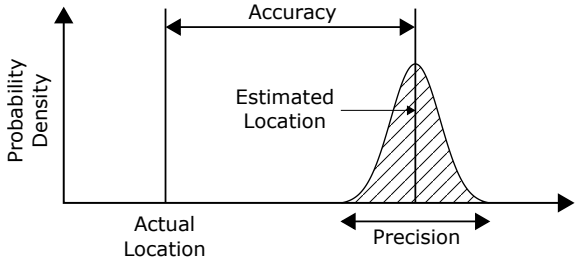


Figure 2.8: Auditory localisation accuracy and precision. Adapted from Łętowski and Łętowski (2012).

the sound source. The misplacement between the perceived location of the sound source and its actual location can be defined as the sum of the following two error metrics: *constant localisation error* and *random localisation error*. They represent the *accuracy* and *precision*, also known as *resolution*, of auditory localisation, respectively. The localisation accuracy and precision concept is illustrated in Figure 2.8.

The spatial precision of the auditory system can be associated with the minimum difference in the direction of the sound source that causes a change in the perceived position of an auditory event. Blauert and Allen (1997) described this attribute as *directional auditory localisation blur*. It is also called the Minimum Audible Angle (MAA) when obtained in sound source discrimination experiments. The lower limit of the MAA in the horizontal plane is about 1° (Mills, 1958; Perrott and Saberi, 1990) and is observed for sound sources located in front of the listener. The horizontal MAA increases to about 10° for the lateral directions. This confirms that auditory spatial resolution in the horizontal plane depends on the discrimination thresholds of interaural time and level differences, which change more rapidly in the function of azimuth for sources located ahead of the listener compared to the lateral region.

In the case of sources located in the median plane, the auditory localisation performance depends on the directional filtering of the pinna and body. The auditory localisation in the median plane is most precise for sound sources placed in front of the listener. Perrott and Saberi (1990) reported MAA of approximately 3° in these directions. A further study by Middlebrooks (1999) reported that the RMS

localisation error in a free field in all directions of the median plane is approximately 23°.

For sources located on diagonal planes, the spatial resolution depends on interaural differences and spectral changes (Perrott and Saberi, 1990; Grantham et al., 2003). Auditory localisation performance degrades in the presence of other sound sources (Langendijk et al., 2001). The accuracy of distance estimation tends to be poor, with the judged distance consistently underestimated Zahorik (2002a). For more comprehensive information on spatial hearing and auditory localisation, the reader is referred to the research report by Łętowski and Łętowski (2012) and a textbook authored by Van Opstal (2016).

2.4 Conclusion

This chapter has provided a comprehensive overview of auditory perception, which forms the foundation for the experiments presented in the further chapters. The reader was introduced to the basics of sound propagation, the human auditory system and spatial hearing. The mechanisms of perception of sound source direction and distance as well as auditory localisation performance were discussed. This chapter also introduced two crucial concepts related to headphone listening, i.e. externalization and binaural lateralization. Overall, this chapter has provided a solid foundation for understanding the principles which will be crucial for interpreting the results of the experiments presented in the subsequent chapters of this dissertation.

Chapter 3

Spatial Audio Techniques

Following the introduction of human auditory perception principles in Chapter 2, this chapter focuses on audio techniques which use the aforementioned principles. Figure 3.1 shows the general classification of these techniques. Suitable audio capture, coding, and reproduction methods must be used to deliver plausible spatial audio scenes to the listener. This chapter highlights techniques applicable to the scope of this thesis, followed by the introduction of spatial audio evaluation methods.

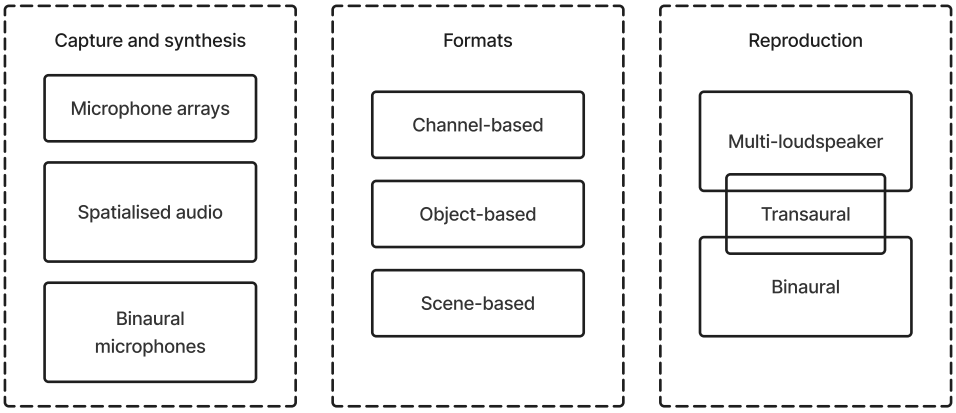


Figure 3.1: General classification of spatial audio techniques including audio capture and synthesis methods, spatial audio formats and reproduction methods.

3.1 Multi-Loudspeaker Reproduction

Multichannel loudspeaker systems create spatial auditory events by sending coherent signals to two or more loudspeakers. The most straightforward multichannel systems are known as *stereophonic* and consist of two loudspeakers. Such systems can create the illusion of sound sources located along the horizontal line between the two loudspeakers. Adding additional loudspeakers around the listener in the horizontal plane has led to the development of *surround* systems. An example of a surround loudspeaker is the 5.1 configuration, as defined in ITU-R (2022b), which is a popular

mixing standard in film postproduction. For more information on such systems, the reader is directed to (Roginska and Geluso, 2017, Chapter 6).

Spatial audio reproduction with height requires loudspeakers positioned at different heights. In recent years, a Dolby Atmos format has emerged, which employs speakers at elevated positions. This format combines the following two methods of sound spatialisation: *channel-based* and *object-based* audio.

The first is based on discrete loudspeaker signals created during the mixing process. The limitation of channel-based audio is that to achieve the intended effect, the loudspeakers used for reproduction should be located accordingly to the standardised layout (ITU-R, 2022a), which was used during material production. If the target loudspeaker layout does not match the original one, it is possible to derive a new set of loudspeaker signals by downward mixing content, although this might not be optimal. The down-mixing process is simply a summation of loudspeaker signals using specific weights.

In contrast, object-based audio represents individual audio objects that are stored along with the associated spatial metadata (Geier et al., 2010). This allows rendering of the virtual audio sources using arbitrary loudspeaker layouts, as the final loudspeaker feeds are produced during the rendering process considering each object's intended spatial position. Another feature of such systems is the possibility of content manipulation by the listener, e.g. to boost the film dialogue volume. For more information on object-based audio, the reader is directed to (Roginska and Geluso, 2017, Chapter 8).

There are alternative methods for spatial audio delivery over multiple loudspeakers that focus on the sound field reconstruction rather than panning between multiple loudspeakers. These methods are Wave Field Synthesis (Berkhout et al., 1993) and Ambisonics, further discussed in Section 3.4. At this point, it is pertinent to mention the Audio Definition Model (ADM) (ITU-R, 2019), an open standard that describes additional data that accompany audio produced with the methods mentioned above. This data can be later used in the rendering process to process the audio streams correctly.

3.2 Binaural Recordings

The basic idea behind the binaural technique is to recreate the previously recorded auditory experience by reproducing the sound pressure at each eardrum (Møller, 1992). Assuming that binaural localisation cues (ITD, ILD) and spectral cues are correct, the experience should be equal to listening to the actual sound source. This can be achieved by capturing and playing back *binaural recordings* or using *binaural synthesis*.

The typical binaural recording is made using a pair of small microphones placed inside the ear canal of each ear. The microphones can be attached to a real human head or an artificial head with the shape of an average human head, which usually includes a model of a pinna and nose. An example of such a device is the KU100

(a) In-ear microphone.

(b) KU100 manufactured by Neumann.

(c) KEMAR manufactured by GRAS.

Figure 3.2: Binaural microphones.

binaural microphone manufactured by Neumann[1]. Some binaural microphones include a simulated torso, e.g. KEMAR manufactured by GRAS[2]. Figure 3.2 shows an in-ear microphone and both artificial heads. In virtual acoustic environment reproduction, binaural recordings have limited use, as they do not allow the user to move their head to explore the scene. Therefore, the common term *static binaural* describes this type of recording. The binaural recording technique is typically limited by the generic morphological characteristics of the head used for recording.

Binaural audio playback is usually done using headphones, which allows for controlled signal distribution to each ear, eliminating the acoustic influence of the environment and the cross-talk, which would be present in regular loudspeaker-based reproduction. In addition to the headphone-based reproduction method, there is a particular case of binaural reproduction using two or more loudspeakers, the so-called *cross-talk cancellation* method, also known as the *transaural* method (Moller, 1988; Takeuchi and Nelson, 2002).

3.3 Binaural Synthesis

Binaural synthesis extends the binaural recording technique by allowing for the spatialisation of arbitrary sound sources. One of the first attempts to produce a simulation of free-field listening using digital techniques was made by Wightman and Kistler (1989a). The transfer function between a specified point in the free field and a point close to the eardrum must first be obtained to reproduce the experiment. This function depends on the angle of incidence of the source and its distance. It is commonly referred to as Head-Related Transfer Function (HRTF). It represents the

[1] `https://en-de.neumann.com/ku-100`

[2] `https://www.grasacoustics.com/products/product/749-45bc.html`

spectro-temporal filtering of an incident acoustic wave caused by the listener's pinna, head, and torso morphology. In other words, HRTF describes sound transmission from a point located in a free field at a certain distance and angle of incidence to a point located inside the ear canal (Møller, 1992). A set of HRTFs or HRIRs (HRTF equivalent in the time domain) is required to synthesise binaural signals corresponding to sound sources at different locations.

3.3.1 HRTF-based Spatialisation

HRTF-based signal processing is a crucial element of all binaural reproduction systems. A binaural signal can be synthesised by multiplying in the frequency domain the source signal and the respective HRTF for each ear. The optimal approach to spatial audio reproduction through binaural synthesis is to use the individual's HRTFs to preserve the subject-specific binaural signal attributes. If it is impossible to obtain an individual HRTF set for the person, the generic HRTF set must be used.

An alternative to direct HRTF convolution is to use HRTFs as virtual loudspeakers. Here, the same signals that would be utilised in a real-world 3D loudspeaker array are convolved with the HRTFs corresponding to specific loudspeaker positions to give a virtualised presentation of the array over headphones. The accuracy of sound field reproduction then becomes dependent on the spatialization method used, for example, Vector-Base Amplitude Panning (Pulkki, 1997), Wave Field Synthesis (Berkhout et al., 1993) or Ambisonics (Gerzon, 1973).

The binaural synthesis of acoustical environments requires measurements or simulations of Binaural Room Impulse Responses (BRIRs), which contain the contribution of the room acoustics. This type of synthesis can provide a high degree of perceptual realism (Lindau et al., 2007). Early reflections can be particularly important for the externalisation of sound sources (Begault et al., 2001).

3.3.2 HRTF Measurements

The established methods for obtaining user-specific HRTFs can be grouped into the following categories: acoustic measurements, mesh-based simulations or predictions based on ear images or anthropometric measurements. The acoustic measurement-based method is the most widely established for audio research and has been perceptually validated in many studies, e.g. (Wightman and Kistler, 1989b; Bronkhorst, 1995).

There are three approaches to HRTF measurements, distinguished by the point where the measurement probe is placed in the ear canal (Møller, 1992): at the ear drum, at the entrance to the open ear canal and at the entrance to the blocked ear canal. HRTFs, by definition, are measured in a free field, e.g. in an anechoic chamber. For a comprehensive review of HRTF measurement methods, the reader is directed to (Li and Peissig, 2020).

Figure 3.3: 3DOF head tracker.

3.3.3 Head Tracking

In opposition to binaural recordings, binaural synthesis can employ head tracking. It is required for dynamic binaural synthesis to provide information on the listener's movement. If the signals fed to each ear are modified accordingly, i.e. providing localisation cues based on the head displacement and rotation, presented virtual sound sources will remain in their positions in space. Perception of the location of the virtual sound source is improved with head tracking-enabled binaural audio systems (Wightman and Kistler, 1999; Begault et al., 2001).

There are multiple head-tracking solutions available. The simplest ones are the three-degrees-of-freedom (3DOF) trackers, which provide information on the rotation of the listener's head. An accurate approximation of head orientation can be derived using the data fusion algorithm based on measurements from the following sensors: accelerometers, gyroscopes, and magnetometers. Figure 3.3 shows a miniature 3DOF head tracker[3] placed on a headphone headband. Optical tracking is usually required to provide additional tracking of head displacement. Such systems are called the six-degrees-of-freedom (6DOF) tracking and are typically employed in virtual and augmented reality headsets or are available as more complex standalone solutions.

3.3.4 Equalisation

HRTF-based spatialisation requires controlled compensation of the frequency response of headphones used for binaural reproduction. It is desirable to use individual headphones compensation filters (Pralong and Carlile, 1996). When individual or headphone-specific filters are not available, it is best to use diffuse-field equalised HRTF sets. Diffuse-field equalisation (DFE) is introduced in order to remove the direction-independent component of HRTF magnitudes (Common Transfer Function). In many cases, this leads to an improvement in timbral balance when listening through standard headphones, which typically are roughly equalised to match the average diffuse field responses of the human head and pinna (Møller et al., 1995; Larcher et al., 1998).

[3] `https://github.com/trsonic/nvsonic-head-tracker`

3.4 Ambisonics

Ambisonics was introduced as a multi-loudspeaker sound reproduction technique in all three dimensions, also known as *periphony* (Gerzon, 1973, 1980). In contrast to object- and channel-based approaches, Ambisonics approximates the full-sphere sound field at the listener's ears. It is based on the spherical harmonic representation of the sound field. The theory of Ambisonics is well documented, and the reader is directed to (Zotter and Frank, 2019; Daniel et al., 2003; Kearney, 2010) for a good explanation of the topic. This section provides only a brief review.

3.4.1 The Microphone Analogy

As Kearney (2010) explains, Ambisonics can be considered an extension of the widely adopted recording technique, where a directional microphone signal can be obtained through the superposition of signals from two coincident in-space microphones: omnidirectional (pressure-sensitive) and bidirectional (velocity-sensitive, figure-8 pattern). Such a *virtual microphone* is oriented along a single axis, and its directivity characteristic can be manipulated by adjusting the weights of the omnidirectional and figure-8 microphone signals. In this scenario, the on-axis direction of the virtual microphone coincides with the direction of the figure-8 microphone.

If we consider placing two additional figure-8 microphones at the same point in space as the first two discussed microphones, oriented orthogonally to each figure-8 microphone main axis, we can create a virtual microphone based on the superposition of all four microphone signals. These signals are called Ambisonic components. They represent specific portions of a sound field, as defined by the microphone directivity functions. A multichannel signal that carries Ambisonic components representing a full sound field is often called *B-Format*, a legacy term introduced in the 1970s. A virtual microphone signal derived from Ambisonic components can be oriented in any direction, and its directivity pattern can be modified.

3.4.2 Spherical Harmonics

The three-dimensional hierarchical basis functions used in Ambisonics are known as *spherical harmonics* (Daniel et al., 2003). Following the naming convention used by Armstrong and Kearney (2021) the spherical harmonics shown in Figure 3.4 can be identified by their *degrees* and *indices* corresponding to the rows and columns of the figure respectively. Ambisonic *order* corresponds to the highest degree of spherical harmonics used in a given set. Ambisonic sound field representations employing 0th and 1st-degree spherical harmonics exclusively are known as First Order Ambisonics (FOA), whereas representations extended by using the higher degrees of spherical harmonics are called Higher Order Ambisonics (HOA). Only a limited number of Ambisonic components can be used in practical scenarios. The required number of components N for a periphonic system with truncation order M equals:

$$N = (M + 1)^2 \tag{3.1}$$

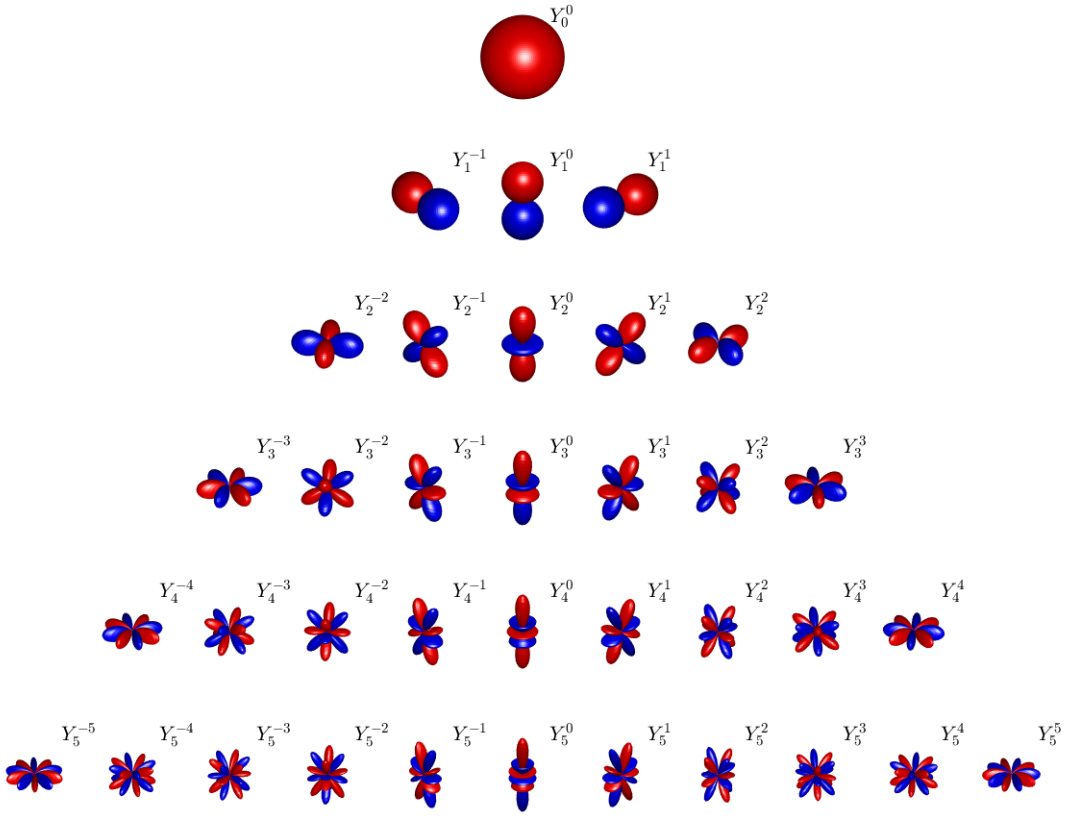


Figure 3.4: 0th- to 5th-degree spherical harmonics normalised using the Schmidt semi-normalisation.

Spherical harmonic coefficients can be calculated as

$$Y_m^i(\theta,\phi) = N_m^{|i|} \times P_m^{|i|}(\sin\phi) \times \begin{cases} \cos(|i|\theta) & \text{if } i \geq 0, \\ \sin(|i|\theta) & \text{if } i < 0, \end{cases} \tag{3.2}$$

where (θ, ϕ) is the direction of the arriving plane wave (θ is the azimuth angle, ϕ is the elevation angle); $m = 0, 1, 2, ...$ is the spherical harmonic degree; $i = -m, ..., m$ is the index, $N_m^{|i|}$ is normalisation term, $P_m^{|i|}(\sin\phi)$ is the associated Legendre polynomial with the Condon-Shortley phase undone. Common normalisation terms include SN3D (Schmidt semi-normalisation), N3D (orthogonal normalisation) and O3D (orthonormal normalisation). These terms can be computed as follows:

$$N_m^{|i|^{SN3D}} = \begin{cases} 1 & \text{if } m = 0, \\ \sqrt{\frac{2(m-|i|)!}{(m+|i|)!}} & \text{if } m > 0, \end{cases} \tag{3.3}$$

$$N_m^{|i|^{N3D}} = \begin{cases} 1 & \text{if } m = 0, \\ \sqrt{(2m+1)\frac{2(m-|i|)!}{(m+|i|)!}} & \text{if } m > 0, \end{cases} \tag{3.4}$$

$$N_m^{|i|^{O3D}} = \begin{cases} \frac{1}{\sqrt{4\pi}} & \text{if } m = 0, \\ \sqrt{\frac{(2m+1)}{4\pi}\frac{2(m-|i|)!}{(m+|i|)!}} & \text{if } m > 0. \end{cases} \tag{3.5}$$

The use of SN3D normalisation is preferred for the production, exchange and playback of Ambisonic content, as it attenuates higher-degree components and prevents their amplitude from exceeding the amplitude of the 0th-degree signal. SN3D normalisation and the ACN channel ordering scheme are specified by the AmbiX format (Nachbar et al., 2011). The ACN ordering specifies 0-indexed channel numbers corresponding to the Ambisonic components as follows:

$$ACN = m^2 + m + i. \tag{3.6}$$

3.4.3 Encoding and Manipulation

A single-channel audio signal can be encoded to Ambisonics by matrix multiplication with spherical harmonic coefficients representing the position of the source on the surface of a sphere. Ambisonics has been widely adopted for immersive audio applications since it can be easily manipulated and transformed (Kronlachner, 2014b). For example, Ambisonic scene rotation is used to facilitate stable sound sources in dynamic binaural rendering. Due to the wide variety of Ambisonic audio plugins, performing these operations in real-time without any prior mathematical and coding expertise is possible.

3.4.4 Microphone Capture

Although Ambisonic components of 0th- and 1st-order are related to the spherical representation of the sound field by spatial characteristics that match typical microphone directivity patterns (omnidirectional and figure-8), such an array can only be approximated in reality, as it is impossible to place multiple electroacoustic transducers at the same point in space. However, Ambisonic microphones created by the combination of omnidirectional and figure-8 microphones can deliver considerably good results. This microphone technique is called *native B-Format* (Benjamin and Chen, 2005).

The most popular group of Ambisonic microphones are tetrahedral arrays, first proposed by Craven and Gerzon (1977). Such microphones typically employ cardioid or subcardioid capsules aligned with tetrahedron faces, as shown in Figure 3.5. Raw signals from the capsules represent spatially sampled portions of the sound field (sometimes referred to as A-Format). Using a conversion matrix and a set of correction filters, it is possible to derive 0th- and 1st-order Ambisonic components from these signals.

Another group of Ambisonic microphones are higher-order spherical arrays. HOA components are derived on the basis of the pressure distribution on the solid sphere. An example of an HOA microphone is the Eigenmike EM32[4], as shown in Figure 3.5, which employs 32 pressure-sensitive (omnidirectional) capsules.

[4] `https://mhacoustics.com/products`

(a) 1st-order tetrahedral microphone (SoundField ST250).

(b) 4th-order spherical microphone (Eigenmike EM32).

Figure 3.5: Ambisonic microphones.

3.4.5 Loudspeaker Rendering

In opposition to traditional channel-based surround sound techniques, the production of Ambisonic audio content does not require detailed knowledge of the layout of the loudspeaker reproduction system. Ambisonic recordings can be reproduced on any multichannel loudspeaker configuration. However, the best performance is achieved with regular loudspeaker arrays. Similarly to the encoding process, decoding the Ambisonic stream to loudspeaker feeds can be done using matrix multiplication. The decoding matrix can be calculated using a Mode-Matching method as a pseudo-inverse of the encoding matrix for the known loudspeaker positions. The minimum number of loudspeakers should be greater than the number of Ambisonic components.

A limited number of spherical harmonic components leads to a truncated sound field representation. Therefore, higher orders have to be used to achieve higher spatial resolution. Increasing the truncation order M results in a higher spatial aliasing frequency f_{alias} and a larger sweet spot (Daniel et al., 2003). Timbral distortions in Ambisonics also depend on the sound-field reconstruction accuracy, which improves with increasing Ambisonic order. Above f_{alias} Ambisonic decoding introduces timbral alterations in the encoded signals (McKenzie et al., 2018) degrading overall timbral fidelity.

In simple terms, spatial aliasing is a direct consequence of the displacement of human ears from the theoretical centre of the virtual loudspeaker array. Moreau et al. (2006) define f_{alias} as

$$f_{alias} = \frac{c\mathrm{M}}{4r(\mathrm{M}+1)\sin\left(\frac{\pi}{2(\mathrm{M}+1)}\right)} \tag{3.7}$$

which can be approximated as

$$f_{alias} = \frac{c\mathrm{M}}{2\pi r}, \tag{3.8}$$

where r is the head radius. Table 3.1 shows the spatial aliasing frequencies for different Ambisonic orders and head radius $r = 8.5$ cm. Some decoding schemes involve frequency-dependent weighting of Ambisonic components for psychoacoustic optimisation, as described by Daniel et al. (1998).

Table 3.1: Spatial aliasing frequencies.

Ambisonic order	1	2	3	4	5	6	7
f_{alias} (Hz)	642	1284	1927	2569	3211	3853	4496

Nonlinear optimisation methods can be used to optimise decoding matrices for irregular arrays. Craven (2003) and Wiggins (2007) used heuristic methods to generate Ambisonic decoding matrices. The main focus of their research was the five-speaker horizontal layout as defined in (ITU-R, 2022b), although the methods could be extended to other loudspeaker configurations. Wiggins proposed the use of the Tabu Search algorithm to find the optimal solution based on the optimisation of the velocity and energy vector criteria (Gerzon, 1992), while Benjamin et al. (2010) proposed the use of the NLopt library (Johnson, 2014) for the same purpose.

A widely established method for decoding Ambisonics to irregular loudspeaker arrays is the AllRAD approach (Zotter and Frank, 2012). An explanation of this and other decoding methods can be found in (Zotter and Frank, 2019).

3.4.6 Binaural Rendering

Although Ambisonics was initially proposed as a technique for multichannel loudspeaker reproduction, using modern DSP processing, Ambisonics can be binaurally rendered over headphones incorporating generic or individual HRTF sets (Jot et al., 1998; Noisternig et al., 2003). Conventionally, the binaural signal is obtained through the convolution of decoded virtual loudspeaker feeds and respective HRIRs of each ear. Most binaural rendering systems supporting Ambisonics use computationally efficient spherically decomposed HRTF sets, i.e., precomputed combinations of virtual loudspeakers and HRTFs (Gorzel et al., 2019).

A benefit of Ambisonic-based binaural rendering is the ease of 3DOF head-tracking implementation. To create a static virtual environment around the listener, head movements should be counteracted by controlled rotation of the Ambisonic scene. This can be done by a simple matrix transformation of the Ambisonic signal. In such a case, the directions of virtual loudspeakers, simulated by HRTFs, remain unchanged, whereas the loudspeaker feeds to the virtual loudspeakers are updated in real-time.

A traditional Ambisonics to binaural renderer can be considered as a set of time-invariant filters capable of reconstructing the perceptually relevant characteristics of the HRTF set used for its design. A bank of these filters can be described as a three-dimensional matrix h_{LR}^{SH} consisting of time-domain filters designated to process

each Ambisonic component for each ear. These filters are called in this thesis SH-HRIRs. Ambisonic signals convolved with these filters result in a binaural mix that can be delivered to a listener through headphones. Three-dimensional matrices h_{LR}^{SH} can be easily saved as multichannel WAV files and used with multichannel convolver plugins, e.g. mcfx_convolver[5].

Some implementations of binaural Ambisonic renderers use two-dimensional matrices h^{SH} containing single ear data for symmetrical processing of ear signals in order to reduce required computational resources. Such a matrix is used in Resonance Audio renderer, which employs filters derived from KU100 dummy-head HRIRs obtained from the SADIE database (Kearney and Doyle, 2015). Similarly, the IEM BinauralDecoder[6] uses a matrix derived from a single ear of KU100 HRIR set measured by Bernschütz (2013).

Besides the time-invariant linear filters, another group of Ambisonic renderers uses information derived from the sound field to update decoding parameters in short time windows. These renderers utilise parametric methods to divide the sound scene into discrete sources and diffuse sound components, allowing for an up-scaled spatial resolution of the original Ambisonic scene. Nevertheless, the processed scene has to be decoded for the listener using either a set of discrete HRIRs or the SH-HRIRs discussed in this chapter. For more information on parametric decoding of Ambisonics, readers are referred to the work of Pulkki et al. (2017); Politis et al. (2018).

The methods for designing Ambisonics to binaural filters have been studied since the late 1990s. Traditionally, filter matrices are derived from HRIR sets using the virtual loudspeaker approach (McKeag and McGrath, 1996; Jot et al., 1998; Noisternig et al., 2003) or by spherical harmonic decomposition in frequency bands (Zotkin et al., 2009). However, order truncation introduces significant HRTF reconstruction errors at frequencies above the spatial aliasing limit f_{alias}. This results in inaccurate ITD and ILD cues and a significant distortion of the spectral cues present in the original HRTFs, as discussed in Section 3.4.5.

Several HRTF manipulation techniques have been proposed to optimise the binaural rendering of Ambisonic signals at high frequencies. Possible methods aimed at mitigating the reconstruction errors above f_{alias} include sub-sampling of HRIR measurement grid (Bernschütz, 2014), spectral equalisation (Ben-Hur et al., 2017; McKenzie et al., 2018), frequency-dependent time alignment of HRIRs (Zaunschirm et al., 2018), SH-domain tapering (Daniel et al., 1998; Hold et al., 2019) or bilateral Ambisonic rendering using ear-aligned HRTFs (Armstrong et al., 2018a; Ben-Hur et al., 2021).

A comprehensive study by Engel et al. (2022) evaluated different methods of improving binaural rendering accuracy at high frequencies using perceptual models. Among the evaluated methods were spatial subsampling (Bernschütz, 2014), equalisation (Ben-Hur et al., 2017), SH-domain tapering following Hold et al. (2019) and the Magnitude Least Squares (MagLS) method proposed by Schörkhuber et al. (2018). However, the study did not include the binaural filter design method implemented

[5] `https://github.com/kronihias/mcfx/blob/master/CONVOLVER_CONFIG_HOWTO.txt`
[6] `https://plugins.iem.at/docs/plugindescriptions/#binauraldecoder`

within Google Resonance, which is used as a baseline rendering scheme across this dissertation. The method combines spatial subsampling with SH-domain tapering via applying Max-Re weighting above f_{alias}. The results of the study found the MagLS method provided the most accurate rendering of conventional order-truncated Ambisonic scenes.

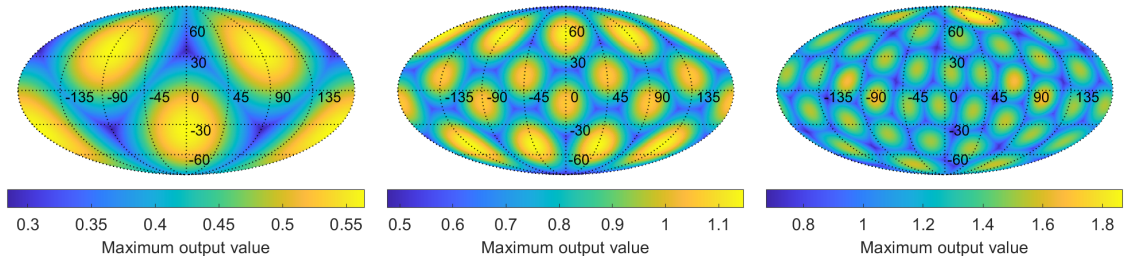
There is a limited number of perceptual studies comparing the MagLS decoders against other methods. A listening test study by Lee et al. (2019) evaluated the MagLS decoders calculated at different Ambisonic orders against a standard 1st-order least squares decoder calculated using a set of HRIRs in a "cube" layout without any optimisation. The study looked at the timbral and spatial fidelity of Ambisonic renders separately. All renders created using the MagLS decoders were rated higher than the traditional decoder in both perceptual domains. A more recent study by Lübeck et al. (2020) included the MagLS method among other optimisation strategies to assess the perceived differences in rendering of anechoic drum samples encoded at 3rd-, 5th- and 7th-order Ambisonics using Ambisonic RIRs. The results showed that the MagLS method provided a similar degree of improvement compared to other optimisation strategies included in that study. Therefore, it is not clear which method is superior to others for the most common scenarios, which is rendering of 1st-, 3rd- and 5th-order Ambisonic scenes using generic HRTF sets. Further research focusing on the perceptual evaluation of MagLS and other methods is required to confirm the findings of the objective study by Engel et al. (2022).

3.4.7 Low-bitrate Coding

Due to the increasing popularity of immersive content streaming, there is a rising demand for efficient audio compression algorithms optimised for spatial audio techniques, including Ambisonics. In some cases, the use of Ambisonics reduces the bandwidth and computational requirements needed to deliver immersive audio content compared to traditional multichannel surround formats and object-based approaches (Brettle and Skoglund, 2016). MPEG-H (Herre et al., 2015) and Opus (Valin et al., 2013) are the only perceptual codecs that officially support Ambisonics. The Opus codec is employed by both YouTube VR and HOAST.

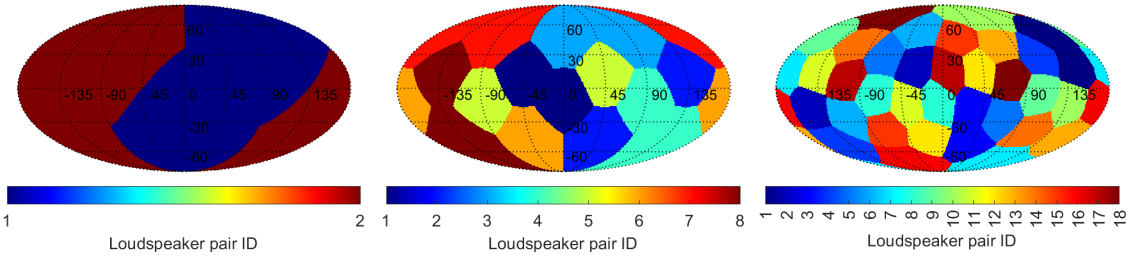
The recent versions of the Opus codec[7] implement two types of channel mappings for encoding Ambisonics (Skoglund and Graczyk, 2018). Channel Mapping Family 2 (CMF2) codes each Ambisonic component as an independent Opus stream, i.e. each channel is encoded separately. This direct uncoupled method does not take advantage of codec features such as coupled stereo mode. Channel Mapping Family 3 (CMF3) utilises these features through projection-based signal decomposition, where channels correspond to virtual loudspeakers and are effectively coupled together in pairs when coded, and then a demixing matrix is used to separate Ambisonic components upon decoding. The publicly available version of Opus contains the 1st- and 3rd-order Ambisonics matrices for Channel Mapping Family 3. For this research, a patch was added to the official Opus 1.3.1 code to support 5th-order Ambisonics which can be provided upon request.

[7] `https://opus-codec.org/`



(a) 1OA (b) 3OA (c) 5OA

Figure 3.6: Maximum output value of the Channel Mapping Family 3 projection matrix multiplied by spherical harmonics evaluated across the sphere.



(a) 1OA (b) 3OA (c) 5OA

Figure 3.7: Virtual loudspeaker pairs used for coupled bitrate compression.

Further investigation of the Channel Mapping Family 3 can be done by analysing the projection (mixing) matrices embedded into the Opus codec source code. Figure 3.6 shows the distribution of maximum values of the virtual loudspeaker feed vector across all directions. This vector is obtained from the multiplication of the vector of spherical harmonics evaluated at a single direction and the CMF3 projection matrix. An increase in the maximum value for certain directions coincides with the virtual loudspeaker coordinates. Figure 3.7 shows the colour-coded pairs of virtual loudspeakers used for coupled bitrate compression (stereo mode). It can be seen that both the virtual loudspeaker layouts and the neighbouring loudspeaker coupling could be possibly improved for the 3rd- and 5th-order Ambisonic projection matrices, however, this remains outside of the scope of this research.

Several studies have been conducted investigating the quality of low-bitrate perceptual coding for surround systems. However, limited research has been published on the quality of compressed 1st- and higher-order Ambisonics. Previous work in this field includes subjective evaluation of Ambisonic scenes compressed with the Opus codec with Channel Mapping Family 2 implementation (Narbutt et al., 2017) and developing a reference objective spatial audio quality metric (Narbutt et al., 2018). No research prior to this work investigated differences in the perceived quality of Opus compressed audio under different reproduction conditions as well as between the two channel mapping families.

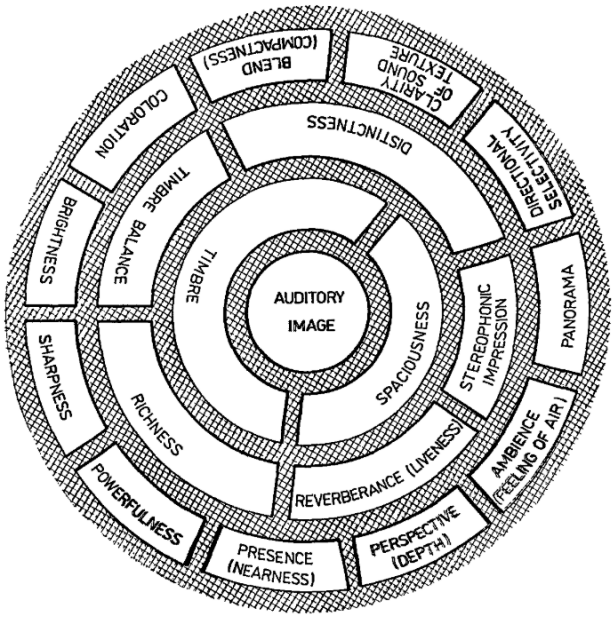


Figure 3.8: Multilevel Auditory Assessment Language proposed by Łętowski (1989).

3.5 Perceptual Evaluation of Spatial Audio

Perceptual experiments play an important role in audio research and the development of high-quality audio systems. While psychoacoustic research is based on listener-oriented tests, the subjective evaluation of spatial audio looks at the evaluation of external objects, e.g. audio hardware or rendering algorithms. Perceptual tests should be carried out in a controlled way so that the results represent the perceived characteristics of the system and can be used to support scientific research or engineering processes. For an in-depth review of established evaluation methods, the reader is directed to the book by Zacharov (2018).

3.5.1 Perceptual Attributes

Early research on the perception of sound quality has been conducted by Helmholtz (2009), who described holistic listening where the auditory image is perceived as a whole without paying attention to its elements. Further research led to the development of analytic approaches. According to Łętowski (1989), auditory events can be discriminated according to the following psychoacoustic attributes: loudness, pitch, duration, spaciousness and timbre. The first three sensations are usually excluded from the auditory assessment of audio systems. Figure 3.8 shows the auditory assessment framework based on the two remaining multidimensional attributes: timbre and spaciousness.

Further research carried out by others led to the development of more sophisticated attributes and definitions focused on spatial audio (Zacharov and Koivuniemi, 2001; Lindau et al., 2014). In 2017 ITU-R published a comprehensive report (ITU-R, 2017)

on the selection and description of attributes for the preparation of subjective tests. Such attributes should be characterised by good discrimination power among stimuli, good agreement among assessors, relevance to the use cases of the assessed system, low redundancy and low correlation with other attributes. The report contains a psychoacoustic attribute classification proposed by Pedersen and Zacharov (2015).

3.5.2 Evaluation Methods

The two primary families of perceptual evaluation methods are discrimination and integrative methods. Discrimination methods are used to establish with certain confidence whether there is a perceived difference between the conditions tested. These methods include paired comparison, ABX, and 2-AFC. Integrative methods establish information on perceived quality or impairment on a scale. An example of such a method is the ITU-R BS.1534-3 (ITU-R, 2015b) recommendation, commonly called MUSHRA, which stands for Multiple Stimuli with Hidden Reference and Anchor. In a MUSHRA test, the assessor is asked to rate specific audio attributes on a Continuous Quality Scale compared to the reference audio condition. The experimental conditions should include a hidden reference as well as low- and intermediate-quality anchors. The ITU-R BS.1116-3 (ITU-R, 2015a) recommendation is designed to assess minor impairments in audio systems. These types of tests are referred to as direct evaluations, as the participants are asked to provide their judgments directly using a designated user interface.

Good auditory localisability is essential for immersion and authenticity. External factors affecting auditory localisation in immersive audio reproduction include audio rendering schemes (Ben-Hur et al., 2020) and HRTFs (Wenzel et al., 1993). Another factor which potentially affects localisation is perceptual audio coding, which has not been studied in this context.

Auditory localisation in spatial audio systems can be evaluated by measuring the listener's performance in sound source localisation tasks. Such tests are also known as indirect evaluations, as the results are inferred from the collected data. Auditory localisation tests employ various response techniques, e.g. perceived direction reporting, visual mapping, physical pointing, and acoustic pointer adjustment. Egocentric pointing methods employ localisation judgement reported in a coordinate system centred on the listener's body. The position of the sound source is indicated by the listener's hand or head pointed in the perceived direction (Majdak et al., 2010). As the egocentric method seems more intuitive, it suffers from uncertainty introduced by the displacement between the intuitive centre of the body and the centre of the listener's head. This problem can be partially mitigated by using the *proximal pointing*, where the listener points in the apparent direction of the sound source in the proximal region of the head (Bahu et al., 2016). In exocentric methods, the listener indicates the perceived direction of the sound source using an external device, e.g. pointing with a stylus on a solid sphere placed in front of their body (Gilkey et al., 1995). Another method of localisation performance evaluation employs a real or virtual acoustic pointer. This method has previously been used to evaluate Ambisonic systems (Bertet et al., 2013; Thresh et al., 2017).

3.5.3 Listening Test Tools

The popular listening test tools are web-based (Schoeffler et al., 2018), created using graphical audio programming environments like Max[8] (Gribben and Lee, 2015), or created using MATLAB (Vazquez, 2015; Ciba et al., 2009). The established tools allow listening tests to be performed according to the methods discussed in the context of stereo or static binaural reproduction. However, spatial audio evaluations often require the playback of multichannel audio simultaneously alongside additional processing in the signal chain. This requires building custom listening test software or introducing software modifications.

Another challenge is to provide participants with an optimal physical test interface, as traditional desktop and laptop computers can introduce acoustic shadowing and reflections while using loudspeaker playback systems. Using compact and wireless interfaces, like tablets, can minimise the interface's influence and make the test more convenient for the assessor. Therefore, it is sometimes necessary to consider the development of a custom listening test interface based on the planned experiment.

3.6 Conclusion

This chapter discussed audio techniques used for creating and delivering spatial audio, focusing on critical concepts on which the research presented in this thesis is based, i.e., Ambisonics, HRTFs and perceptual evaluation of audio systems. A limited amount of research on the problems of the perceived audio quality of low-bitrate compressed Ambisonic scenes has been identified. Another area of research which has not been thoroughly investigated using perceptual methods is the binaural rendering of Ambisonic signals using state-of-the-art methods.

[8] `https://cycling74.com/products/max`

Chapter 4

Evaluation of Timbral Distortion in Bitrate-Compressed Ambisonic Scenes Using Loudspeaker and Headphone-Based Reproduction

The increasing popularity of Ambisonics as a spatial audio format for streaming services poses new challenges to established audio coding techniques. Immersive audio delivered to mobile devices requires efficient bitrate compression that does not affect content quality. The most widely used audio codec for Ambisonics is Opus (Valin et al., 2013), employed by the YouTube VR platform for delivering spatial audio along with 360° videos.

A limited amount of research has been published on the quality of compressed 1st-and higher-order Ambisonics. Researching the degree of audio quality degradation introduced by compression at different Ambisonic orders is pertinent. This chapter, along with Chapter 5 and Chapter 6, presents a group of experiments aimed at investigating the differences in perceived audio quality of signals encoded with the Opus codec at different settings and audio contexts.

Typically, the quality of multichannel perceptual coding systems is evaluated using a single attribute, Basic Audio Quality (BAQ). BAQ is used to judge any and all detected differences between an unimpaired reference and compressed stimuli (ITU-R, 2015b). BAQ in the context of home cinema surround sound reproduction is affected mainly by the timbral fidelity of the assessed audio, and the spatial fidelity accounts for as low as 30% of BAQ rating (Rumsey et al., 2005; Marins et al., 2008). Although using a single measure to quantify perceptual differences between particular systems is convenient, the complex nature of perceptual coding distortions requires a broader set of attributes to provide a meaningful comparison. The experiment presented in this chapter focuses on the perceived level of timbral distortion compared to the reference uncompressed audio. The spatial quality aspect of Ambisonic coding is investigated in Chapter 5.

Another factor after bitrate compression that affects the quality of Ambisonics is the sound field reconstruction accuracy, which degrades with the truncation of

the Ambisonic representation order. As a decrease in order results in a lower spatial aliasing frequency f_{alias}, Ambisonic decoding introduces timbral alterations to the encoded signals above f_{alias}, degrading overall timbral fidelity (McKenzie et al., 2018).

Using a multichannel loudspeaker array allows for the playback of Ambisonic audio without the need to use headphones and HRIR filters. Such a reproduction is technically closer to natural listening, provided that the listener's head is in the centre of the array. However, listening to Ambisonic content on headphones is much more widespread due to the hardware simplicity and the availability of audio streaming services supporting binaurally rendered Ambisonics. This work seeks to evaluate the impact of the rendering method on the perceived level of signal degradation introduced by Ambisonic order truncation and low-bitrate coding. The codec used in this study is Opus with Channel Mapping Family 3, which has been discussed in Section 3.4.7.

4.1 Methods

Listening tests were carried out based on the ITU-R BS.1534 (MUSHRA) (ITU-R, 2015b) recommendation to assess the degree of timbral distortion introduced by compression at different Ambisonic orders. Due to the large number of experimental conditions in this test, the mid-quality anchor specified in the MUSHRA guidelines was not included in the assessment. The test participants were asked to rate the level of timbral similarity on a continuous quality scale in relation to the reference audio sample, disregarding the spatial fidelity of the scenes.

4.1.1 Participants

All participants were masters and PhD students in audio engineering with experience in critical listening, although some participated in the sound quality assessment tests for the first time. Participants were instructed on how to perform the test by reading the information sheets and receiving individual demonstrations. They were instructed to keep their heads in the centre of the loudspeaker rig and limit head rotations throughout the test. All participants gave their informed consent to be included in the study. The protocol was approved by the Physical Sciences Ethics Committee of the University of York (approval code: Rudzki021018). Table 4.1 shows the number of participants who took part in each phase of the experiment.

4.1.2 Test Stimuli

The test material consisted of eight sound scenes (see Table 4.2) divided equally into simple and complex groups. The simple scene Ambisonic stimuli were created by encoding a single monophonic audio file to a location straight in front of the listener. The following samples were used: pink noise, a vocal sample, and two instrument

Table 4.1: Number of participants who completed the tests grouped by the rendering method and audio content type used. Values in brackets indicate the number of participants who passed the post-screening (described further in Section 4.1.6).

Reproduction Method	Content Type	Number of Participants
Loudspeakers	Simple	21 (19)
	Complex	16 (15)
Binaural (Individual HRTFs)	Simple	15 (11)
	Complex	14 (11)
Binaural (Generic HRTFs)	Simple	19 (15)
	Complex	19 (15)

Table 4.2: Audio material used for the timbral distortion assessment.

Symbol	Scene type	Description
A	Simple	Vocals (Suzanne Vega - "Tom's Diner")
B	Simple	Castanets (EBU)
C	Simple	Glockenspiel (EBU)
D	Simple	Pink Noise
E	Complex	Electronic Music 1
F	Complex	Electronic Music 2
G	Complex	Acoustic Music 1
H	Complex	Acoustic Music 2

samples from the EBU SQAM dataset[1]: Castanets and Glockenspiel. The vocal sample was an excerpt from the song "Tom's Diner" by Suzanne Vega, which has been previously used in audio codec research (Brandenburg and Henke, 1993). Complex Ambisonic scenes were music compositions consisting of multiple monophonic audio files encoded at different locations on the sphere. The first two scenes consisted of sampled and synthesised sounds. The second two were based on anechoic male vocal and instrumental recordings with simulated room reflections and reverberation added using the IEM RoomEncoder and IEM FdnReverb plugins respectively. The complete set of experimental stimuli can be obtained upon request.

4.1.3 Test Conditions

Each trial consisted of 11 conditions: nine compressed stimuli at different bitrates and orders, a hidden reference and a hidden anchor. Table 4.3 shows the bitrates of the codec under investigation. The 5th-order Ambisonics (5OA) uncompressed signal was used as the hidden reference. The anchor was created by low-pass filtering at 3.5 kHz the 1st-order Ambisonics (1OA) signal compressed at 16 kbps per channel (kbps/ch).

[1] `https://tech.ebu.ch/publications/sqamcd/`

Table 4.3: Investigated bitrates (kbps) at different Ambisonic orders.

	Bitrate per channel	Total bitrate		
		1OA	3OA	5OA
Compressed	16	64	256	576
Compressed	32	128	512	1152
Compressed	64	256	1024	2304
Uncompressed	768			27648

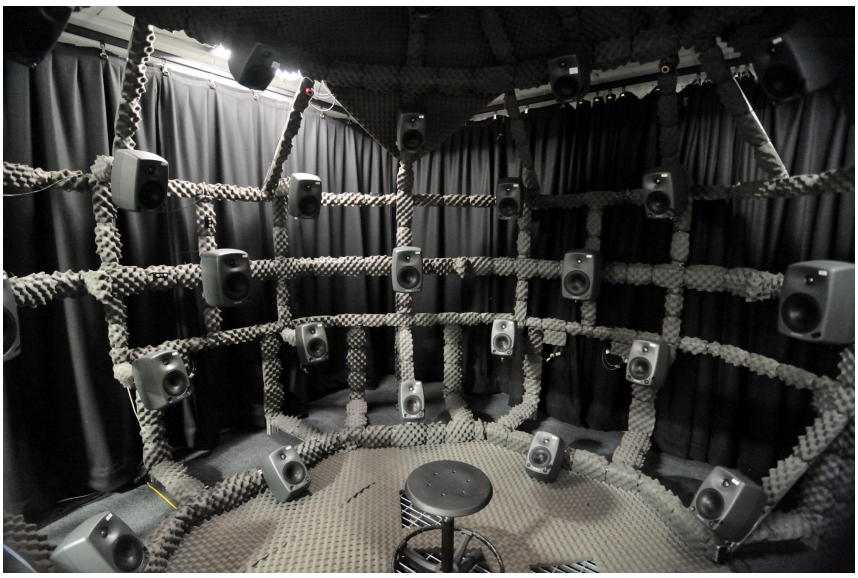


Figure 4.1: 50-channel spherical loudspeaker array at the AudioLab, University of York.

4.1.4 Spatial Audio Rendering

The evaluation was performed using multi-loudspeaker and dynamic binaural rendering methods inside an acoustically treated room. Loudspeaker reproduction used a 50-channel full-sphere array based on the Lebedev quadrature (see Figure 4.1). The rendering of the 5th-order Ambisonic scenes was done using all 50 loudspeakers (Lecomte et al., 2016). As shown in Figure 4.2, 1st- and 3rd-order scenes were rendered using subsets of loudspeakers based on the octahedron and the 26-point Lebedev grid, respectively. Dual-band decoding was implemented by prefiltering the Ambisonic input with a set of shelf filters[2] to apply Max-Re correction weightings above f_{alias} before feeding the decoder. AmbiX[3] Ambisonic decoder configuration files were obtained from the SADIE II database[4].

To create binaural signals, loudspeaker feeds were convolved in real-time with diffuse-field equalised HRTF sets obtained from the SADIE II database (Armstrong et al., 2018b). Individual and generic HRTF sets were used. The individual HRTF-

[2] `https://github.com/resonance-audio/resonance-audio/tree/master/matlab/ambisonics/shelf_filters/`

[3] `https://matthiaskronlachner.com/?p=2015`

[4] `https://york.ac.uk/sadie-project/ambidec.html`

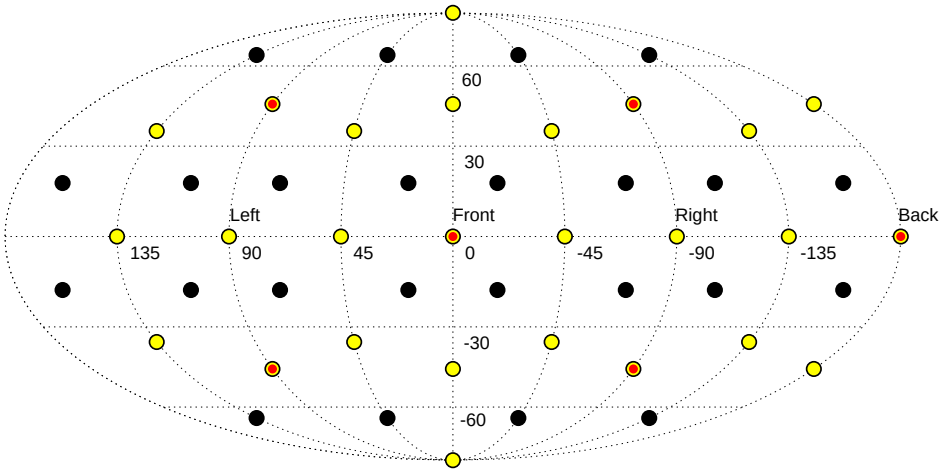


Figure 4.2: Loudspeaker configurations for decoding Ambisonics. 50-point Lebedev configuration (black), 26-point Lebedev subset (yellow) and octahedron subset (red) used for the rendering of 5th-, 3rd- and 1st-order Ambisonic scenes respectively.

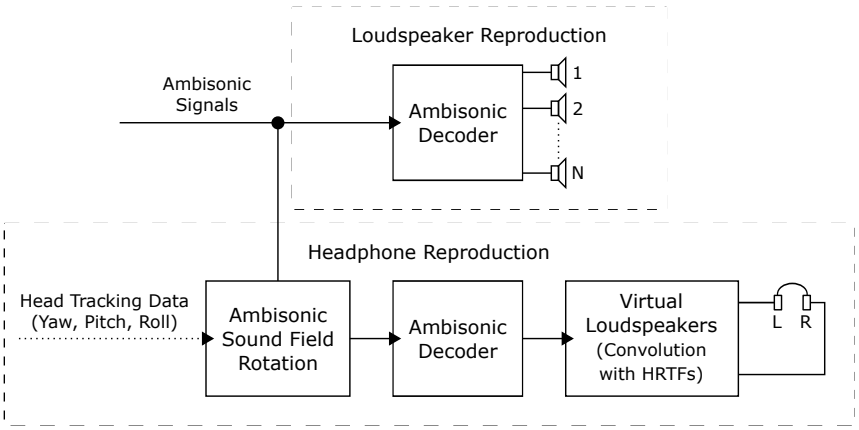


Figure 4.3: Block diagram illustrating the audio signal chain used for Ambisonic rendering over loudspeakers and headphones.

based evaluation required the participation of subjects who were included in the database. The generic HRTF evaluation was done using a Neumann KU100 binaural microphone HRTFs. Sennheiser HD 650 headphones were used for the binaural tests. This headphones model provides a consistent frequency response between the coupling and decoupling of headphones with ears (Adams and Boland, 2010). The frequency response of the headphones was equalised using inverse filters derived from measurements performed using the KU100 dummy head. Dynamic binaural rendering was done using an Optitrack optical motion tracking system[5] consisting of six Flex-3 infrared cameras and reflective markers attached to the headphone headband. Figure 4.3 shows the audio signal chain for the Ambisonic rendering used in this study.

The sound pressure level in the centre of the loudspeaker array was calibrated

[5] `https://optitrack.com`

for each loudspeaker using an automated SPL calibration script. Based on the pink noise scene stimulus, the SPL for reproduction in the centre of the array was set to 65 dBA. The differences in loudness between the test samples were compensated using an ITU-R BS.1770-3 (ITU-R, 2012) compliant analysis prior to encoding into Ambisonics. The binaural reproduction level was adjusted to match the loudspeaker reproduction level through calibration with a KU100 binaural microphone.

4.1.5 Data Collection

Listening test software for loudspeaker presentation was created using the visual, audio programming environment Max[6]. Headphone-based tests were conducted using dedicated listening test software (see Appendix A) and the DAW Reaper as an audio engine. In both cases, a tablet-based MUSHRA user interface was used.

4.1.6 Post-Screening of Participants

The MUSHRA guidelines specify criteria for post hoc exclusion of participants. This experiment adopted a similar procedure. Participants who rated the hidden reference condition below the score of 80 in more than one trial within each test group were excluded. Table 4.1 shows the number of participants who passed the post-screening.

4.2 Results

This section presents the results of the listening tests. The nonparametric 95% confidence intervals are denoted in the figures by whiskers. The intervals have been computed based on the interquartile range (IQR) using the following equations:

$$IQR = Q_3 - Q_1, \tag{4.1}$$

$$\left[Q_2 - 1.57\frac{\mathrm{IQR}}{\sqrt{n}}, Q_2 + 1.57\frac{\mathrm{IQR}}{\sqrt{n}}\right], \tag{4.2}$$

where Q_1 and Q_3 are the first and third quartiles of the data, Q_2 is the median (second quartile), n is the sample size, and 1.57 is a constant multiplier corresponding to the 95% confidence level. This method, which is used to calculate the size of a notch in a boxplot (McGill et al., 1978), is recommended by the MUSHRA recommendation for visualisation and exploratory data analysis.

4.2.1 General Comparison

Figure 4.4 shows the median scores aggregated across all audio scenes and reproduction methods. This comparison reveals the general relationship between low-bitrate coding, Ambisonic order truncation, and timbral impairments. The anchor conditions were rated as the lowest fidelity (highest timbral distortion), and the hidden reference

[6] `https://cycling74.com/products/max/`

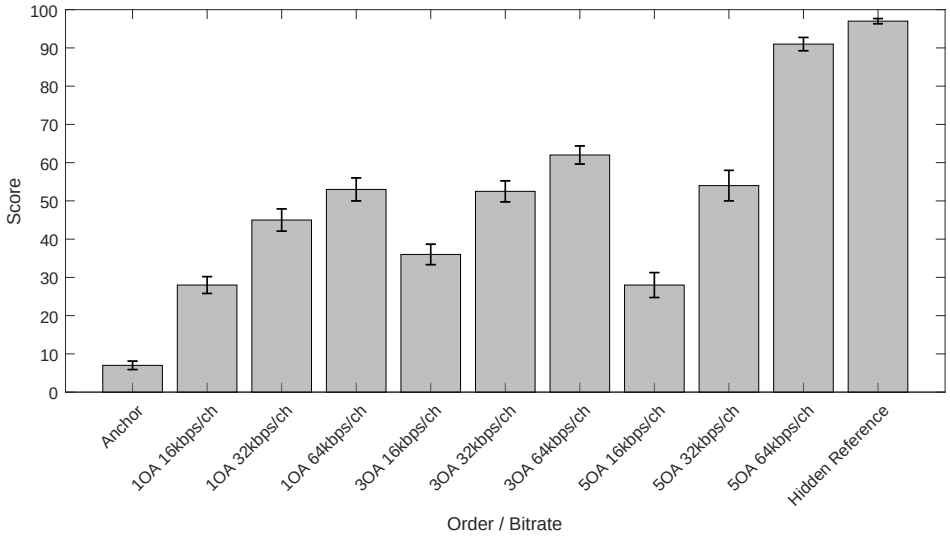


Figure 4.4: Median scores aggregated over all audio scenes and reproduction methods. The whiskers indicate nonparametric 95% confidence intervals.

conditions as the highest fidelity (lowest timbral distortion). It can be seen that the perceived timbral distortion increases with decreasing codec bitrates and Ambisonic orders.

The differences between the median scores of 16, 32 and 64 kbps/ch conditions are similar for the 1st and 3rd-Ambisonic orders. However, the 5th-order scores exhibit higher spread, where the 16 kbps per channel bitrate is rated low, and the scenes encoded using 64 kbps/ch bitrate score very high, close to the uncompressed reference signal. This might be explained by the fact that the 5th-order conditions were not affected by the Ambisonic order truncation, as the reference condition was also a 5th-order scene. At the remaining 1st and 3rd-order conditions, the audio scenes were affected by both low-bitrate coding and truncation.

4.2.2 Audio Scenes

Figure 4.5 shows the median scores for both types of audio scenes aggregated over all reproduction methods. Simple scenes were rated higher than complex ones for the 1st-order Ambisonics conditions. The opposite is observed at the 3rd and 5th-order conditions. This suggests that the ratings were affected by the combined effects of order truncation and low-bitrate coding. As complex scenes consisted of sounds located in multiple directions, the ratings were affected by order truncation more, which caused inaccuracy in rendering binaural cues. Therefore, the low-order complex scenes are rated lower than the simple ones consisting of a single source in front of the listener.

Any impaired condition with a similar distribution of ratings as the hidden reference can be considered perceptually transparent. Only the highest bitrate conditions at 5th Ambisonic order are rated similarly to the hidden reference, suggesting that the codec is perceptually transparent at this bitrate for complex scenes and close to

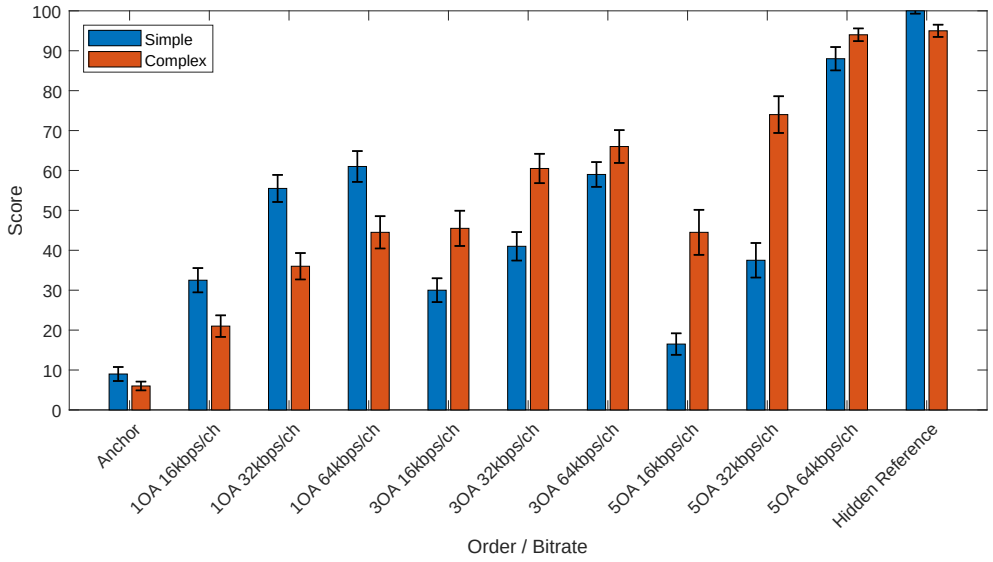


Figure 4.5: Median scores for each audio scene type aggregated over all reproduction methods. The whiskers indicate nonparametric 95% confidence intervals.

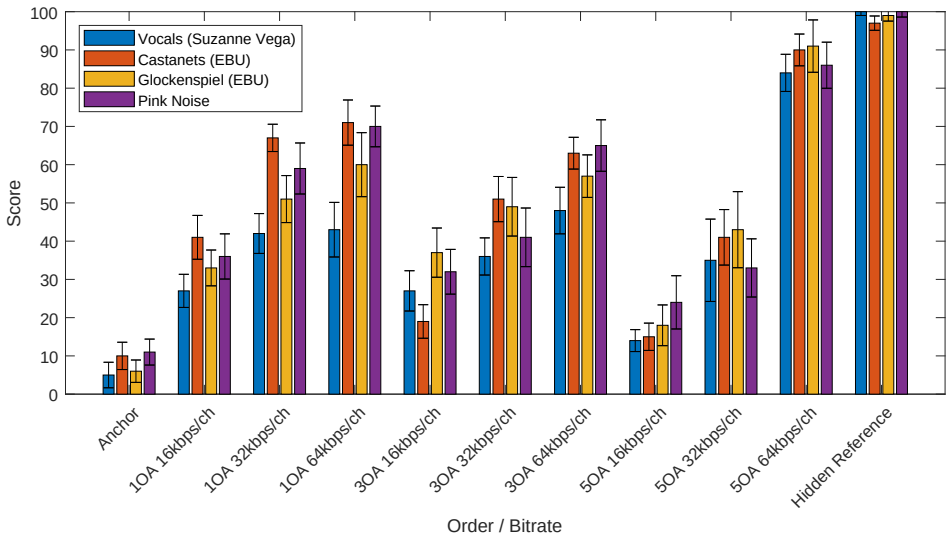
transparent for simple scenes.

Relatively small differences in ratings can be observed within each group of audio content, as shown in Figure 4.6. Although there are some exceptions, e.g. the Castanets scene was rated higher than the Vocals scene.

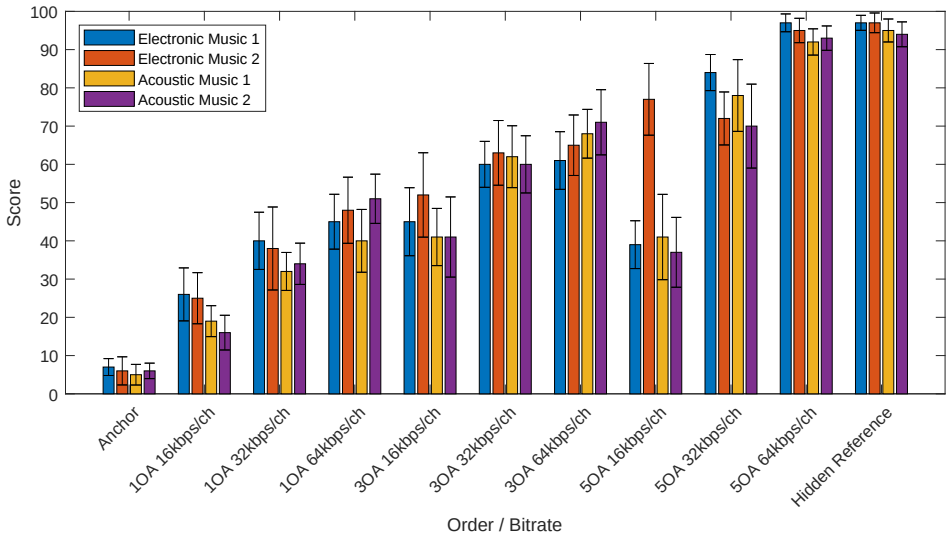
4.2.3 Rendering Methods

Figure 4.7 shows the median scores for each reproduction method aggregated across all audio scenes. Differences in scores obtained using different methods can be observed under specific experimental conditions. In such cases, the scores obtained using binaural reproduction with individual HRTFs are lower than those obtained using loudspeakers or generic HRTFs. This might suggest that the impairment introduced by order truncation and low-bitrate coding is more pronounced when individual HRTFs are used for binaural reproduction compared to the other methods.

Further statistical analysis was conducted using a Wilcoxon rank-sum test to test the hypothesis of equal medians for two independent unequal-sized samples. In this case, the test is used to identify significant differences in the median scores obtained using different reproduction methods in each experimental condition. Table 4.4 shows the p-values obtained for all audio scene ratings, respectively. It can be observed that the median scores obtained using generic and individual HRTFs differ for almost all conditions.



(a) Simple Scenes



(b) Complex Scenes

Figure 4.6: Median timbral fidelity scores for evaluated test conditions grouped by audio scene type. The whiskers indicate nonparametric 95% confidence intervals.

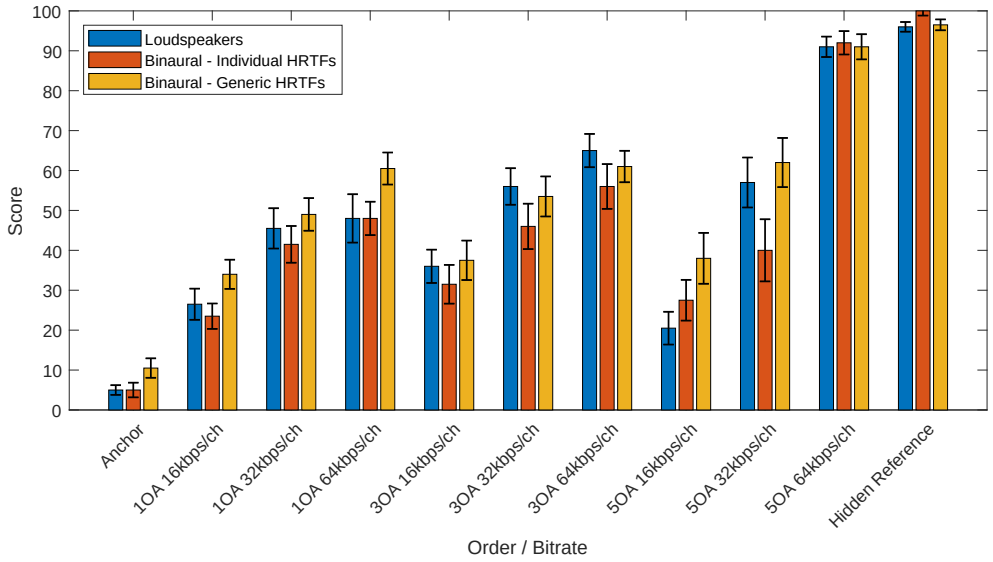


Figure 4.7: Median scores for each reproduction method aggregated over all audio scenes. The whiskers indicate nonparametric 95% confidence intervals.

Table 4.4: p-values obtained using a Wilcoxon rank-sum test for score distributions obtained using different reproduction methods. LSPK, BINGEN, and BININD denote loudspeaker reproduction, binaural reproduction using generic HRTFs and binaural reproduction using individual HRTFs, respectively.

	LSPK / BININD	LSPK / BINGEN	BININD / BINGEN
Anchor	0.5407	0.0000	0.0000
1OA 16kbps/ch	0.1752	0.0011	0.0000
1OA 32kbps/ch	0.2067	0.1727	0.0049
1OA 64kbps/ch	0.6667	0.0130	0.0003
3OA 16kbps/ch	0.0532	0.7646	0.0307
3OA 32kbps/ch	0.0001	0.1651	0.0141
3OA 64kbps/ch	0.0082	0.1457	0.2029
5OA 16kbps/ch	0.2098	0.0000	0.0049
5OA 32kbps/ch	0.0179	0.1374	0.0002
5OA 64kbps/ch	0.5496	0.9367	0.6343
Hidden Reference	0.0536	0.6013	0.0155

4.3 Discussion

The results obtained in this study clearly show the relationship between timbral distortion, order truncation and codec bitrate. A significant difference was observed between the rating scores for simple and complex scenes, suggesting that Ambisonic order does not affect the timbral fidelity of simple scenes as much. A likely explanation for this is that the single sound source in the simple scene tests was always panned at a location straight in front of the listener as opposed to multiple sources panned at various locations in the complex scenes. Therefore, the rendering of that source was not affected by the binaural cue distortions introduced by order truncation. Future revisions to the test procedure could avoid this by panning the source to directions distributed across the sphere.

For complex scenes, reducing both Ambisonic order and bitrate increases timbral distortions. However, the 1st-order condition at the highest bitrate per channel and the 3rd-order condition at the lowest bitrate per channel received similar ratings, although they share the same total bitrate of 256 kbps. As a typical Ambisonic recording would be a complex scene, this study suggests that it is possible to implement streaming of 3rd-order Ambisonics using a lower bitrate in place of 1st-order Ambisonics at a higher bitrate, preserving the same total bitrate and perceived level of timbral impairment.

Significant differences in median scores have been identified between different reproduction methods for all conditions except the one with the highest bitrate at the 5th-order. Scores obtained using binaural reproduction with individual HRTFs exhibit the highest spread between conditions, suggesting that this type of reproduction results in the best discrimination of timbral distortion. Possible explanations for this include the following: generic HRTF rendering reduces the ability to perceive timbral distortion differences, and rendering using loudspeakers is affected by the acoustics of the listening environment reducing the clarity of sound.

Further studies should include uncompressed reference conditions at respective orders to disentangle the effects of order truncation and low-bitrate coding.

4.4 Conclusion

This study evaluated the perceived timbral distortion in Ambisonic audio compressed with the Opus codec using the Channel Mapping Family 3. A strong relationship has been found between the codec bitrate, order truncation, and timbral fidelity. The results suggest that the user experience would significantly improve with spatial audio streaming services implementing at least 3rd-order Ambisonics over 1st-order. The Opus codec with Channel Mapping Family 3 enabled allows streaming of complex Ambisonic content with fair quality using a reasonable 256 kbps total bitrate.

The results suggest that using binaural reproduction with individual HRTFs leads to better discrimination of the perceived timbral distortions than the other two methods used. By focusing on timbral fidelity, this work provides a basis for further research on other aspects of the spatial audio quality of low-bitrate compressed Ambisonic scenes.

Chapter 5

Auditory Localisation in Bitrate-Compressed Ambisonic Scenes

The previous chapter evaluated the perceived timbral distortion in Ambisonic audio compressed with the Opus codec. However, perceptual audio coding can also introduce spatial distortion leading to the degradation of localisation cues. This chapter's primary focus is the evaluation of auditory localisation within Ambisonic scenes with respect to perceptual low-bitrate coding and different reproduction methods. Specifically, subjective differences between Ambisonic scenes encoded with the Opus codec at different bitrates and Ambisonic orders are investigated regarding the localisation precision of virtual sound sources presented over loudspeakers and headphones.

5.1 Auditory Localisation in Ambisonics

The relation between Ambisonic order and auditory localisation error has been researched in several experiments. Previous studies used synthesised Ambisonic scenes as well as scenes recorded with Ambisonic microphones (Braun and Frank, 2011; Bertet et al., 2013) reproduced using loudspeaker arrays (Power et al., 2012) and binaural rendering (Thresh et al., 2017). It has been shown that localisation error depends on the Ambisonic order as well as the incidence of the virtual sound source. Furthermore, binaural reproduction produces more front-back confusion errors than loudspeaker-based reproduction (Thresh et al., 2017). Both loudspeaker and headphone reproduction methods have not been directly compared using Higher Order Ambisonics and individual HRTF-based rendering.

There is currently a limited amount of research published on the quality of compressed spatial audio, particularly on the compression of first- and higher-order Ambisonics. The recent version of the Opus codec (Valin et al., 2013) implements Channel Mapping Family 3, which allows for Ambisonic signal coupling (Skoglund and Graczyk, 2018). A brief characterisation of Opus channel mapping families is provided in Section 3.4.7. Previous work by Narbutt et al. (2017, 2018) includes

subjective evaluation of Ambisonics compressed with Opus 1.2 codec with Channel Mapping Family 2 implementation and the development of a reference objective spatial audio quality metric. They use a MUSHRA paradigm to assess the localisation degradation and demonstrate quality degradation between equivalent bitrates at different orders. The absolute extent of localisation precision is not shown. These studies also focus on static and generic HRTF binaural listening conditions. The localisation performance within Ambisonic scenes compressed with the Opus codec using Channel Mapping Family 3 has not been researched.

5.2 Methods

The purpose of the experiment presented in this chapter was to subjectively assess the spatial distortion introduced by Ambisonic order truncation and perceptual coding of the Ambisonic scenes using different bitrates. The method of adjustment (H. Langendijk, 1997) was used for the auditory localisation tests. Participants were asked to move an artificially reproduced virtual acoustic pointer to the perceived direction of a reproduced target sound source using a physical controller, shown in Figure 5.1. The audio playback of the pointer and target scenes was controlled by the participants and programmed to ensure that both stimuli were never presented simultaneously. The azimuth and elevation step encoders adjusted the rotation of the rendered acoustic pointer with a single-degree precision. It is important to note that the experiment was designed to examine the perceived differences between the uncompressed and low-bitrate compressed scenes, not the absolute localisation error. Assessing the relative localisation of virtual sound sources removes the need for using real sound sources as the localisation anchors. Therefore, evaluation of binaurally rendered signals reproduced using ear-occluding headphones is possible.

Listening test software for loudspeaker presentation was created using the visual, audio programming environment Max[1]. Headphone-based tests were conducted using dedicated listening test software (see Appendix A) and the DAW Reaper as an audio engine.

[1] `https://cycling74.com/products/max/`

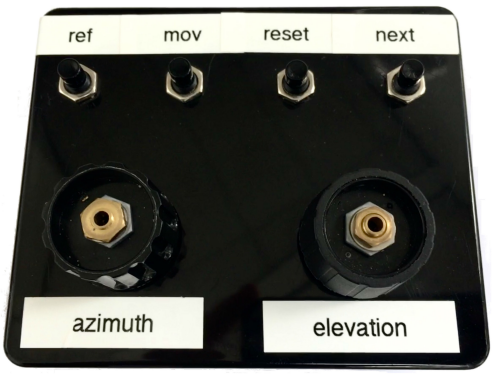


Figure 5.1: Physical controller designed for the auditory localisation test.

Table 5.1: Target sound source directions during the localisation performance test.

Direction	1	2	3	4	5	6
Azimuth (°)	0	180	72	-36	-144	108
Elevation (°)	90	-18	18	-18	18	-18

5.2.1 Test Stimuli

The acoustic pointer consisted of a one-second pink noise burst encoded into 5th-order Ambisonics, appropriate to the spatial resolution of the Ambisonic reproduction systems used in the experiment. The low bitrate compressed scenes presented during the simple scene evaluation consisted of one-second pink noise bursts placed in six static target directions: above, behind and on the sides of the listener. The coordinates of the investigated directions are listed in Table 5.1. These directions were chosen to match the context of the experiment, that is, 360° video streaming, where spatial audio is often used to direct users' attention in the virtual space.

The complex scene stimuli consisted of reference pink noise bursts and a modified Ambisonic soundscape from the Eigenscape dataset (Green and Murphy, 2017) recorded with a 4th-order Ambisonic microphone[2]. The HOA microphone signal was used to provide different input signal conditions for the Opus codec compared to the simple scene material and to mimic the typical audio content of 360 videos. The used excerpt of the forest soundscape did not include any prominent spatially defined sounds that could influence the perception of the target sound direction. The level of the soundscape was empirically adjusted to prevent participants from being significantly distracted from the task. Because the spatial resolution of the Ambisonic soundscape was limited to the 4th-order, additional 5th-order background noise was added. This 5th-order noise signal was synthesised using 36 uniformly distributed virtual loudspeakers fed with decorrelated Brownian noise, which was chosen because its power spectrum differs from the pink noise used as the virtual sound source in this experiment. The target sound sources consisting of pink noise bursts (4 times a repeated sequence of 2 s burst with 250 ms rise and fall times followed by 500 ms of silence) were panned at the specified directions shown in Table 5.1.

The required 1st and 3rd-order test stimuli were extracted from the 5th-order simple and complex scenes and subsequently compressed using Opus encoder at different bitrates and Channel Mapping Family 3 enabled. The resulting test stimuli set consisted of 60 scenes for simple and complex scene tests (a multiplication of six target sound source directions and ten system conditions). The investigated system conditions are shown in Table 5.2. The uncompressed 5th-order Ambisonic scenes were used as the reference condition.

5.2.2 Spatial Audio Rendering

The audio rendering chain employed in this experiment follows the one described in Section 4.1.4. The sound pressure level for simple scene stimuli reproduced in the

[2] `https://mhacoustics.com/products`

Table 5.2: Investigated bitrates (kbps) at different Ambisonic orders.

	Bitrate per channel	Total bitrate		
		1OA	3OA	5OA
Compressed	16	64	256	576
Compressed	32	128	512	1152
Compressed	64	256	1024	2304
Uncompressed	768			27648

Table 5.3: Number of participants who completed the tests grouped by the rendering method and audio content type used.

Reproduction Method	Content Type	Number of Participants
Loudspeakers	Simple	21
	Complex	16
Binaural (Individual HRTFs)	Simple	15
	Complex	14
Binaural (Generic HRTFs)	Simple	19
	Complex	19

array was set to a target level of 65 dBA. The loudness of the complex scene stimuli for the localisation test was subjectively aligned to match the loudness of the simple scene test. The binaural reproduction level was adjusted to match the loudspeaker reproduction level through calibration with a KU100 binaural microphone.

5.2.3 Participants

The experimental group consisted of MSc and PhD audio engineering students and senior researchers with experience in critical listening. Some participants took part in the sound quality assessment tests for the first time. All participants were instructed on how to perform the tests by reading an information sheet and receiving individual demonstrations. The localisation test included a training phase consisting of three test tasks with uncompressed stimuli. The responses gathered during the training were not exported for further analysis. Participants were instructed to keep their heads in the centre of the loudspeaker rig, look towards the front of the loudspeaker rig and limit head movements, although their heads were not physically constrained. The limitation of head movements was to ensure that the target virtual sound sources were located around the participant's head and the participants didn't rely exclusively on dynamically changing interaural cues to localise both target and pointer sources throughout the test.

All participants gave their informed consent to be included in the study. The protocol was approved by the Physical Sciences Ethics Committee of the University of York (approval code: Rudzki021018). Table 5.3 shows the number of participants who completed the tests, separated into subgroups by the reproduction method and audio content type used.

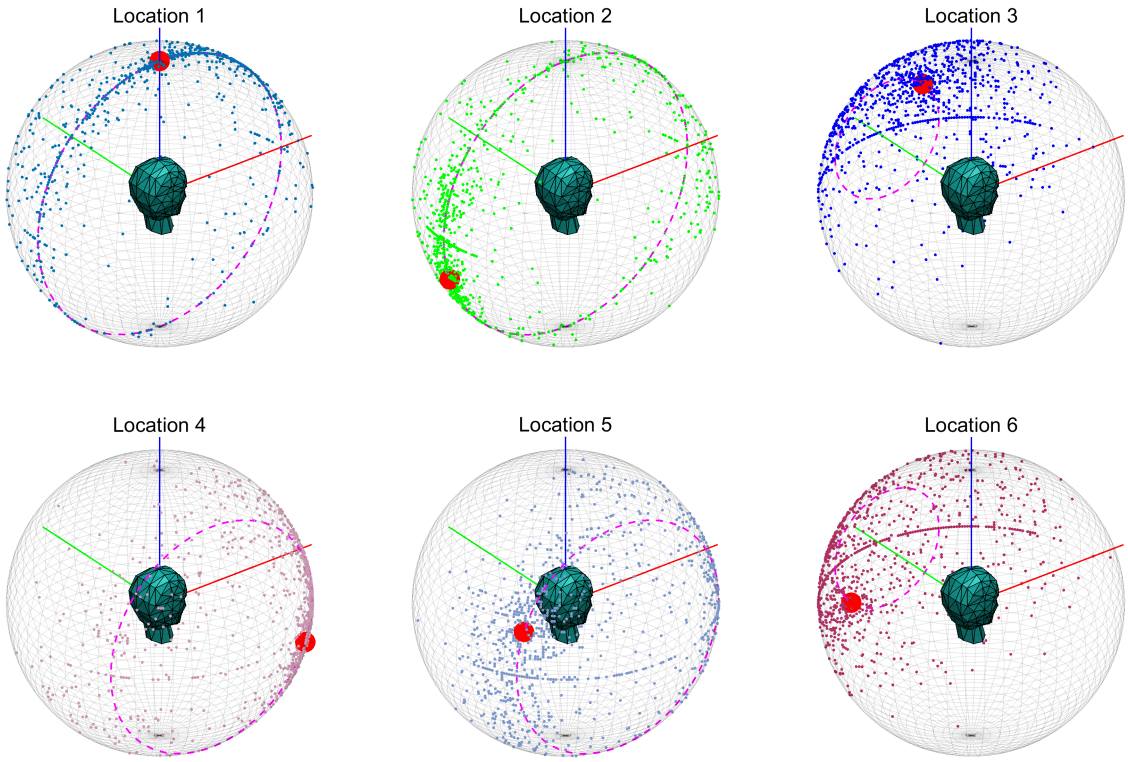


Figure 5.2: Distributions of the acoustic pointer directions recorded across all experimental conditions and participants grouped by target source direction. The red dots symbolize the directions of the target sound sources. The dashed magenta circles represent the respective cones of confusion on the sphere. The axes denote directions relative to the listener: red – front, green – left, blue – top.

5.3 Results

The collected directional data represents the auditory localisation of the virtual acoustic pointer adjusted to match the perceived direction of the virtual target sound sources. Each listening test consisted of 60 individual tasks. All investigated low-bitrate compression conditions were compared within each of the six subgroups.

Figure 5.2 shows the distribution of the acoustic pointer directions on the sphere set by the participants. Each sphere represents data gathered for different directions of the target sound source presented using all three reproduction methods at different Ambisonic orders and compression bitrates. Figure 5.3 shows the same data plotted using equirectangular projection.

The initial direction of the pointer for each task was set in front of the listener. Since participants operated the azimuth and elevation controls (see Figure 5.1) independently, it can be seen that the distributions of the recorded pointer directions are slightly skewed toward the horizontal and median planes. This suggests that the responses may have been affected by the collection method. Virtual sound source target directions one and two correspond to the median plane directions of incidence. The respective recorded acoustic pointer indications are distributed close

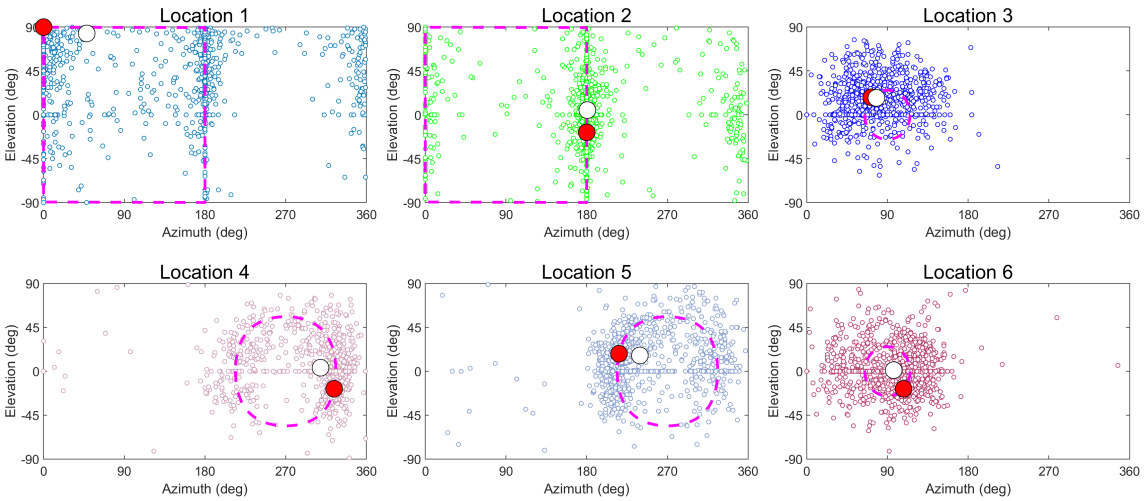


Figure 5.3: Distributions of the acoustic pointer directions recorded across all experimental conditions and participants grouped by target source direction. The red dots symbolize the directions of the target sound sources. The dashed magenta lines represent the respective cones of confusion. The unfilled circles represent the mean direction of the pointer directions.

to the intersection of the median plane with the unit sphere. The listener can match the perceived elevation of these sources and the acoustic pointer using exclusively spectral and dynamic localisation cues. Directions four and five correspond to slightly elevated and laterally shifted directions. Collected acoustic pointer indications are distributed along the intersection of respective cones of confusion with the unit sphere. Directions three and six correspond to slightly elevated and strongly laterally shifted directions close to the interaural axis. Based on the visual observation, the distributions of pointer indications are relatively concentrated around the target sound source directions with a slight skew towards the respective cones of confusion.

Further analysis was conducted using the great-circle distance, which can be calculated as the shortest angular distance between each pointer and the corresponding target directions on the unit sphere. Analysis of the horizontal and vertical localisation error components was performed. However, the differences between codec conditions were less evident than when using the combined error metric. To minimise the directional bias introduced by the pointing interface and focus on the random localisation error (localisation precision, not accuracy), the great-circle distance was calculated between pointer locations and the mean pointer direction within each analysed subset.

Figure 5.4 shows the set of probability density functions (Shimazaki and Shinomoto, 2010) of the localisation error obtained experimentally at different Ambisonic orders/codec bitrates and rendering methods, whereas Figure 5.5 shows the set of probability density functions of the localisation error for different types of reproduction methods and audio scene types. It can be seen that the general shape of the presented distributions corresponds to the shape of the von Mises-Fisher distribution (Fisher et al., 1993) plotted as the probability density function of the distance between each spherical mean and each sample. However, the experimental

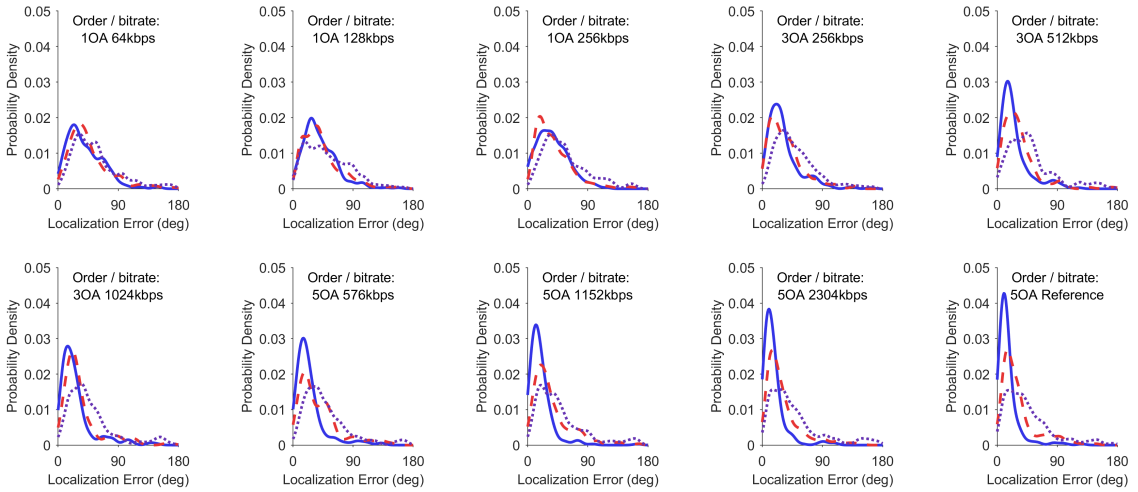


Figure 5.4: Probability density functions of the localisation error for different spatial audio rendering methods at different codec bitrates and Ambisonic orders. The continuous line represents loudspeaker reproduction, the dashed line denotes binaural with individual HRTFs, and the dotted one denotes binaural with generic HRTFs. The kernel bandwidth of 6° was chosen empirically to show differences in estimated distributions.

distributions exhibit multi-modal characteristics caused by the cone of confusion and data collection biases. This limits the use of statistical tests based on parameterised spherical data distributions for a unified analysis of the results. Instead, the Kruskal-Wallis rank-based nonparametric test was used to investigate the spherical concentration (Kruskal and Wallis, 1952; Verdebout, 2015) of participant responses under different experimental conditions.

The following experimental variables were tested: participants, virtual target sound source direction, codec bitrates, Ambisonic orders, audio content type, and audio reproduction method (loudspeaker-based vs. individualised vs. generic HRTFs). A significant difference has been found in the overall localisation task performance between participants in each of the three experimental phases using different spatial

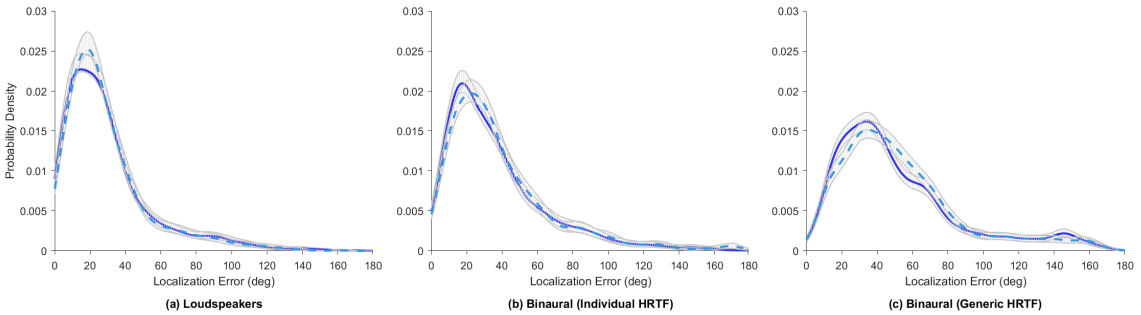


Figure 5.5: Probability density functions of the localisation error for different types of reproduction methods and audio scene types. The continuous line denotes simple scenes and the dashed line denotes complex scenes. The kernel bandwidth of 6° was chosen empirically.

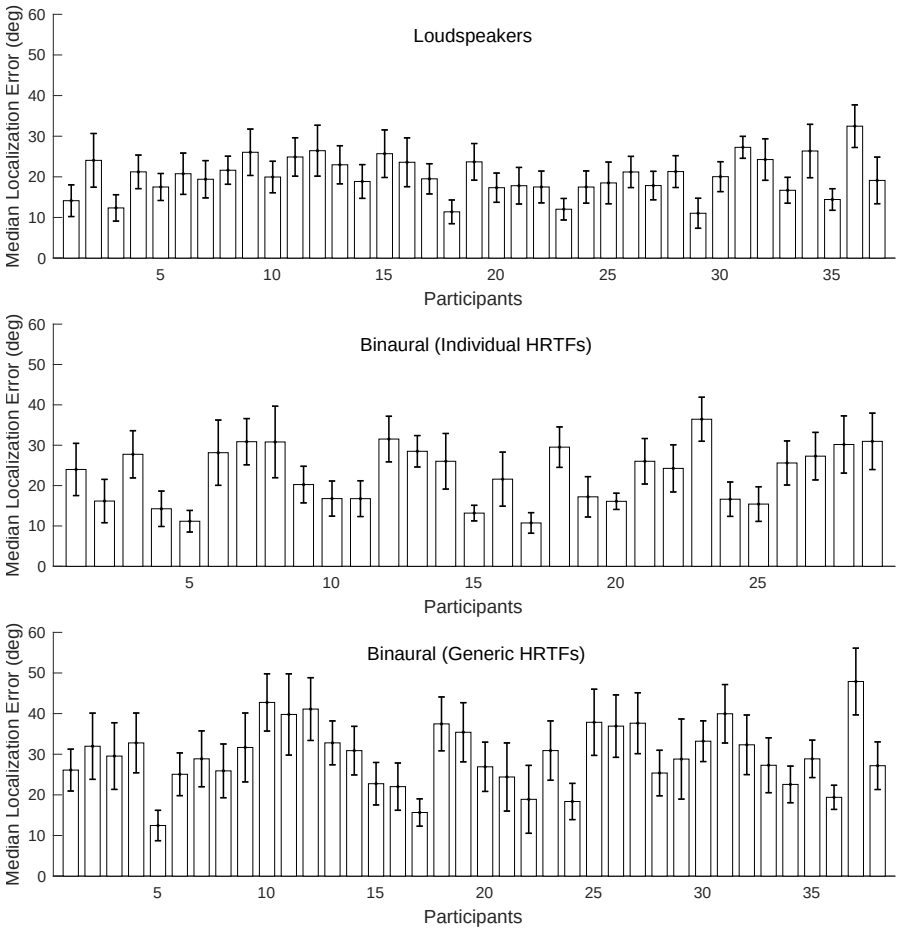


Figure 5.6: Median localisation error of each participant at different reproduction methods. The whiskers indicate nonparametric 95% confidence intervals.

audio rendering methods ($p < .01$). The overall localisation error median for each participant is shown in Figure 5.6. The effect of the position of the virtual target sound source was significant ($p < .01$). However, no clear trends in data were identified. The localisation error median for each direction at different rendering methods is shown in Figure 5.7. The effect of codec bitrate was analysed in nine subgroups, grouped by Ambisonic order and rendering method. It was found to be significant ($p < .01$) in two groups: 3rd-order and 5th-order scenes reproduced using the loudspeaker array. Significant differences have been found between compressed scenes grouped by Ambisonic order for each of the three rendering methods ($p < .01$). The effect of scene type on participant responses was investigated in the raw data, and the test result was close to the 95% confidence limit ($\chi^2 = 3.92, p = .048$). Detailed analysis was carried out in 30 subgroups, grouped by different codec bitrates and Ambisonic orders, and spatial audio rendering methods. The difference between simple and complex scene content has not been found to be significant ($p > .01$) in 29 of the 30 subgroups. The rendering method significantly affected the localisation error ($p < .01$).

Figure 5.8 shows the median localisation error at different codec bitrates, Am-

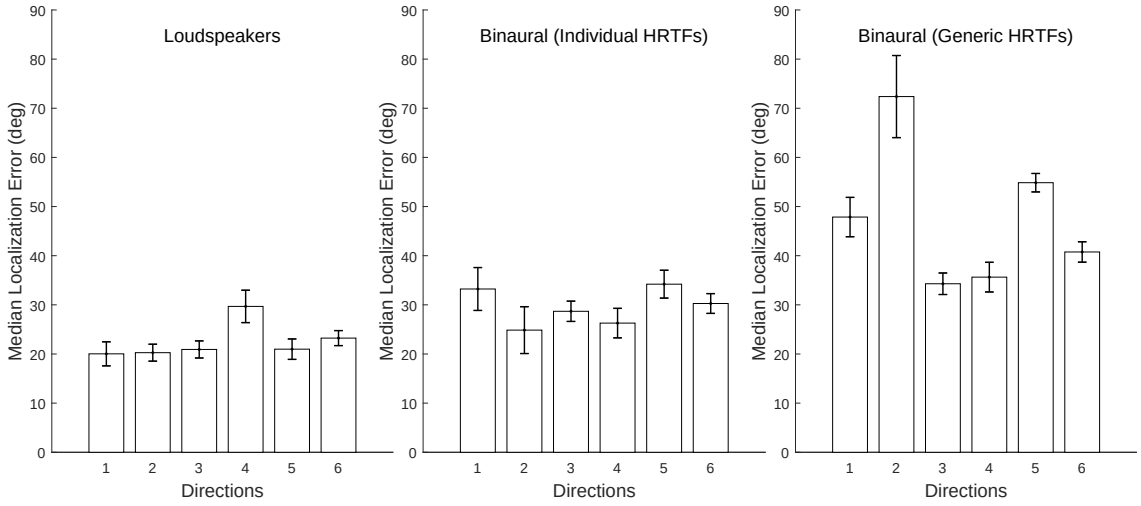


Figure 5.7: Median localisation error for each virtual sound source direction at different reproduction methods. The whiskers indicate nonparametric 95% confidence intervals.

bisonic orders, and rendering methods. It can be seen that the generic HRTF reproduction resulted in higher localisation errors compared to the loudspeaker-based and individual HRTF reproduction. A decrease in localisation error was observed with the increase of Ambisonic order using all three reproduction methods. This effect is most prominent in the loudspeaker-based tests, where the difference between 1st- and 3rd-order is much more significant than the difference between 3rd- and 5th-order. The differences in median localisation error caused by different bitrates within the same Ambisonic order can be observed. However, they are not significant in most cases. We can infer from the results that the localisation precision depends slightly on the bitrate (within the examined bitrate values).

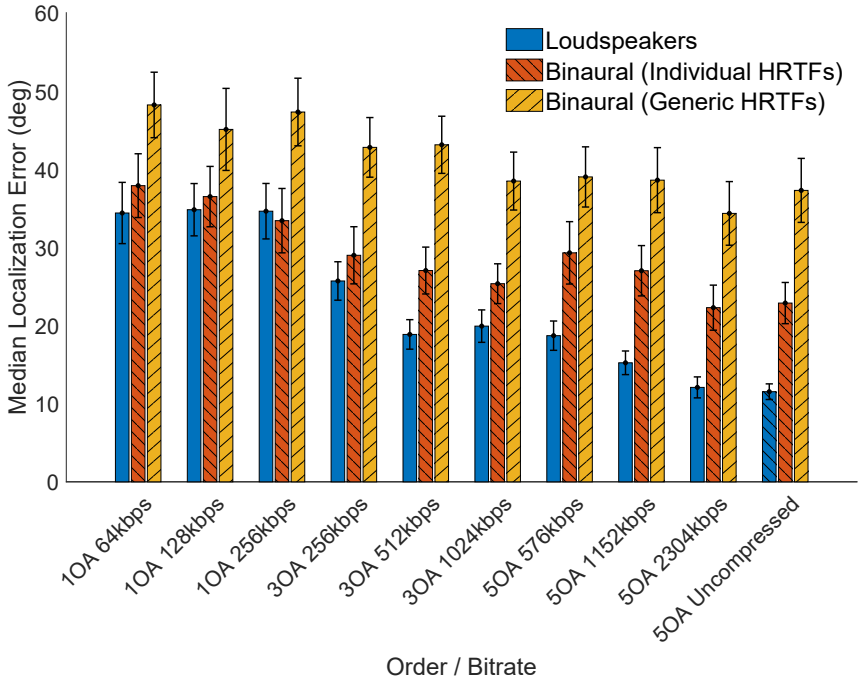


Figure 5.8: Median localisation error at different codec bitrates and Ambisonic orders. The whiskers indicate nonparametric 95% confidence intervals.

5.4 Discussion

Multiple factors have contributed to the localisation error measured during the experiment. Firstly, the limited resolution of the human auditory localisation. This study was focused on directions where the localisation blur is relatively high, which may have contributed to the high variance in the responses. The measured median localisation error for the 5th-order uncompressed reference reproduced over loudspeaker array was about 11°, comparable with other studies focusing on localisation in horizontal and vertical planes combined (Makous and Middlebrooks, 1990).

Secondarily, the limited-order Ambisonic representation of the target and acoustic pointer scenes has contributed to the localisation error. The results prove that the Ambisonic order largely defines localisation precision, as higher orders present more precise spatial resolution. The median localisation error for the 1st-order scenes was about 34° corresponding to the results obtained by Braun and Frank (2011) for the virtual sound source presented over the loudspeaker array. The localisation error obtained in the same study for the 4th-order Ambisonic virtual sources was about 10°, which corresponds indirectly to our 5th-order uncompressed condition result. In the experiment by Bertet et al. (2013), the localisation error for 4th-order Ambisonic virtual sources measured in the horizontal plane varied from 5° to 14° at the lateral positions. It is important to note that the current experiment used virtual sound source presentation for both target and pointer sounds to facilitate the headphone-based tests, whereas the study by Bertet was carried out with an Ambisonic pointer and real sound source target. A direct comparison of the results

with any of the referenced studies is impossible due to the differences in reproduction systems and test frameworks.

Another source of error comes from the limitations of the reproduction methods used. Participants in the loudspeaker tests localised the sound sources with the highest precision. Similar precision was obtained using binaural reproduction with individually measured HRTFs at 1st Ambisonic order. At 3rd- and 5th-order, the localisation error in headphone-based tests was higher than in the loudspeaker-based phase. Binaural reproduction of Ambisonic scenes employing the generic HRTF set resulted in the highest localisation error at all tested signal conditions. This phenomenon requires further investigation, as the loudspeaker and headphone test data were obtained with different groups of participants. It is worth noting that the lowest bitrate in 3rd-order Ambisonic presentations (256 kbps) produced a significantly improved localisation precision in the loudspeaker case than the highest bitrate condition for 1st order (also 256 kbps). In both headphone listening cases, there is no significant difference between the aforementioned bitrates and orders. These results contradict those found by Narbutt et al. (2018), which show a significant degradation with the lowest bitrate at 3rd-order. These differences might be attributed to the use of head-tracking, rather than static binaural presentations, and different localisation-performance test paradigms.

The results of this study show that auditory localisation in low-bitrate compressed Ambisonic scenes is not significantly affected by codec parameters. Although the differences between localisation errors for the same orders were not statistically significant, based on the visible trends (see Figure 5.8) localisation precision degrades slightly with a decrease in bitrate. The study presented in Chapter 4, which focused on timbral fidelity of the Opus compressed spatial audio, revealed significant differences between bitrates and Ambisonic orders.

The effect of an additional soundscape present in test stimuli was investigated. However, no significant difference was observed between simple and complex content presentations. The 5th-order spatially diffused sound scene was added to investigate if the spatial distortion of the single sound source presentation within the scene will be affected by feeding additional non-directional information to each encoded channel. This condition was supposed to mimic a recording done with an Ambisonic microphone, although maintaining the highest possible spatial resolution of the single sound source by synthesising the Ambisonic sound field. The impact of the sound scene complexity on the localisation error has not been revealed.

The chosen data collection method might have contributed to a high variance in the participants' overall localisation performance. The median time to complete each task by a participant was about 26 seconds. The average duration of the test session was about 45 minutes, with a single break in between. Some participants reported mild psycho-physical fatigue after the experiment, suggesting that responses collected using a less challenging test methodology could give more consistent results. As the participants' heads were not constrained and the optical tracking system did not record the head movements, it is unknown to what degree the small head rotations contributed to the measured localisation performance.

Another factor affecting localisation precision in the experiment is the acoustic

pointer response collection method. The auditory localisation precision measured using the acoustic pointing method may give higher error estimates than the source discrimination methods used for the MAA measurements, which is focused more on the change of the perceived acoustic signal rather than spatial analysis of the sound field (Makous and Middlebrooks, 1990). However, once the virtual pointer and target directions are perceptually matched, participants might compare both signals using features other than perceived spatial locations. The degree to which the presented mechanism has contributed to the experimental results remains unknown. Further studies should consider different response collection techniques adequate to the proposed application of the coding system.

The continuation of research should look into developing efficient indirect perceptual evaluation methods for assessing binaural-based spatial audio systems, including bitrate compression schemes. Given the importance of the frontal region for immersive content consumption in VR and AR and teleconferencing services, the localisation accuracy and precision should be investigated for this region in more depth.

5.5 Conclusion

The study presented in this chapter builds upon the study presented in Chapter 4, focusing on the localisation precision of binaural-based Ambisonic reproduction using low-bitrate compression over different Ambisonic orders. The tests were conducted using headphone-based reproduction employing individualised and generic HRTF sets as well as loudspeaker-based presentations for comparison. The results suggest that strong bitrate compression will not affect the auditory localisation in scenes encoded using Opus compression compared to uncompressed Ambisonic presentations.

The results suggest that using higher-order Ambisonic content instead of 1st-order Ambisonics will improve localisation within the scenes, especially when using personalised binaural rendering or multi-loudspeaker reproduction.

Chapter 6

Context-Based Evaluation of the Opus Audio Codec for Spatial Audio Content in Virtual Reality

Chapter 4 and Chapter 5 evaluated Opus compression using different rendering methods. However, they did not include a variety of audio stimuli. Streaming of VR content poses a challenge to existing compression schemes, as content can vary from simple 360° non-interactive soundscapes to fully interactive occupational training simulations. Such diverse contexts may dictate spatial audio processing requirements, e.g. higher bitrates of compression or specific compression algorithms. The Opus codec is one such algorithm that caters for a wide variety of audio applications, from Voice-over-IP to streaming live music performances, and remains the focus of the study presented in this chapter.

The study presented in Chapter 4, which also evaluated the codec using the MUSHRA paradigm, was carried out in a standard non-immersive listening room environment using a 2D tablet-based data collection method without any accompanying visual stimulus. However, the consumption of VR content differs from such conditions. Therefore, the study presented in this chapter was conducted using a VR environment containing synchronised audio and visual stimuli, making the evaluation conditions more similar to the intended use case.

Previous studies by Narbutt et al. (2017, 2018) focused on two perceptual attributes: Listening Quality and Localisation Accuracy for Opus-compressed Ambisonic stimuli, while the studies described in Chapter 4 and Chapter 5 focused on timbral distortion and localisation precision respectively. Although this study builds on the previous research, it aims to evaluate a single perceptual attribute, Basic Audio Quality (BAQ), which encompasses both the timbral and spatial aspects of perceived quality impairments. The tested codec parameters are bitrate and channel mapping family. Previous studies evaluated either Channel Mapping Family 2 or 3, while this study directly compares the two. A brief characterisation of the Opus channel mapping families is provided in Section 3.4.7.

Another difference between the current study and the ones described in Chapter 4 and Chapter 5 is that this study does not evaluate the Ambisonic order truncation.

Each experimental trial encompasses a reference condition of the same Ambisonic order as the evaluated conditions. This ensures that the only system evaluated in the study is the codec used in different contexts.

6.1 Methods

This section presents the methods for designing and implementing the carried out multi-stimulus listening test, including content creation. The study is based on the ITU-R BS.1534-3 (MUSHRA) (ITU-R, 2015b) recommendation but does not strictly follow it, as the recommendation was never designed for head-tracked VR presentations.

6.1.1 Content Rationale

Four different contexts were evaluated: gaming, music, soundscapes and teleconferencing. Firstly, gaming is arguably the most common scenario in which VR is experienced, demonstrated by the sheer number of VR headsets marketed as accessories to gaming consoles or as independent, all-in-one gaming devices.

Ambisonics has been useful for delivering game audio, e.g. in racing simulators (Deleflie and Goodwin, 2007). Opus can be used for compressing game assets and streaming networked virtual gaming/VR experiences such as interactive Metaverse concerts or viewing of e-Sports, not to mention the widespread streaming of recorded game content on YouTube.

Music is another context in which spatial audio has entered the market. For example, Apple currently supports the Dolby Atmos spatial audio format on its music streaming platform[1]. Music listening is likely to become a more prevalent context within VR.

Soundscapes are another important context to evaluate. Not only have they been used in previous studies and evaluations of spatial audio in conjunction with VR (Fela et al., 2022), but soundscape-based virtual environments have been used in health and well-being applications, for example, to measure the behavioural response in children with Autism Spectrum Disorder (ASD) (Johnston et al., 2019a), therapeutic treatment (Wang and Anagnostou, 2014), disaster awareness (Fino et al., 2017) and as educational tools (Max and Burke, 1997).

Teleconferencing has become an integral part of everyday life for many people in recent years since the COVID-19 pandemic. Accompanying this shift in the way colleagues interact comes with many challenges that impede the efficiency and naturalness of conversation. For instance, latency can cause participants in a teleconference to talk over each other, while lack of any sort of spatial dimension or feeling as though other participants are not sharing the same physical space can cause other problems, such as distraction or disinterest. VR teleconferencing may help improve, to some extent, the key characteristics of media naturalness discussed by Karl et al. (2022). The use of spatial audio in VR for teleconferencing could

[1] `https://support.apple.com/en-gb/HT212182`

help to improve co-location by immersing participants within the same perceived environment, VR environments could help with facial expression and body language by giving participants full body avatars; spatial audio could also help with speech intelligibility by making it easier to discern who is talking, and finally, the issue of synchronicity can be improved by making the streaming as efficient as possible.

6.1.2 Content Production

Despite the fact that the test scenes represented the different contexts of gaming, music, soundscapes and teleconferencing, to facilitate a fair comparison between the contexts and test conditions, the material had to be pre-rendered. This meant that apart from three-degrees-of-freedom (3DOF) head-tracked rotation of the scenes, no other interaction was facilitated for this study. In other words, participants could not actually play the game in game scenes, play instruments or sing in music scenes, or speak or interact with the avatars in teleconference scenes. Therefore, all scenes were 360° videos representing specific contexts in which the MUSHRA-style test interface could be overlaid.

The foundations of the scenes were created using the Unity[2] platform. JavaScript Object Notation (JSON) files were used to export information about the sound-emitting objects, such as their position relative to the listener at time intervals. The 360° videos were rendered using Unity's proprietary video recorder.

Since 1st, 3rd and 5th-order Ambisonic scenes were evaluated in this study, synthesised Ambisonic tracks were made by spatialisation of monophonic tracks. Without upmixing lower-order Ambisonic content, this was the only way to produce original 5th-order Ambisonic content, as such high-order Ambisonic microphones were not commercially available. Once a single channel track was created for each of the objects in a scene, the positional data from the respective JSON files were used to spatialise the tracks using Ambisonics.

The monophonic tracks were produced as WAV files at 48 kHz sample rate and a 16-bit depth. Some of them were produced entirely using VST plugins in Logic Pro X, whilst others incorporated royalty-free sounds. All test stimuli that were produced for this study, Ambisonic audio files, 360° videos, and a tracklist of sound effects for all outsourced audio have been made publicly available for download (Lee, 2022).

In order to read the mono WAV files and corresponding JSON files for each object, a Max[3] patch was used. Max objects can be used to run plugins, and this was necessary in order to spatialise the monophonic tracks through the use of IEM's Ambisonic plugin suite[4]. The Room Encoder allowed for the simulation of early reflections and the Stereo Encoder for general Ambisonic encoding.

1st, 3rd and 5th-order Ambisonic scene files were created for each of the eight scenes mentioned in Section 6.1.3. Before compressing with Opus, the loudness of all Ambisonic scenes was normalised to an arbitrary value of -31 LUFS. The loudness

[2]Unity – `https://unity.com`

[3]Max – `https://cycling74.com/products/max/`

[4]IEM Plug-in Suite – `https://plugins.iem.at`

was measured based on binaurally rendered audio. The rendering used in this step followed the rendering chain used in the listening test described in Section 6.1.7.

6.1.3 Description of Scenes

Two different scenes were made for each of the four contexts. All scenes had a duration of 12 seconds and are described as follows:

- *GameCar*: First-person perspective of a driving game in which the participant is in the driving seat. Different coloured cars race while colliding with each other and the walls.
- *GameFPS*: First-person shooter-style game where the participant is in a room with an alien and a UFO as enemies. A gun can be seen shooting at enemies until they are destroyed.
- *MusicBlues*: Scene in which the participant is in a room surrounded by various sounding instruments, e.g. piano, brass and percussion.
- *MusicMallets*: Scene in which the participant is in a room surrounded by various sounding instruments, e.g. keyboard, vibraphone, strings and percussion.
- *SoundscapeFarm*: Scene in which the participant is in a farmer's field, surrounded by various low poly objects: environmental objects, stationary sound-emitting farm animals and a moving tractor.
- *SoundscapeOasis*: Scene in which the participant is in a desert oasis. The participant stands next to water and camels. There are also some nearby palm trees and a propeller plane passing in the distance.
- *TeleconferenceOne*: Scene in which the participant is at a table with four animated mannequin avatars having a discussion.
- *TeleconferenceTwo*: Scene in which the participant is at a table with three animated mannequin avatars having a discussion.

The test soundtracks were intentionally not made over dense with sound effects to reduce the cognitive load of the user and to allow them to focus better on the timbre and spatial positioning of sounds. To reduce test duration and potential listener fatigue, scenes were divided between two test sessions. Each scene was spatialised using 1st, 3rd and 5th-order Ambisonics and each Ambisonic order was assessed separately, therefore Test Sessions 1 and 2 comprised 12 trials each.

6.1.4 Opus Compression and Anchor Creation

The Opus parameters tested were bitrate and channel mapping family. Each Ambisonic reference file was compressed at 16, 32 and 64 kilobits per second per channel (kbps/ch) and for both channel mapping families, 2 and 3. This meant that the

total amount of compressed files, for each scene, at each Ambisonic order, was six. These six different compression conditions composed each trial, along with a hidden reference and mid-range and low anchors. The low anchor was a low-pass filtered version of the reference Ambisonic audio with a cut-off frequency of 3.5 kHz; the mid-range anchor had a cut-off frequency of 7 kHz. Therefore, each trial consisted of nine different conditions.

6.1.5 Participants

Expert listeners were employed as participants. To ensure that the listeners were suitable candidates, people with extensive listening test experience, i.e. professionals and PhD students who work in the field of music and/or audio, were chosen to complete the tests. All participants gave their informed consent to be included in the study. The protocol was approved by the Physical Sciences Ethics Committee of the University of York (approval code: Lee240621). A total of 23 participants took part in this experiment.

6.1.6 Listening Test Environment

The Spatial Audio Listening Test Environment (SALTE) software developed by the author was used for this experiment. It is described in detail in Appendix B. The software consists of a dedicated standalone app containing test control and binaural audio rendering modules. A separate app, SALTE for VR[5], was used to render visual content on a standalone VR headset (Oculus Quest 2). The VR app was developed using Unity. All of the 360° video content had to be uploaded to the Unity project in a Streaming Assets folder so that it would be installed onto the headset and, therefore, directly accessible by the headset when rendering the visuals.

Oculus Quest runs the Android operating system, so the final step was to produce an Android application package (APK) which could be installed on the headset to display 360° videos and connect to the SALTE desktop application wirelessly. Therefore, the VR app needed to send OSC data containing head-tracking information from the headset to the desktop SALTE audio renderer so that the Ambisonic scene could be counter-rotated. Other information, such as the ratings for conditions and the actual test interface display, needed to be passed between the desktop and the headset so that they could move and respond in coordination. The 360° video also needed to be in sync with any audio being output by SALTE, which was achieved by the prior measurement of the time taken for video playback to start and delaying the audio playback by the same amount of time so that both played simultaneously. This was one of the reasons that the selection of stimulus excerpt could not be adjusted by the participant in real-time, as it could have caused video and audio to become out of sync.

[5] `https://github.com/trsonic/SALTE4Quest-XRIT`

6.1.7 Binaural Rendering

The SALTE audio rendering module was configured to process Ambisonic audio and output binaural renderings of each scene while simultaneously processing head-tracking data from the VR headset so that the Ambisonic scene could be rotated in real-time. The rendering chain employed in this experiment follows the generic HRTF one described in Section 4.1.4.

6.1.8 Test Setup

To set up a MUSHRA test in SALTE, a configuration JSON file is used to specify the reference and condition audio files for each trial and to set other parameters of the test. This includes any additional gain applied to the audio files, whether headphones are used instead of loudspeakers, which HRTFs will be used for the subsequent binauralisation and the location where these HRTFs can be found on the disk. The 360° video file name for each trial is also set in the config file. Once the correct config file was selected, the participant had to take note of their randomised subject ID, which was used as an anonymous marker. The test results were exported into a single CSV file.

Once these settings were configured for the SALTE desktop program, the participant had to put their Oculus Quest on, run the APK file and follow the on-screen instructions. The IP address of the participant's Quest is shown, and this is then input into the SALTE desktop program to link the SALTE desktop and SALTE for VR programs together. Once the headset and computer were linked, the participant could begin the test.

Upon clicking 'Begin', the participant was presented with the test rating interface overlaid onto the 360° video for the first trial. The participant was then able to click on each of the conditions to listen to them. Pressing on the corresponding playback buttons would play, pause or stop the audio. The sliders above each of the conditions could be dragged up and down to rate the corresponding condition between 0-100 based on how similar it was to the reference audio. The participant would move between all of the trials using the 'Next' or 'Previous' buttons until they were happy with the ratings they had given to each of the conditions in every trial. A 'Finish' test button would then appear on the final trial, which, upon clicking, ended the test whilst saving the ratings in the CSV file that the participant had created in the test setup. Figure 6.1 shows the SALTE for VR test interface with all of the described features, such as sliders and playback buttons.

The listening tests were originally planned for distribution to participants so that they could be completed in the comfort of their own homes. However, shortly after the initial test distribution, COVID-19 restrictions had relaxed enough for the remainder of the tests to be set up and completed at the University of York's AudioLab. This meant that the remainder of the participants no longer had to download, install and set up the SALTE and SALTE for VR programs on their home devices; this streamlined the test process and allowed for a more efficient collection of a larger quantity of data sets from multiple participants.

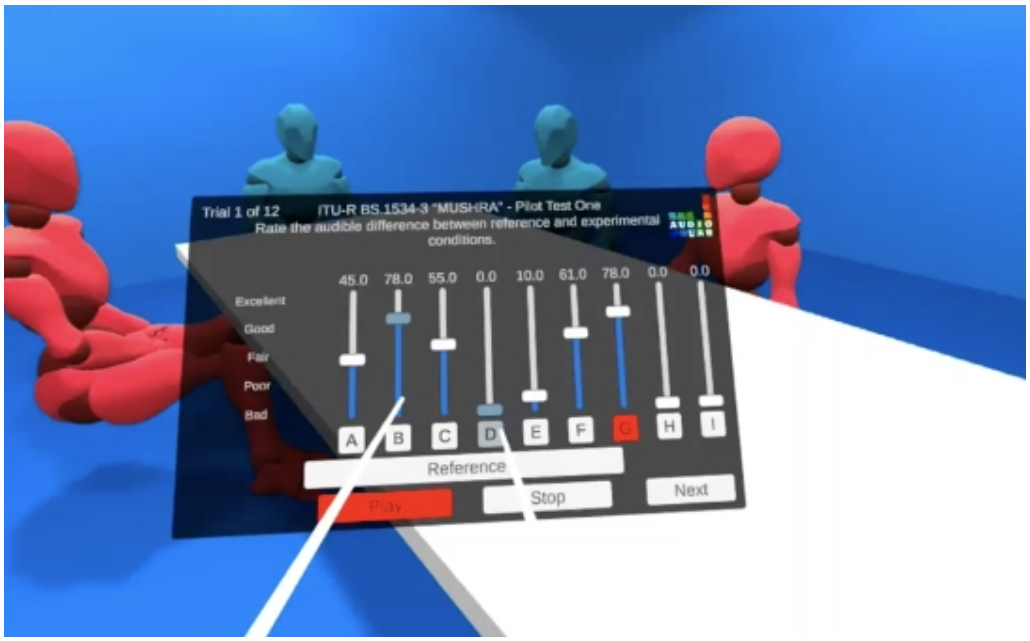


Figure 6.1: A screenshot of the SALTE for VR MUSHRA test interface overlaid on top of the TeleconferenceOne scene.

6.1.9 Data Analysis

To reduce the number of independent variables so as to make the analysis less convoluted, the data collected using stimuli encoded for Ambisonic orders was assessed separately. This study does not address differences in perceived audio quality between different Ambisonic orders.

The first null hypothesis to be investigated is that there is no significant difference between the Basic Audio Quality (BAQ) of compressed and uncompressed stimuli. The second null hypothesis is that, for each bitrate of compression, the channel mapping family has no effect on the BAQ rating. The independent variable in this example is the channel mapping family, within a certain bitrate of compression, and the dependent variable is the BAQ rating. The third null hypothesis is that for each of the codec parameters, the BAQ rating is independent of scene context.

The following criteria were applied in the post-screening of participants: Responses of the assessor collected within a single listening test session were excluded if they rated the hidden reference condition for more than 20% of the test items lower than a score of 90 or if they rated the mid-range anchor for more than 33% of the test items higher than a score of 95. The above criteria are slightly relaxed in comparison to the MUSHRA guidelines, due to the interactive nature of this study. For instance, looking in one direction when listening to a hidden reference and a different direction when listening to the actual reference could affect the rating of this condition by the listener.

The responses collected in both listening test sessions were analysed separately with the above exclusion principles being applied. Therefore some participants might have been excluded, e.g. from the first session, but not the second one. These criteria

resulted in the exclusion of seven assessors due to missed hidden reference and a further four assessors due to mid-range anchor rated too high.

6.1.10 Test for Normality

The Shapiro-Wilk test was used in order to determine whether the data was normally distributed. All data was divided into 24 Order-Scene groups. Each group contained ratings for each of the conditions; as mentioned in Section 6.1.4, there were a total of nine conditions in each trial, meaning there were 216 different distributions tested using the Shapiro-Wilk test. Out of the 216 distributions that were analysed, 150 were not normally distributed. Therefore, it was determined that nonparametric statistical analysis will be conducted using the Kruskal-Wallis test and subsequent comparison of median ranks.

6.2 Results

The conditions in the figure and tables are labelled as shown in Table 6.1. The descriptions of the evaluated conditions can be found in Section 6.1.4.

6.2.1 General Comparison

Figure 6.2 shows median BAQ ratings aggregated over all eight scenes, differentiated by Opus codec parameters and Ambisonic orders. For each test scene, the conditions were evaluated against a reference track rendered at the same Ambisonic order. Therefore, differences in perceived BAQ between different Ambisonic orders are not revealed. The nonparametric 95% confidence intervals have been computed based on the standard formula used to calculate the size of a notch in a boxplot McGill et al. (1978). It can be seen that for every Ambisonic order, the low anchor garnered the lowest BAQ rating of all conditions. The mid-range anchor, 16 kbps/ch Channel Mapping 2 and 16 kbps/ch Channel Mapping 3 conditions all received the next lowest BAQ rating, and there were no significant differences between these conditions at

Table 6.1: Evaluated conditions and their identifiers.

Identifier	Condition
cm2_16kbpspch	16 kbps/ch Channel Mapping Family 2
cm2_32kbpspch	32 kbps/ch Channel Mapping Family 2
cm2_64kbpspch	64 kbps/ch Channel Mapping Family 2
cm3_16kbpspch	16 kbps/ch Channel Mapping Family 3
cm3_32kbpspch	32 kbps/ch Channel Mapping Family 3
cm3_64kbpspch	64 kbps/ch Channel Mapping Family 3
hid_reference	Hidden Reference
low_anchor	Low anchor
mid_anchor	Mid-range anchor

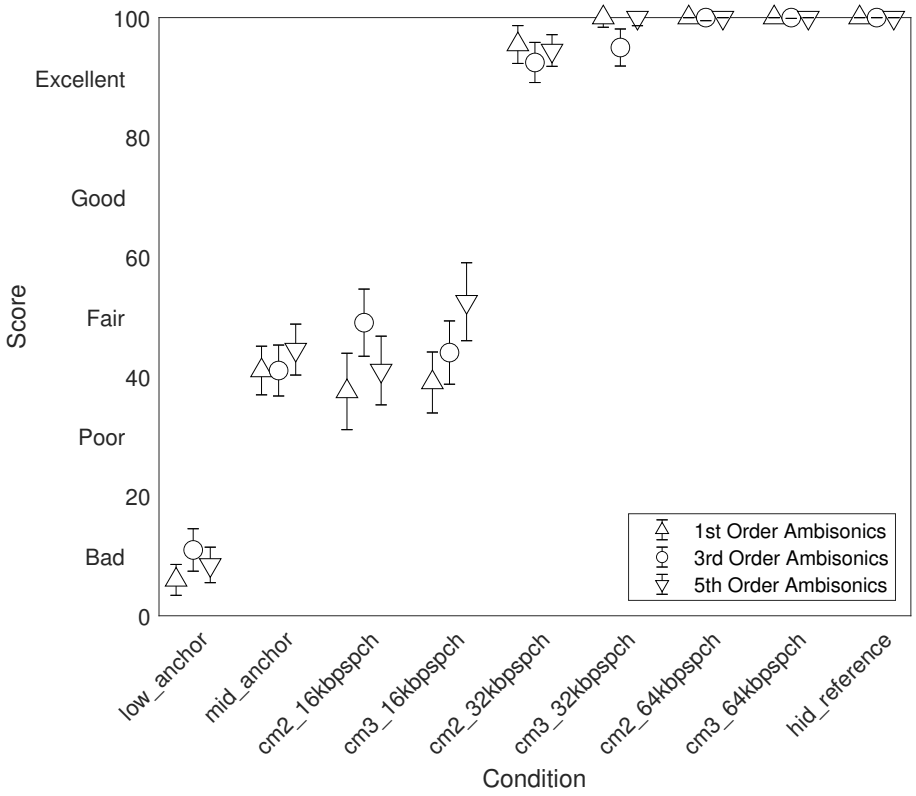


Figure 6.2: Median BAQ ratings for all conditions aggregated over all contexts. Whiskers denote 95% confidence intervals.

respective Ambisonic orders. 32 kbps/ch Channel Mapping 2 conditions received lower median BAQ values for all Ambisonic orders than 32 kbps/ch Channel Mapping 3 conditions. However, the only significant difference between these two conditions can be seen for the 5th-order Ambisonics, as the whiskers do not overlap. This helps to disprove the second null hypothesis, that the use of Channel Mapping Families 2 and 3 results in the same perceived quality of encoded audio. There were no significant differences at any Ambisonic order between 64 kbps/ch Channel Mapping 2, 64 kbps/ch Channel Mapping 3 and the hidden reference conditions, partially supporting the first null hypothesis. In general, Channel Mapping 3 garnered higher BAQ ratings in most instances, except for 3rd-order Ambisonic scenes compressed at 16 kbps/ch.

6.2.2 Perceived Audio Quality Impairment

This section addresses the problem of finding codec parameters which do not cause perceived degradation of BAQ through a comparison of all experimental condition scores against uncompressed stimuli scores. Table 6.2 shows p-values obtained from pairwise comparisons between hidden reference and each of the remaining conditions using the Wilcoxon rank sum test. Each of the bitrates is analysed to test the null hypotheses separately.

Table 6.2: p-values obtained from pairwise comparisons between uncompressed audio (hidden reference) and each of the remaining conditions using Wilcoxon rank sum test. Assuming a p-value threshold of 0.05 the table cell colours denote the following findings: none - rating score distributions are significantly different; dark grey - is significantly the same; light grey - neither.

	low_ anchor	mid_ anchor	cm2_ 16kbpspch	cm3_ 16kbpspch	cm2_ 32kbpspch	cm3_ 32kbpspch	cm2_ 64kbpspch	cm3_ 64kbpspch
GameCar_1OA	0.000053	0.980273	0.079739	0.00023	0.718707	0.663576	0.754917	0.791644
GameCar_3OA	0.000036	1	0.000227	0.000031	0.399109	0.862624	0.798117	0.791674
GameCar_5OA	0.000878	0.138096	0.001282	0.000002	0.260397	0.017076	0.455617	0.695352
GameFPS_1OA	0	0	0.000008	0.000007	0.11014	0.769304	0.777101	0.961729
GameFPS_3OA	0	0	0.000004	0.000001	0.001556	0.076657	0.279069	0.536784
GameFPS_5OA	0	0	0.000001	0	0.007141	0.042497	0.346427	0.097076
MusicBlues_1OA	0.000001	0.000001	0.000028	0.000006	0.014397	0.043275	0.916158	0.632725
MusicBlues_3OA	0.000001	0.000001	0.000028	0.000028	0.008258	0.00152	0.018066	0.079725
MusicBlues_5OA	0.000001	0.000001	0.000021	0.000004	0.014247	0.000259	0.325821	0.179751
MusicMallets_1OA	0	0	0	0	0.025387	0.003929	0.422191	0.910876
MusicMallets_3OA	0	0	0	0	0.000275	0.076628	0.572773	1
MusicMallets_5OA	0	0	0	0.000001	0.003531	0.257501	0.412137	0.45305
SoundscapeFarm_1OA	0.000002	0.000017	0.000096	0.000019	0.017676	0.663629	0.837836	0.978746
SoundscapeFarm_3OA	0.000001	0.000068	0.000007	0.000129	0.000876	0.092739	0.151459	0.48852
SoundscapeFarm_5OA	0.000002	0.000215	0.000004	0.000018	0.048918	0.215485	0.942572	0.823946
SoundscapeOasis_1OA	0	0	0.000001	0	0.000063	0.295024	0.860367	1
SoundscapeOasis_3OA	0	0	0.000003	0	0.000435	0.023186	0.740057	0.334596
SoundscapeOasis_5OA	0	0	0	0	0.000005	0.000001	0.178139	0.916615
TeleconferenceOne_1OA	0.000001	0.000001	0.000001	0.000001	0.019058	0.006702	0.361567	0.361415
TeleconferenceOne_3OA	0.000002	0.000002	0.000002	0.000011	0.034945	0.10601	0.910914	0.918453
TeleconferenceOne_5OA	0.000001	0.000001	0.000001	0.000006	0.003675	0.013402	0.389423	0.609312
TeleconferenceTwo_1OA	0	0	0	0	0.019373	0.083438	0.166298	0.153641
TeleconferenceTwo_3OA	0	0	0	0.000001	0.046467	0.402262	0.211677	0.322942
TeleconferenceTwo_5OA	0	0	0	0	0.414746	0.405687	0.701748	0.789881

64 kbps/ch Bitrate

It can be seen from Table 6.2 that there are no significant differences present between the hidden reference and any of the stimuli compressed at 64 kbps/ch using Channel Mapping Family 3. For some of the stimuli, the ratings are significantly the same as for the hidden reference, suggesting the perceptual transparency of Opus compression at these particular settings. This finding, therefore, supports the first null hypothesis that there is no significant difference between the median BAQ rating of the uncompressed reference and Opus compressed stimuli at the 64 kbps/ch bitrate using Channel Mapping Family 3.

The ratings of stimuli encoded at 64 kbps/ch using Channel Mapping Family 2 follow the same trend in many cases, as there is only one trial of the 24 where the median BAQ rating was significantly different from the hidden reference rating. This finding, therefore, also supports the first null hypothesis in all trials but the Music Blues scene encoded using 3rd-order Ambisonics.

32 kbps/ch Bitrate

32 kbps/ch compressed material for both channel mapping families had instances where it was significantly different to the hidden reference, disproving the first null hypothesis. This bitrate showed greater variation between channel mapping families than the other bitrates. To elaborate, Channel Mapping Family 2 was significantly different from the hidden reference more times than its counterpart, Channel Mapping Family 3. At this bitrate, Channel Mapping Family 2 was significantly different from the hidden reference in 19 of the 24 trials, whereas Channel Mapping Family 3 was significantly different from the hidden reference in 10 of the 24 trials. This finding suggests that there is a difference between channel mapping families, thereby rejecting the second null hypothesis.

16 kbps/ch Bitrate

16 kbps/ch compressed material was significantly different to the hidden reference in all cases but one, where Channel Mapping Family 2 showed neither significance for the Game Car 1OA scene. The first null hypothesis is therefore disproven at this bitrate in most cases, as there was a significant difference between the compressed stimuli and the hidden reference. These results also support the second null hypothesis at this bitrate as both channel mapping families were rated significantly different from the hidden reference in all trials but one, which was the only difference between the channel mapping families at this bitrate.

Other Findings

An interesting finding that Table 6.2 presents clearly is that the mid-range anchor was never significantly different and even rated significantly the same as the hidden reference twice across the trials utilising the Game Car scene. This makes sense upon reflection as the uncompressed reference audio consisted of low-pass filtered audio for engine and road sounds typically found in first-person car games. The mid-range

Table 6.3: p-values obtained from pairwise comparisons between Channel Mapping Families 2 and 3 using Wilcoxon rank sum test. Assuming a p-value threshold of 0.05, the table cell colours denote the following findings: none - rating score distributions are significantly different; dark grey - is significantly the same; light grey - neither.

	16kbpspch	32kbpspch	64kbpspch
1OA	0.474799	0.007192	0.951408
3OA	0.034928	0.001441	0.112376
5OA	0.373840	0.555900	0.516494

anchor applies a 7.5kHz low pass filter to the uncompressed audio, which would have had little to no effect if the audio was already low pass filtered below this frequency during the stimuli production phase.

At first glance, it is difficult to tell whether specific contexts were affected differently by Opus compression, but it can be seen from Table 6.2 that certain contexts had more conditions with median BAQ ratings significantly different from the hidden reference than other contexts. This contradicts the third null hypothesis that for each of the codec parameters, the BAQ rating is independent of the scene context. The context where most conditions were rated differently from the hidden reference was music, which could suggest that this content is more difficult to encode than others. Further analysis is required to fully determine whether or not the third null hypothesis has been proven or disproven by these results.

6.2.3 Channel Mapping Family Effect on BAQ

This section addresses whether there are perceived differences in BAQ ratings between Channel Mapping Families 2 and 3 at each bitrate. Table 6.3 shows p-values obtained from pairwise comparisons between the two channel mapping families for all contexts combined at separate bitrates and Ambisonic orders using the Wilcoxon rank sum test.

For 64 kbps/ch, the second null hypothesis holds true as no significant difference could be found between the two channel mapping families at this bitrate at any Ambisonic order - 1st-order Ambisonic content at this bitrate even showed that the channel mapping families were significantly the same. For 32 kbps/ch, the second null hypothesis is rejected for 1st and 3rd-order Ambisonics because there was a significant difference between the channel mapping families - no significance, either way, was determined at this bitrate for 5th-order Ambisonics. Finally, for 16 kbps/ch, the second null hypothesis is rejected for 3rd-order Ambisonics because there was a significant difference between the channel mapping families - no significance, either way, was determined at this bitrate for 1st and 5th-order Ambisonics.

6.2.4 Stimulus Context Effect on BAQ

This section addresses the problem of determining whether there is any difference in BAQ ratings between contexts at each bitrate. Each condition from each of the trials was combined except for the separate Ambisonic orders. Table 6.4 shows p-values

Table 6.4: p-values obtained using Kruskal-Wallis test on data grouped by different contexts at separate conditions and Ambisonic orders. Assuming a p-value threshold of 0.05, the table cell colours denote the following findings: none - rating score distributions are significantly different between different contexts; light grey - no significant difference between the contexts.

	low_ anchor	mid_ anchor	cm2_ 16kbpspch	cm3_ 16kbpspch	cm2_ 32kbpspch	cm3_ 32kbpspch	cm2_ 64kbpspch	cm3_ 64kbpspch	hid_ reference
1OA	0.883807	0.000002	0.000253	0.450940	0.007015	0.081333	0.413270	0.503503	0.244155
3OA	0.217983	0.000001	0.000704	0.203719	0.018235	0.085015	0.222632	0.008767	0.066835
5OA	0.001135	0.000002	0.000840	0.008950	0.002021	0.000002	0.937564	0.663838	0.532762

obtained using the Kruskal-Wallis test for all contexts at separate conditions and Ambisonic orders. Each of the bitrates was analysed and used to test the third null hypothesis separately.

For 64 kbps/ch Channel Mapping Family 2, the third null hypothesis holds true as no significant difference could be found between the contexts for this condition as at any Ambisonic order; 64 kbps/ch Channel Mapping Family 3 showed a significant difference between contexts for 3rd-order Ambisonic content only. For 32 kbps/ch Channel Mapping Family 2, the third null hypothesis is rejected as there were significant differences found between the contexts for this condition at all Ambisonic orders; 32 kbps/ch Channel Mapping Family 3 showed a significant difference between contexts for 5th-order Ambisonic content only. Finally, for 16 kbps/ch Channel Mapping Family 2 the third null hypothesis is rejected as there were significant differences found between the contexts for this condition at all Ambisonic orders; 16 kbps/ch Channel Mapping Family 3 showed a significant difference between contexts for 5th-order Ambisonic content only. In general, Channel Mapping Family 3 showed the least variation in median BAQ between contexts at most bitrates and Ambisonic orders.

6.3 Discussion

A similar study conducted by Fela et al. (2022), involved the evaluation of 360° video with 4th-order Ambisonic audio reproduced over a loudspeaker array. The study also featured varied content with different contexts. However, all of the scenes in that study were actual recordings, not virtually produced, and different orders of Ambisonics were not tested. Their study assessed video and audio separately, then assessed video and audio combined, whereas this study only focused on audio impairments; video remained constant in a given trial and was not assessed independently.

The Ambisonic audio in Fela et al.'s study was compressed at the same compression rates of 16, 32 and 64 kbps/ch using FFmpeg with AAC-LC encoder, whereas this study used Opus audio codec with its channel mapping 2 and 3 families. One finding of the current study was that audio compressed at 64 kbps/ch was not significantly different to uncompressed audio in most cases. This could suggest that Opus preserves audio quality at this bitrate better than the codecs used in Fela et al.'s study, where there was a more obvious reduction in quality when audio was

compressed at 64 kbps/ch from the original PCM signal. This finding aligns well with the results of the study described in Chapter 4, which suggested that the codec is perceptually transparent at 64 kbps/ch bitrate for 5th-order complex scenes. Finally, Fela et al. used a different metric to assess stimuli in their tests, Mean Opinion Score (MOS), whereas this study used Basic Audio Quality (BAQ) as the metric to assess the stimuli.

Due to the large number of variables in this study, data analysis was quite challenging, and many different approaches could have been taken. In further work, it would be useful to focus on a specific context at a particular Ambisonic order and give listeners fewer stimuli to compare at once. For example, cutting down from the current nine stimuli in each trial and instead having just five stimuli: one hidden reference, both anchors and Channel Mapping Families 2 and 3 for just one bitrate of compression. This may help to reduce the spread exhibited in the results gathered in this study which could give rise to a significant difference between the two different channel mapping families.

The trials presented could also be tested with no visuals. If there is a significant difference, such as a higher rating/less difference between compressed audio and the hidden reference with visuals enabled, this could suggest that lower bitrates can be used when the VR media contains visuals. Conversely, the perceived effects due to the increased complexity of the visuals could also be investigated, such as complex animations and more realistic assets.

Further exploration could also involve changing the content; for example, using real-life 360° videos and simultaneously recorded Ambisonic audio. A comparison could then be made on the effects of Opus compression on real or virtual scenes and whether one favours a certain channel mapping family over another; the only context where this would be challenging is in "Game" scenes, as most VR games are created with virtual content. Using a microphone array capable of capturing 5th-order (or higher) Ambisonics would be the preferred recording option over up-mixing lower-order Ambisonic content.

Further work could look to improve the test procedure. For example, conducting a pilot test for the pre-screening of each participant as described in ITU-R (2015b) could reduce the number of participants that had to be excluded from the final results. Also, a more sophisticated way to keep the visuals and audio in sync is desirable, in order for participants to be able to play certain parts of the scene and set a playback loop. In this study, the participants had to listen to the stimulus from the beginning when they switched conditions in order to keep audio and video synchronised. This study could also be extended to investigate other binaural rendering chains, for example, different methods for rendering Ambisonics binaurally and different HRTFs.

6.4 Conclusion

The study presented in this chapter contributes new information about which bitrates of Opus compression and channel mapping families provide perceptual transparency when dealing with Ambisonic audio for different VR contexts and Ambisonic orders. The four contexts presented were Gaming, Music, Soundscapes and Teleconferencing.

The Opus parameters investigated were the channel mapping family and the bitrate of compression.

The first null hypothesis investigated was: *There is no significant difference between the Basic Audio Quality (BAQ) rating of compressed and uncompressed stimuli.* This null hypothesis holds true for 64 kbps/ch as no significant difference was found between this bitrate and the uncompressed hidden reference. For 32 kbps/ch there were trials where the compressed stimuli garnered a median BAQ significantly different to the uncompressed audio and therefore disproves this null hypothesis in most of the trials at this bitrate. For 16 kbps/ch, in most trials, the compressed stimuli garnered a median BAQ significantly different to the hidden reference therefore also disproving this null hypothesis at this bitrate.

The second null hypothesis was: *the channel mapping family, within a certain bitrate of compression, has no effect on the BAQ rating that a stimulus is awarded.* This null hypothesis holds true for 64 kbps/ch as there was no significant difference between the channel mapping families at any Ambisonic order. For 32 kbps/ch there were significant differences between the channel mapping families in 1st and 3rd-order Ambisonic content which disproves this null hypothesis for this bitrate at these Ambisonic orders. For 16 kbps/ch there were significant differences between the channel mapping families in 3rd-order Ambisonic content which disproves this null hypothesis for this bitrate at this Ambisonic order.

The third null hypothesis investigated was: *for each of the codec parameters, the BAQ rating is independent of the scene context.* This null hypothesis holds true for some bitrates and some Ambisonic orders but is disproven by others; in general, 64 kbps/ch Channel Mapping Family 2 garnered the lowest variation in median BAQ between different contexts, as there was no significant difference between context at any Ambisonic order - this does not necessarily mean that this condition produced the highest BAQ for each, just that it gained the most consistent BAQ rating across all contexts.

The key result is that across all trials, there was no significant difference between stimuli compressed at 64 kbps/ch, using Channel Mapping Family 3, and the hidden reference, making these settings optimal to use if the Basic Audio Quality of the original Ambisonic audio is to be preserved. Channel Mapping Family 2 at this bitrate performed just slightly worse, garnering a BAQ rating significantly different than the uncompressed audio for only one of the scenes. Furthermore, Channel Mapping Family 3 showed no significant difference in median BAQ ratings across evaluated contexts at a higher number of bitrate and Ambisonic order conditions than Channel Mapping Family 2, which suggests it is a more robust compression scheme.

Chapter 7

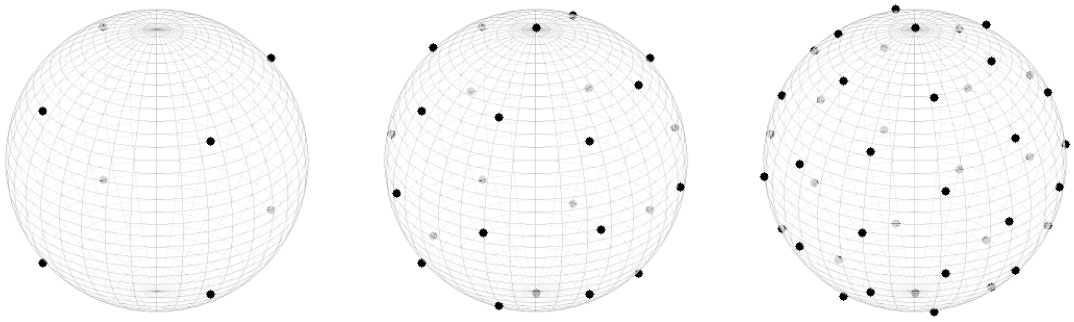
Evaluation of Binaural Ambisonic Rendering Methods

The previous chapters discussed experiments aimed at the evaluation of Opus codec under different conditions. However, another critical factor to consider when discussing the quality of the Ambisonic delivery chain is the rendering method used. Interactive binaural audio rendering has been widely adopted to deliver immersive audio over headphones, for example, in VR games and applications, 360° video streaming and for playback of music produced using immersive audio formats. Some binaural renderers use Ambisonics as their input format or an intermediate sound bed, which is then decoded to the left and right ear headphone signals through respective filters. An excellent example of such a renderer is Google Resonance Audio (Gorzel et al., 2019) which has been used extensively in the VR industry and has set a standard for spatial audio quality in VR. This renderer has also been used as part of the YouTube 360 streaming service to provide a head-tracked rendering of binaural audio. However, since the release of the renderer, several methods have been proposed to improve the quality of the binaural rendering of Ambisonics. Therefore, it is pertinent to investigate these methods, evaluate existing implementations, and look into possible improvements to the state-of-the-art methods.

This chapter focuses on the implementation of the established and alternative methods for designing Ambisonic-to-binaural filters within a single framework and subsequent objective and subjective evaluations of these. The conducted experiments are aimed at providing a basis for suggesting a filter design procedure perceptually superior to the one used for the calculation of the current Resonance Audio binaural filters.

7.1 Evaluated Methods

This section describes the principles and implementation details of the methods evaluated in this chapter. The Google Resonance Audio method is used as a baseline in this comparison. It uses the virtual loudspeaker approach, as defined below in Section 7.1.1. This study extends this method by applying a direction-independent equalisation. The leading contender for the proposed improvement is



(a) Cube. (b) 26-pt Lebedev grid. (c) Pentakis icosidodecahedron.

Figure 7.1: Virtual loudspeaker layouts.

MagLS (Schörkhuber et al., 2018) method, described in Section 7.1.2. The MagLS method is extended in this study by applying the diffuseness constraint, as proposed by Zaunschirm et al. (2018); Zotter and Frank (2019). The objective evaluation also included an alternative approach to MagLS, the AkLS method described below in Section 7.1.3. Table 7.1 lists all evaluated methods, followed by the implementation description of each method.

Table 7.1: Evaluated binaural Ambisonic filter design methods.

Identifier	Method
VL	Virtual Loudspeakers
VL-EQ	Virtual Loudspeakers with Equalisation
MagLS	Magnitude Least Squares
MagLS-diffc	Magnitude Least Squares with Diffuseness Covariance Constraint
AkLS	Linear Phase Approach

7.1.1 Virtual Loudspeakers

The conventional approach to the binaural rendering of Ambisonic scenes is to use a grid of virtual loudspeakers represented by HRIRs. Therefore, it is possible to use mode-matching decoders suitable for decoding Ambisonic signals to regular loudspeaker layouts. For example, the Google Resonance Audio decoder uses directions corresponding to the cube, 26-pt Lebedev grid and pentakis icosidodecahedron vertices for 1st, 3rd and 5th-order Ambisonic rendering respectively (see Figure 7.1).

In such a case, the decoding matrix is obtained by calculating the pseudoinverse of the Ambisonic encoding matrix derived by evaluating the spherical harmonics at the loudspeaker directions (see Section 3.4.2). Firstly, the coefficients are stored in

an encoding matrix

$$Y = \begin{bmatrix} \mathbf{Y}_0^i(\theta_1, \phi_1) & Y_1^i(\theta_1, \phi_1) & \dots & Y_M^i(\theta_1, \phi_1) \\ \mathbf{Y}_0^i(\theta_2, \phi_2) & Y_1^i(\theta_2, \phi_2) & \dots & Y_M^i(\theta_2, \phi_2) \\ \vdots & \vdots & & \vdots \\ \mathbf{Y}_0^i(\theta_Q, \phi_Q) & Y_1^i(\theta_Q, \phi_Q) & \dots & Y_M^i(\theta_Q, \phi_Q) \end{bmatrix}. \tag{7.1}$$

Then a Basic (mode-matching) Ambisonic decoding matrix D is obtained as the pseudoinverse of Y:

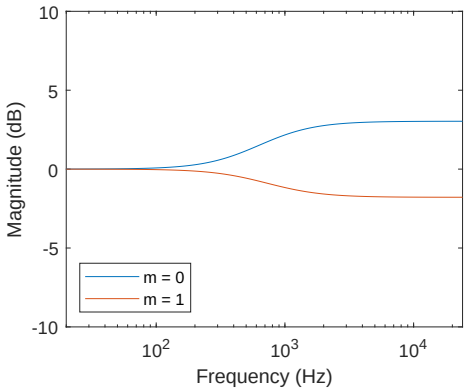
$$D = \text{pinv}(Y) = Y^T(YY^T)^{-1}. \tag{7.2}$$

However, the Basic decoder can reproduce the ITD and spectral cues of the original HRTFs only up to f_{alias}. A certain way to mitigate the reconstruction errors of ear signals at high frequencies is to use a dual-band approach, where frequencies below f_{alias} are decoded using the Basic decoder while frequencies above f_{alias} are decoded using max-rE-weighted decoder, which aims to maximise the energy vector r_E for all directions (Daniel et al., 1998). Dual-band decoding may be implemented by pre-filtering the Ambisonic input with a set of shelf-filters and applying max-rE correction weightings to the high-passed signals before feeding the decoder. Such a strategy is used to calculate the Google Resonance SH-HRIRs. Figure 7.2 shows the magnitude responses of these filters at different Ambisonic orders.

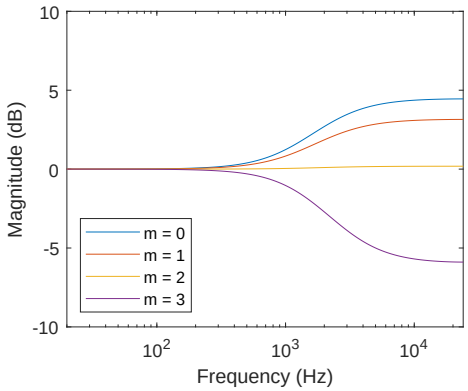
Another degree of improvement can be achieved by applying a global equalisation filter to the binaural signal. Ben-Hur et al. (2017) proposed the use of a Spherical Head Filter based on the spherical head model to compensate for the loss of high-frequency energy at higher Ambisonic orders, whereas McKenzie et al. (2018) proposed a more general approach utilising an inverse filter derived from the diffuse-field response of the calculated SH-HRIRs. In this work, we use a modified Diffuse Field Equalisation approach in which the linear-phase EQ filter is calculated as an inverse of the ratio between reconstructed and original HRTF set diffuse field responses. This allows for preserving the Common Transfer Function of the original HRTF set, which typically is not flat, e.g. for human HRTFs or Bernschütz (2013) KU100 HRTFs, which are diffuse field pre-equalised using the analogue circuitry built-in in the artificial head. Figure 7.3 shows diffuse field magnitude responses of the original and reconstructed HRTFs as well as global equalisation filters for the SH-HRIRs calculated using the virtual loudspeaker method at different Ambisonic orders.

7.1.2 Magnitude Least Squares

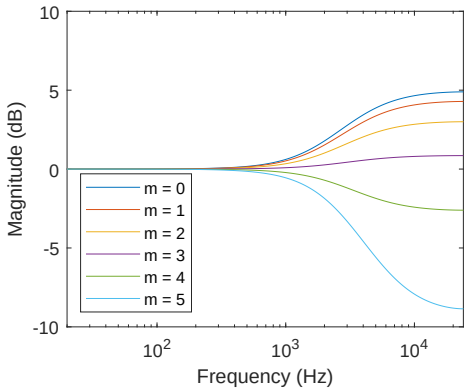
The inaccuracy of high-frequency rendering of Ambisonic signals is inherent to human ears being spaced apart. However, the binaural rendering of Ambisonics differs from using a real loudspeaker array in the sense that the used HRTF set can be freely manipulated prior to using it. This allows one to remove interaural delays from HRTFs at high frequencies while maintaining original ITD cues at low frequencies, where they are the most perceptually significant. Based on this assumption, Zaunschirm et al. (2018) proposed high-frequency time alignment and Schörkhuber et al. (2018) proposed Magnitude Least Squares (MagLS). The latter



(a) 1st-order Ambisonics shelving filters.



(b) 3rd-order Ambisonics shelving filters.



(c) 5th-order Ambisonics shelving filters.

Figure 7.2: Google Resonance shelving filters for applying max-rE weights at high frequencies.

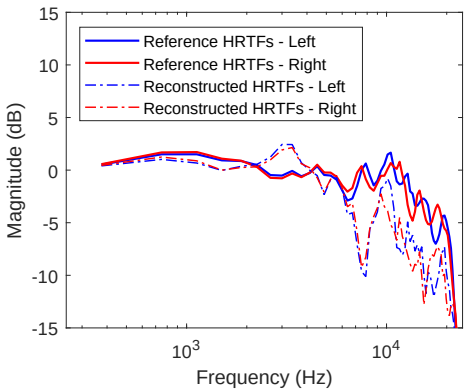
method performed better in the (Schörkhuber et al., 2018) study, which included an objective comparison of both methods. Therefore, the MagLS method remains the focus of this work.

The MagLS method was implemented in this study based on the code attached to the Deppisch et al. (2021) paper. The code available at the time of preparation of this study[1] was modified to correct the issue of ITD inaccuracy in reconstructed HRIRs. The MagLS method was therefore implemented in the following manner. All 2702 HRIR pairs from the KU100 set (Bernschütz, 2013) were used. The original HRIRs of 128 sample lengths were zero-padded to a length of 2048 samples. Then a median group delay was calculated based on the 0th-order component of the Basic decoding matrix. All HRIRs were shifted backwards by the fixed median group delay value, eliminating the global delay. The Least Squares solution was calculated as

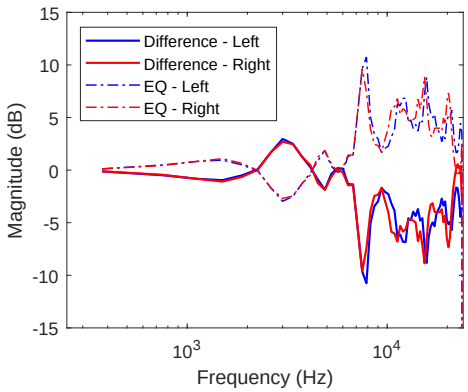
$$w_{LS}(\epsilon) = h(\epsilon) \cdot \text{pinv}(Y), \tag{7.3}$$

$$\epsilon \subset (left, right), \tag{7.4}$$

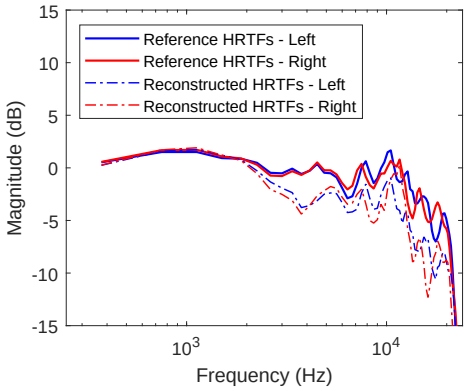
[1] `https://github.com/thomasdeppisch/eMagLS/blob/dbcbdf74d5fb8961b17616c00bfc1f34fc039077/getMagLsFilters.m`



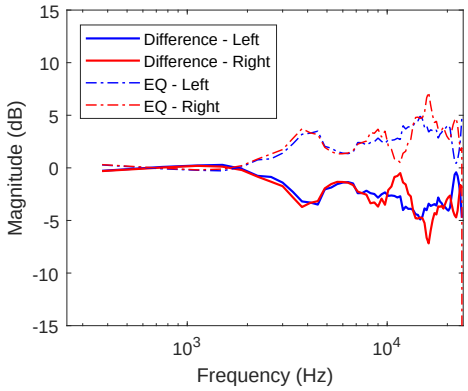
(a) Diffuse field responses of the original HRTF set and reconstructed one using 1OA SH-HRIRs.



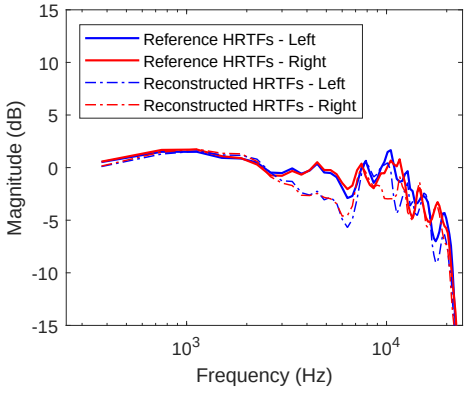
(b) Differences between diffuse field responses of original and reconstructed HRTFs and respective equalisation filters at 1OA.



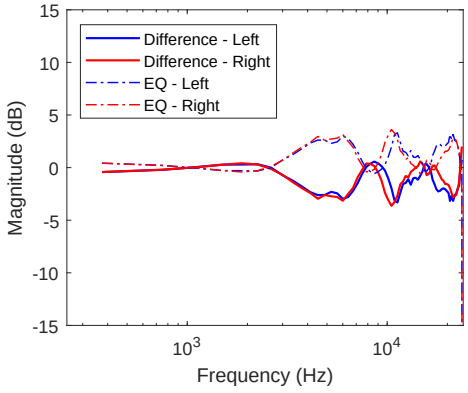
(c) Diffuse field responses of the original HRTF set and reconstructed one using 3OA SH-HRIRs.



(d) Differences between diffuse field responses of original and reconstructed HRTFs and respective equalisation filters at 3OA.



(e) Diffuse field responses of the original HRTF set and reconstructed one using 5OA SH-HRIRs.



(f) Differences between diffuse field responses of original and reconstructed HRTFs and respective equalisation filters at 5OA.

Figure 7.3: Diffuse field responses of original and reconstructed HRTF sets. Differences between the responses and respective equalisation filters for left and right ear signals at different Ambisonic orders.

where h is the HRIR matrix, ϵ denotes the processed ear and Y is the encoding matrix. Then the decoding matrices were transformed into the frequency domain using FFT:

$$W_{LS}(\epsilon) = \mathrm{FFT}(w_{LS}(\epsilon)). \tag{7.5}$$

Based on the revised experiments on sound localisation and ITD sensitivity, the phase difference between pure tones at both ears is noticeable exclusively at frequencies below ca. 1400 Hz (Mills, 1958; Klug and Dietz, 2022). The cut-off frequency f_x was set to 1500 Hz, 1927 Hz, and 3211 Hz for 1OA, 3OA, and 5OA filters, respectively. The 1500 Hz lower limit was imposed to preserve ITD cues at the relevant frequencies.

$$f_x = \max(f_{alias}(M), 1500) \tag{7.6}$$

Below the cut-off frequency, the solution remained unchanged. Above the cut-off frequency, a Magnitude Least Squares solution was calculated as follows:

$$\Phi = \angle(W_{\mathrm{MLS}}(k-1, \epsilon) \cdot Y), \tag{7.7}$$

$$W_{\mathrm{MLS}}(k, \epsilon) = |H(k, \epsilon)|e^{i\Phi} \cdot \mathrm{pinv}(Y), \tag{7.8}$$

where H is the HRTF matrix and k denotes the consecutive frequency bins. The resulting decoder filters are converted to the time domain using inverse-FFT, resulting in h_{LR}^{SH} matrix of the following dimensions: N samples x SH coefficients number x 2 (left and right ear).

Along with the time alignment method Zaunschirm et al. (2018) introduced the use of a covariance constraint (Vilkamo et al., 2013) in order to enhance interaural decorrelation for the rendering of diffuse sound fields at low Ambisonic orders. For a full explanation of the method, the reader is referred to (Zotter and Frank, 2019). The constraint was calculated according to the Deppisch et al. (2021) paper code repository.

7.1.3 Linear Phase Approach

The linear phase approach assumes that the binaural filters can be synthesised based solely on the magnitudes of the original HRTFs and their estimated ITDs. In this work, the method is referred to as AkLS. Similarly to the MagLS method, ITDs are disregarded at frequencies above the cut-off frequency. If the evaluated frequency is below f_x, the linear phase delay is calculated based on the ITD value (see Equation 7.9) and is equal to zero for the remaining frequencies. The cut-off frequency f_x was the same as for the MagLS decoders at each respective order.

$$\Phi = \begin{cases} \pm 2\pi\omega\frac{ITD(\theta,\phi)}{2} & \mathrm{if} \omega < f_x \\ 0 & \mathrm{if} \omega \geq f_x, \end{cases} \tag{7.9}$$

where the sign of Φ depends on the processed ear. The decoder is calculated using Equation 7.10.

$$H(\omega, q, \epsilon) = |H_{orig}(\omega, q, \epsilon)|e^{i\Phi} \cdot \mathrm{pinv}(Y), \tag{7.10}$$

7.2 Objective Evaluation

This section describes the objective evaluation of binaural Ambisonic decoders obtained using the methods specified in Section 7.1. This analysis aimed to identify the methods that perform best objectively to decrease the number of conditions evaluated in further subjective tests.

The numerical evaluation of binaural rendering filters was based on a comparison of the HRIR set reconstructed from SH-HRIRs with the original HRIR set. This approach has been previously used by Wiggins et al. (2001); Engel et al. (2022) and others. This work was carried out using the KU100 HRTF set (Bernschütz, 2013), which consists of HRIR pairs measured using a 2702-point Lebedev grid. The evaluation was carried out for 1st, 3rd and 5th-order filters.

7.2.1 Methods

The following objective metrics have been used to evaluate the proposed binaural Ambisonic decoder design methods.

Diffuse-Field Response

The diffuse-field response (DFR) was calculated by averaging left- and right-ear HRTF magnitudes over the entire HRTF set. Each direction was weighted by its corresponding solid angle value, as in Equation 7.11, where K represents the total number of HRTFs and $\Omega(k)$ is the solid angle vector expressed in sr.

$$H_{DFR} = \sqrt{\frac{1}{4\pi} \sum_{k=1}^{K} |H(k)|^2 \, \Omega(k)} \tag{7.11}$$

Binaural Cues

The ITD for each HRIR pair was calculated as follows. First, each HRIR was filtered with a linear-phase low-pass filter with a cut-off at 1.2 kHz and 128-tap length, as shown in Figure 7.4, as the ITD cue is perceptually relevant mostly at low frequencies. The index of the sample with a maximum absolute value in both IRs was found. Both IRs were truncated 0.5 ms ahead and 5 ms behind the peak sample. This was done to remove any possible room reflections and also to shorten the IRs to speed up the ITD estimation process. Next, the IRs were up-sampled, increasing the time scale resolution by $r = 8$. The lag τ for each IR was estimated by finding the maximum of the cross-correlation function between the IR and its minimum-phase version (Nam et al., 2008).

$$R_{xy}(n) = \sum_{m=-\infty}^{\infty} h(m) \cdot h_{\min}(m - n) \tag{7.12}$$

$$\tau = \arg\max |R_{xy}(n)| \tag{7.13}$$

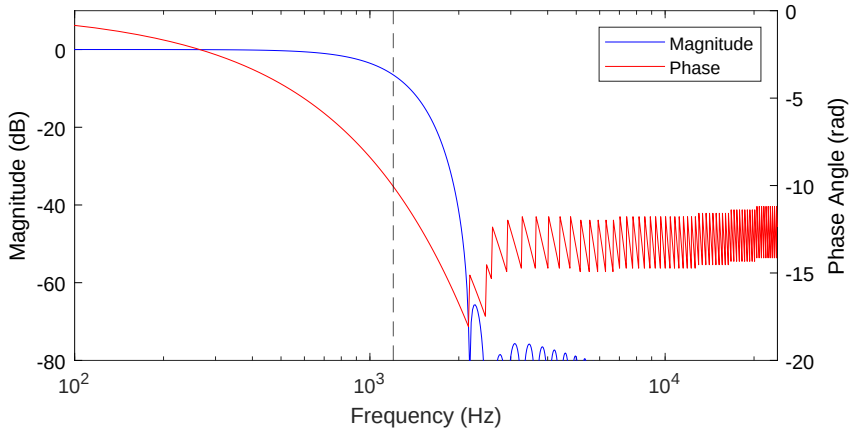


Figure 7.4: ITD estimation algorithm low-pass filter.

Then ITD was estimated as a difference in TOA between left- and right-ear signals taking into account the up-sampling factor r.

ILD was estimated in the frequency domain based on the power spectra ratio between the left and right ear signals. First, the relative weights of high-frequency FFT bins were reduced by calculating the RMS of the spectrum over 40 equivalent rectangular bandwidth (ERB) bands. Following Glasberg and Moore (1990), the ERB bands were calculated as follows:

$$\begin{aligned} Q &= 9.265 \\ L &= 24.7 \\ k &= 1, 2, \ldots, 40 \\ f_{bw}(k) &= L \exp(\frac{k}{Q}) \\ f_c(k) &= Q(f_{bw}(k) - L) \end{aligned}$$

For further ILD calculation, only ERB bands with $f_c > 1500$ Hz were considered.

$$\text{ILD} = 20 \log_{10} \left(\frac{\sqrt{\frac{1}{21} \sum_{k=19}^{40} |H_L(k)|^2}}{\sqrt{\frac{1}{21} \sum_{k=19}^{40} |H_R(k)|^2}} \right) \tag{7.14}$$

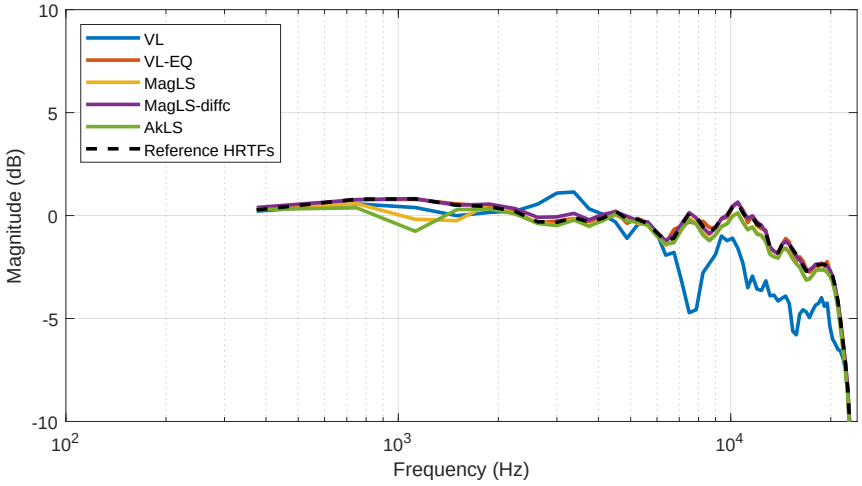
Spectral Difference

Due to Ambisonic order truncation, HRTF spectral cues become distorted. This study calculated the average spectral difference based on the RMS average of the spherical-angle-weighted difference between the reference and reconstructed HRTF magnitude. Such a metric is valuable in diagnosing specific problems with binaural filters, e.g. spatial aliasing affecting certain frequency ranges. However, measuring the perceived spectral colouration of signals using single-value metrics is more complicated. Several

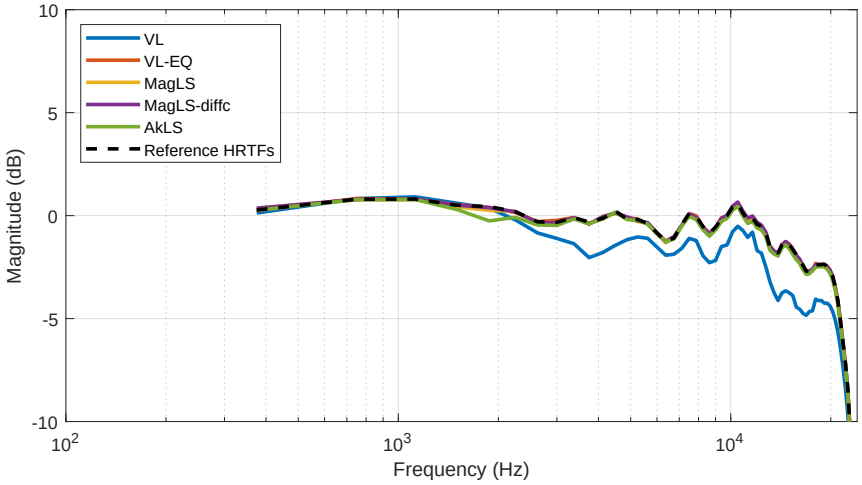
methods have been proposed to tackle this problem, such as BSD, PEAQ and CLL. More recently, McKenzie et al. (2022) proposed an objective signal colouration metric designed specifically for binaural audio, the predicted binaural colouration (PBC) method. The implementation of the method is available as part of the Auditory Modeling Toolbox (Majdak et al., 2022). This method has been used to assess the perceptually relevant differences introduced by different binaural Ambisonic schemes evaluated in this chapter.

7.2.2 Results

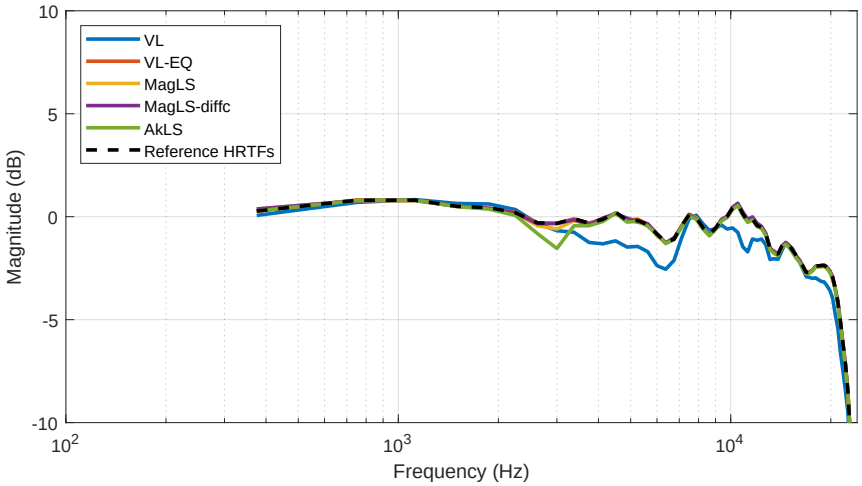
As shown in Figure 7.5, the Virtual Loudspeaker method without EQ (VL) results in diffuse-field responses much different from the original HRTF set one. The other methods result in responses closely following the reference. Figure 7.6 shows estimated ITDs in the horizontal plane. The characteristics look similar across all methods and match the ITD curve of the reference HRTF set. However, there are differences in the distribution of ITD error across all 2702 evaluated points, as shown in Figure 7.7. At 1st-order, the AkLS method performs best, while at higher orders, MagLS methods result in smaller estimated ITD errors. Generally, MagLS is the closest to the reference ILD, as shown in Figure 7.8. It is interesting to observe that MagLS-diffc results in asymmetrical ILD, especially at 1st order. The MagLS performs best in terms of overall ILD error, as shown in Figure 7.9. The average spectral difference is the lowest for AkLS and MagLS at 1st order, as shown in Figure 7.10. At 3rd and 5th-order, MagLS performs best. Figure 7.11 shows the predicted binaural colouration distributions. The MagLS method has the lowest level of colouration across evaluated orders.



(a) 1OA

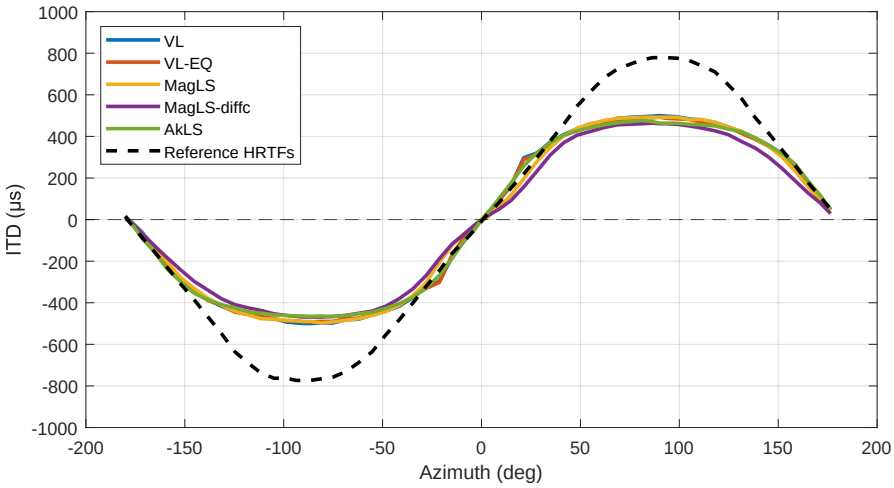


(b) 3OA

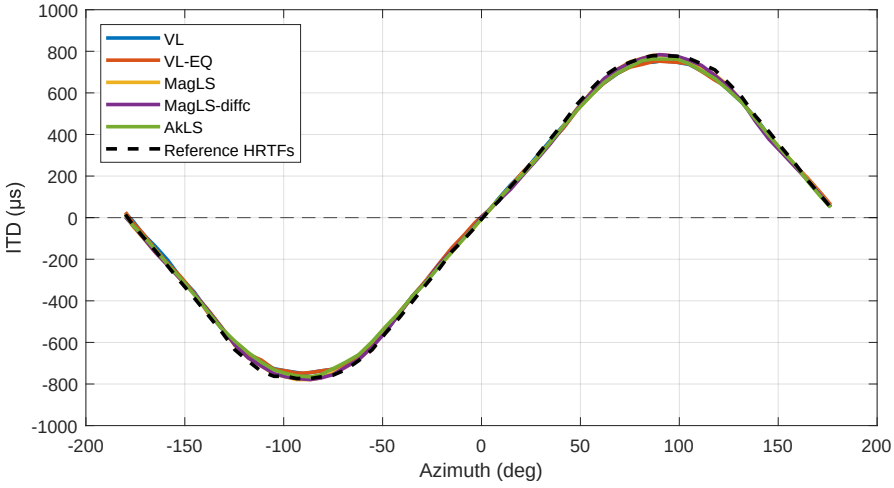


(c) 5OA

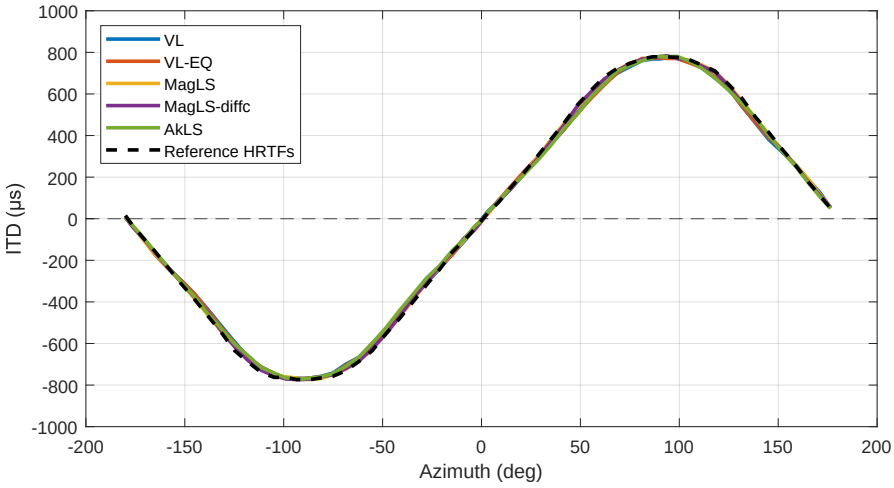
Figure 7.5: Diffuse-field responses of reconstructed HRTFs.



(a) 1OA

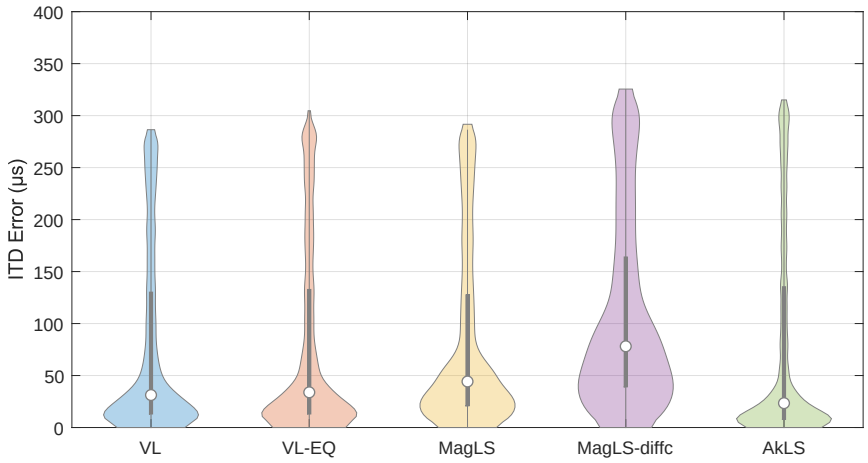


(b) 3OA

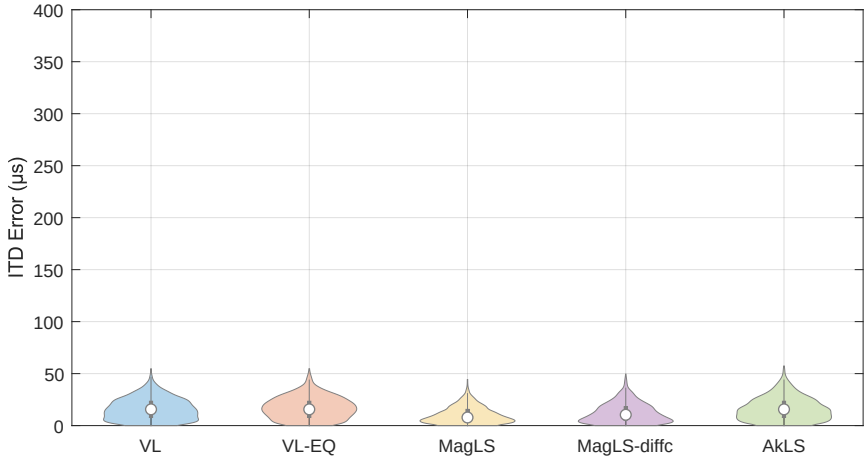


(c) 5OA

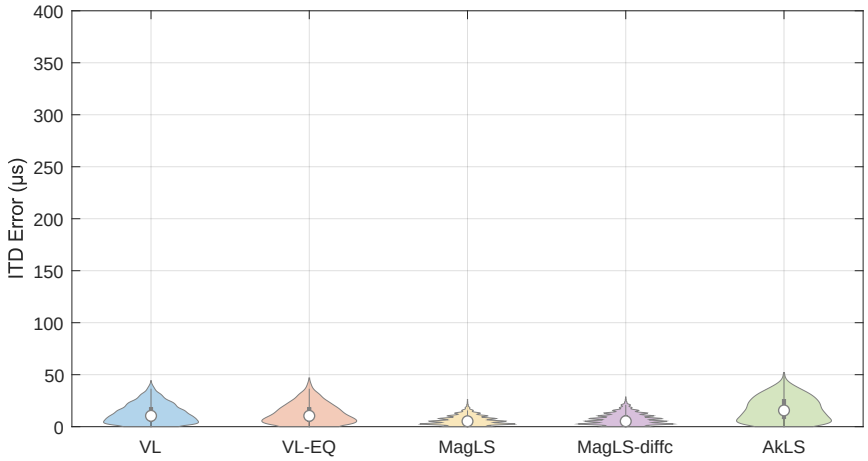
Figure 7.6: Horizontal plane ITDs of reconstructed HRTFs.



(a) 1OA

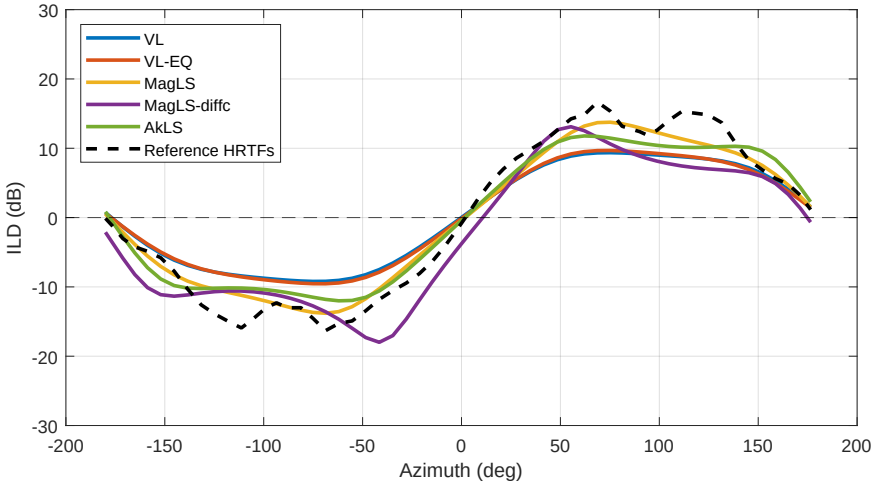


(b) 3OA

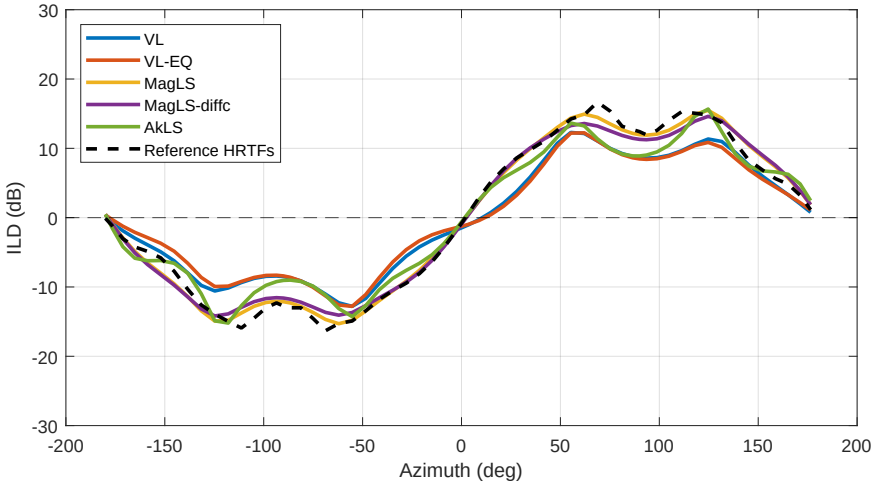


(c) 5OA

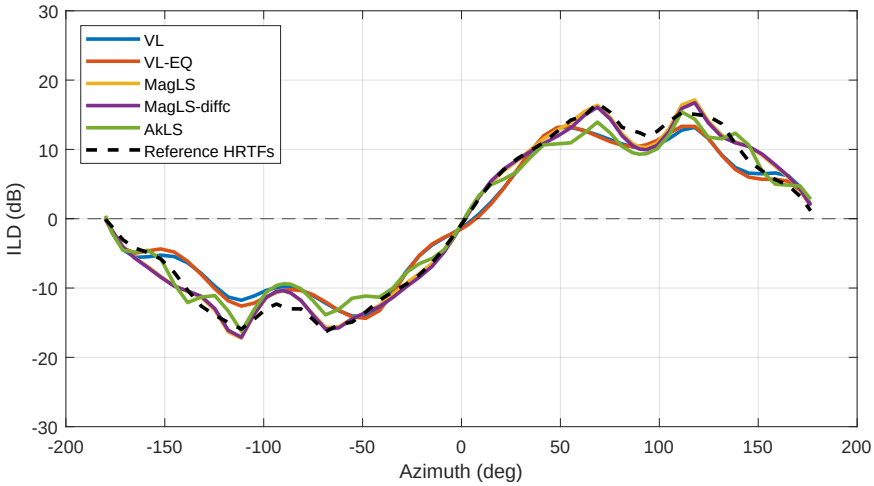
Figure 7.7: Reconstructed HRTFs ITD error distributions.



(a) 1OA

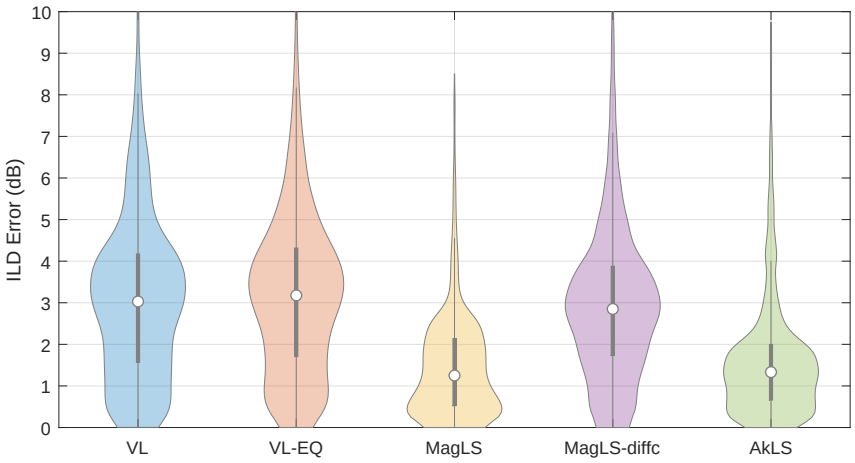


(b) 3OA

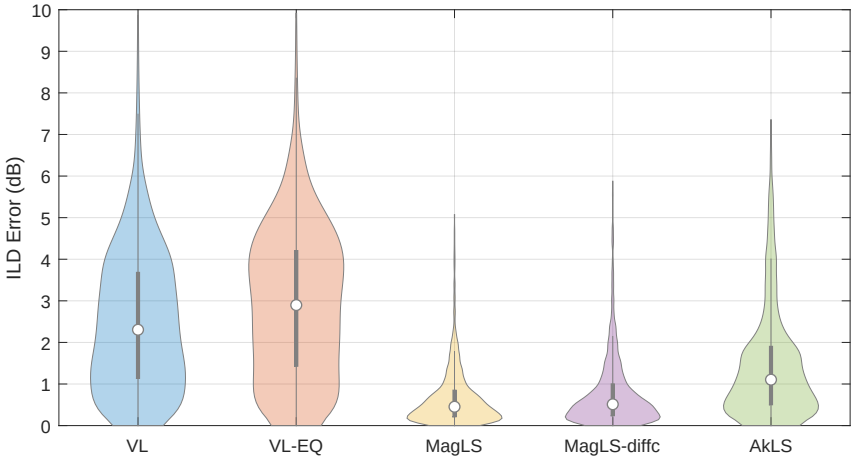


(c) 5OA

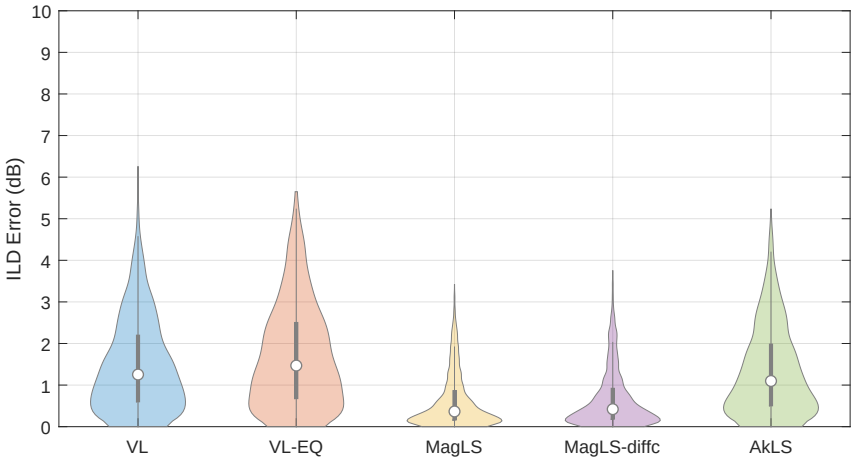
Figure 7.8: Horizontal plane ILDs of reconstructed HRTFs.



(a) 1OA

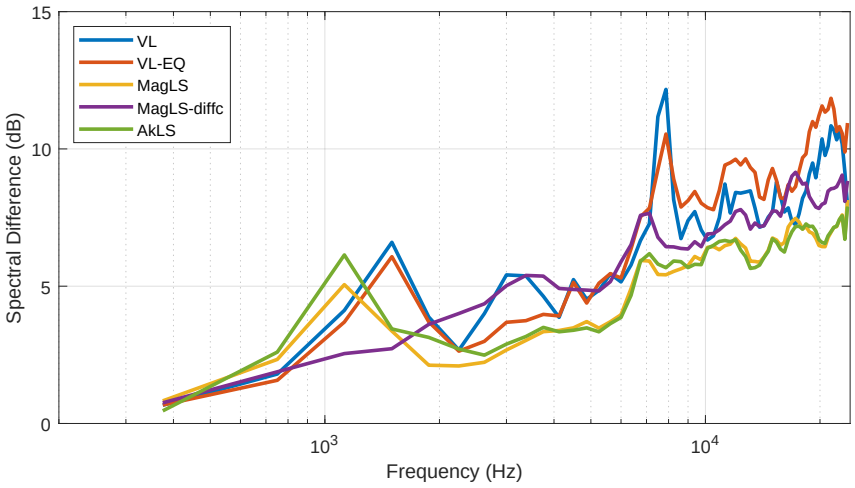


(b) 3OA

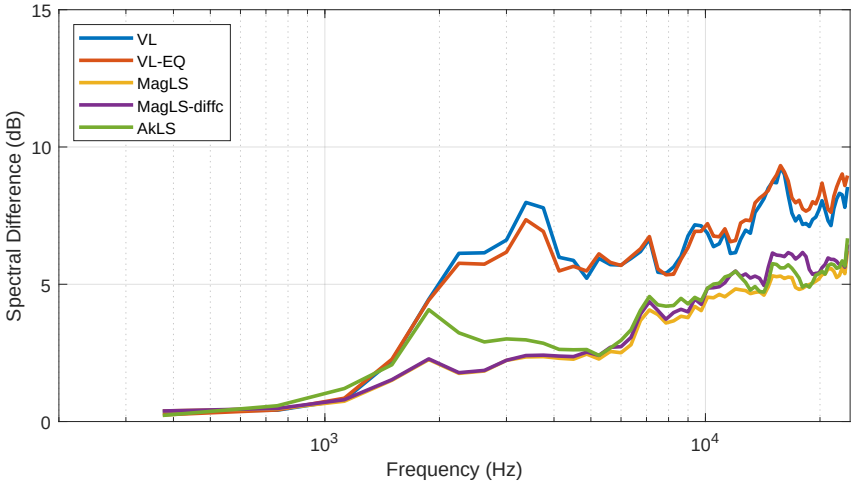


(c) 5OA

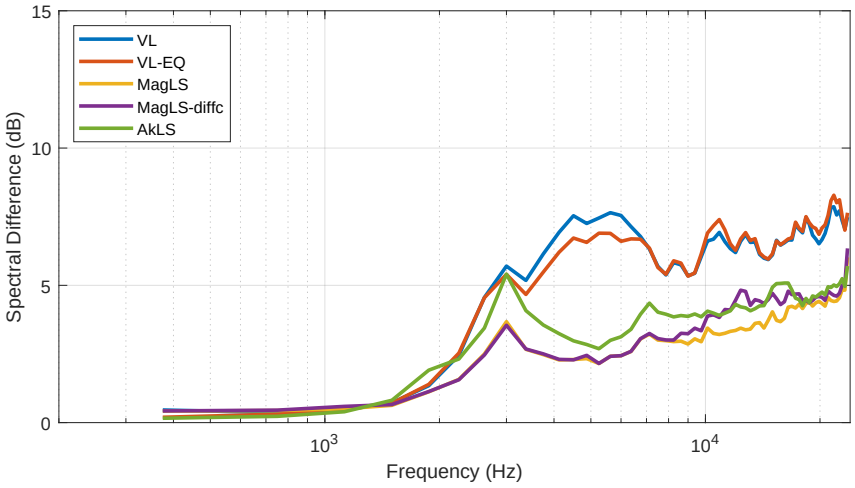
Figure 7.9: Reconstructed HRTFs ILD error distributions.



(a) 1OA

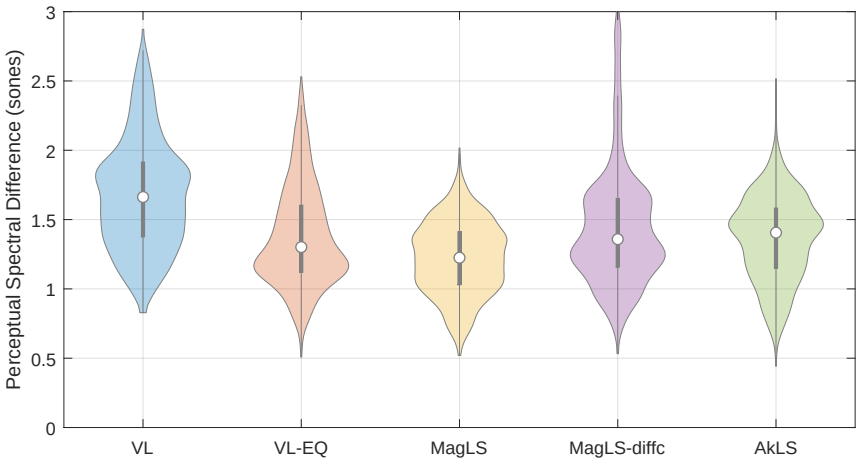


(b) 3OA

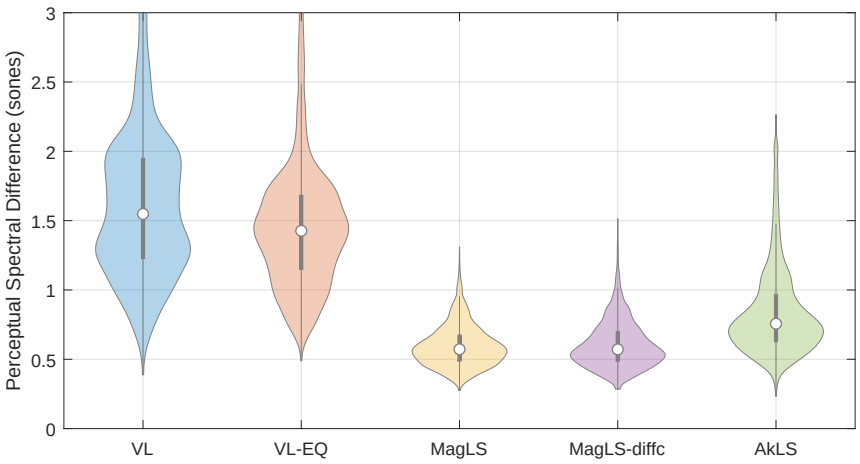


(c) 5OA

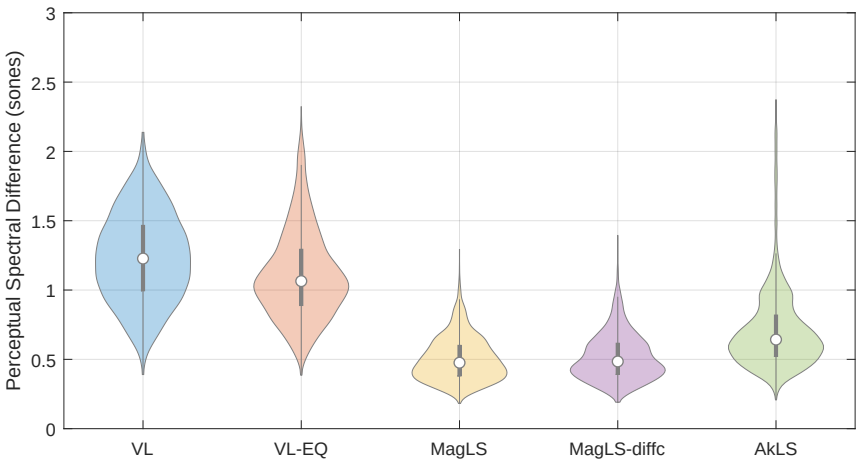
Figure 7.10: Average Spectral Difference.



(a) 1OA



(b) 3OA



(c) 5OA

Figure 7.11: Perceptual Spectral Difference distributions.

7.3 Subjective Evaluation

The subjective study used a subset of methods evaluated objectively. Based on the objective analysis results, the AkLS method is eliminated from the subjective evaluation, as it didn't show any advantage over the MagLS method. Therefore, four different methods (VL, VL-EQ, MagLS and MagLS-diffc) were evaluated at 1st, 3rd and 5th-order Ambisonics.

7.3.1 Methods

Listening tests were carried out based on the ITU-R BS.1534 (MUSHRA) (ITU-R, 2015b) recommendation to assess the degree of BAQ deterioration introduced by Ambisonic processing using different methods for delivering the binaural audio at different Ambisonic orders. The test participants were asked to rate the level of overall similarity on a continuous quality scale in relation to the Reference audio sample, integrating both spatial and timbral aspects of the presented conditions into a single judgement. The Reference condition was created by direct convolution with HRIRs. Due to the large number of experimental conditions in this test, only one anchor was included in the assessment. The anchor was created by applying a low-pass filter to the Reference condition at 7 kHz and summing left and right ear signals.

The listening test software employed for this test was the modified webMUSHRA web-based listening test environment (Schoeffler et al., 2018). The modification of the software allowed for a total of 14 conditions in each trial, instead of the maximum of 12. To make it easier for the participants to switch between each of the experimental conditions and the reference, a copy of the reference condition trigger button was added above all condition trigger buttons spanning across the interface width, as shown in Figure 7.12. The experiment was conducted using Sennheiser HD650 headphones.

The test stimuli set consisted of two types of scenes: simple scenes containing a single sound source and complex scenes consisting of either music or dialogues and sound effects panned around the listener. The simple scenes were created based on two excerpts from the EBU SQAM dataset[2]. Each of the audio samples was panned in three different directions, as listed in Table 7.2. This resulted in a total of six simple scenes. The remaining three scenes were complex, created based on the radio drama excerpts from the S3A ADM dataset (Woodcock et al., 2016) rendered to 7.0.4 loudspeaker setup using the EBU ADM renderer[3] and encoded to Ambisonics. Table 7.3 lists the audio material used for the stimuli creation.

All participants were PhD students and staff of the AudioLab experienced in critical listening and sound quality assessment. Participants were instructed on how to perform the test by reading the information sheets and receiving verbal instructions. All participants gave their informed consent to be included in the study. The protocol was approved by the Physical Sciences Ethics Committee of

[2] `https://tech.ebu.ch/publications/sqamcd/`

[3] `https://github.com/ebu/ebu_adm_renderer/`

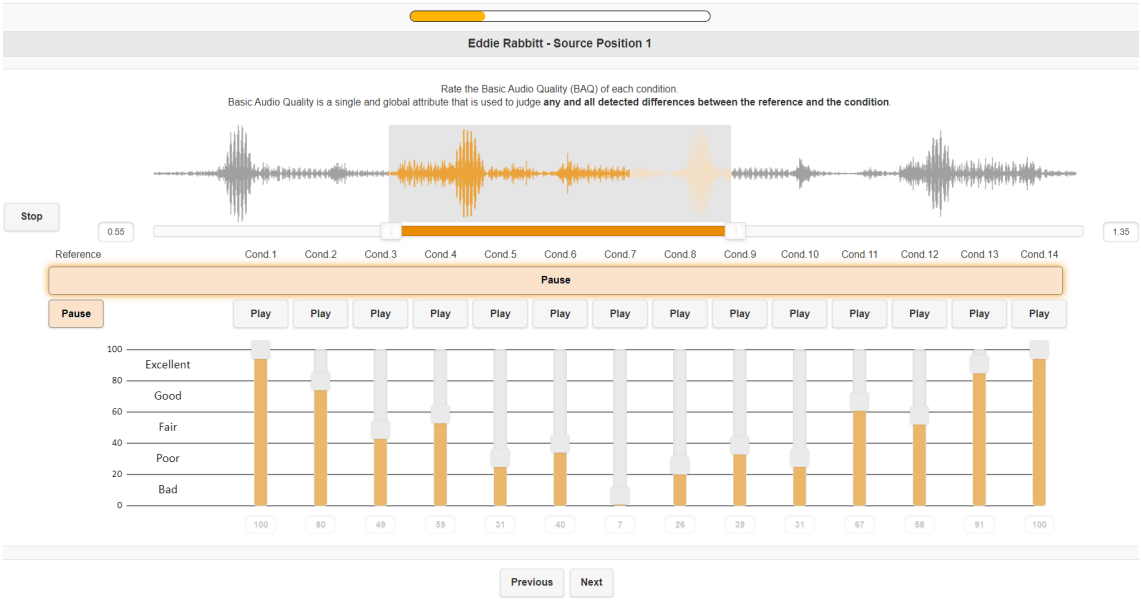


Figure 7.12: Modified webMUSHRA test interface.

Table 7.2: Sound source directions in simple scenes.

Direction	1	2	3
Azimuth (°)	34	88	-62
Elevation (°)	19	6	-5

Table 7.3: Audio material used to produce test stimuli.

Identifier	Scene type	Description
src1_castanets	Simple	Castanets (EBU)
src1_eddie	Simple	Eddie Rabbit (EBU)
scene1_music	Complex	Music Intro (S3A ADM)
scene2_forest	Complex	Forest Soundscape (S3A ADM)
scene3_monster	Complex	Monster Sound (S3A ADM)

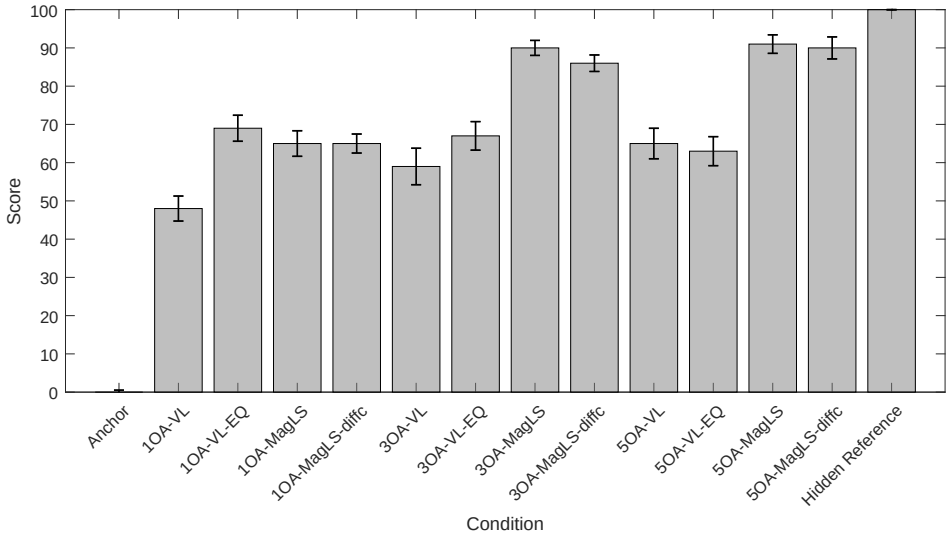


Figure 7.13: Median scores aggregated over all test scenes. The whiskers indicate nonparametric 95% confidence intervals.

the University of York (approval code: Rudzki20230131). A total number of 16 participants (all males) took part in this experiment.

7.3.2 Results

Figure 7.13 shows the median scores aggregated over all experimental scenes. The MagLS method received the highest rating for 3rd and 5th-order Ambisonic reproduction. However, the equalised virtual loudspeaker-based reproduction garnered the highest rating at 1st order. It can be observed that for the MagLS conditions, the difference between 5th-order and 3rd-order rendering is much smaller than the difference between 3rd-and 1st-order one. While increasing order does not produce a difference for the VL-EQ rendering.

Figure 7.14 shows the median scores for different types of audio scenes. At 1st order, the VL-EQ method received slightly higher ratings than both MagLS methods for complex scenes. At higher orders, MagLS is rated either equally to MagLS-diffc or higher, while the VL methods score worse than the MagLS methods.

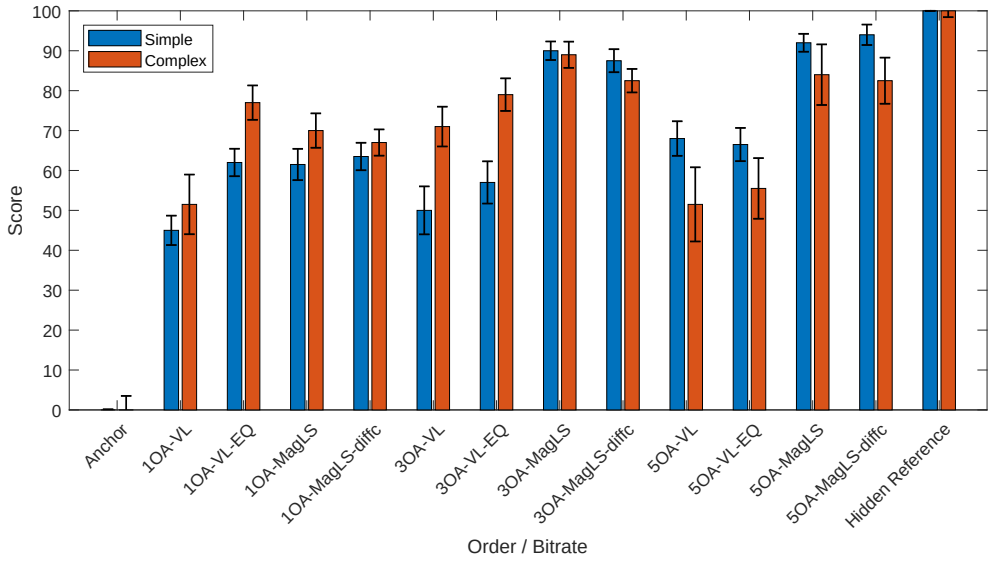


Figure 7.14: Median scores for each audio scene type. The whiskers indicate nonparametric 95% confidence intervals.

7.4 Conclusion

Both objective and subjective evaluations clearly indicated that the MagLS method provides the most perceptually accurate reconstruction of the reference HRTFs at the 3rd and 5th Ambisonic orders. However, the results are not as clear for the 1st-order rendering, where the equalised virtual loudspeaker method garnered the highest score for complex scene stimuli in the listening test. This contradicts the results of the objective analysis, which pointed at MagLS as the optimal filter design method.

The high scores for both MagLS methods at 3rd and 5th orders reconfirm the findings of the experiments in Chapter 4 and Chapter 5, i.e., Ambisonics should be delivered using at least 3rd-order signals for improved perceived quality.

Chapter 8

User Preference Evaluation of Direct-to-reverberant Ratio of Virtual Ambisonic Listening Spaces

In the last Chapter, the binaural decoding of Ambisonics using anechoic filters was evaluated. However, although binaural rendering schemes are based primarily on spatialisation via measured or simulated sets of HRTFs, a convincing experience also requires rendering of the room acoustics component, typically via measured or simulated BRIRs. In the context of audio and video content production and consumption, binaural rendering comes down to a simulation of a set of loudspeakers placed in a typical listening room or recording studio control room. Therefore, audio professionals use a wide range of mostly software solutions aimed at recreating the sound of such rooms over headphones, e.g. Waves Abbey Road Studio 3[1] or APL Virtuoso[2]. On the hardware side, the Smyth Realiser[3] has been a popular and respected system among the audiophile community.

Typically, the binaural rendering of Ambisonics is based on filters derived from anechoic HRIR sets. This allows a good reconstruction of binaural and spectral cues caused by human morphology, as shown in Chapter 7. However, such filters do not contain any acoustic response of the listening space, although Ambisonics was conceptualised originally as a loudspeaker-based reproduction format. Therefore, we could argue that listening to Ambisonic recordings via anechoic filters will provide a substantially different experience than listening on a multichannel loudspeaker array placed in a room with controlled acoustics. This difference might arise from the inaccuracies of HRTF-based headphone rendering, as well as the missing acoustic response of the listening space in the binaural signal. One way to mitigate the disparity coming from the lack of room response is to use Ambisonic rendering filters derived from measured BRIRs, e.g. obtained from the SADIE II database (Armstrong et al., 2018b), or to synthesise them by combining filters obtained using different methods, e.g. simulated SH-domain RIRs and anechoic SH-HRIRs filters.

[1] `https://www.waves.com/plugins/abbey-road-studio-3/`
[2] `https://apl-hud.com/product/virtuoso/`
[3] `https://smyth-research.com/`

The multichannel loudspeaker playback of Ambisonic scenes in a real room arguably gives a more convincing experience in comparison with the headphone playback. Perceived differences include mostly better externalisation and sound envelopment. Employing hybrid binaural filters combining room acoustics with anechoic filters might provide benefits to the headphone user in terms of perceived envelopment, externalisation and overall preference. At the same time, they may also cause significant timbre colouration (Crawford-Emery and Lee, 2014; Giller et al., 2019), decreased localisability, and lack of clarity (as with any reverberant room response added beyond measure).

Based on the official YouTube VR channel[4], the most popular 360 video content falls into the following categories: Music, Virtual Tours, Storytelling, and Gaming. Therefore, it is pertinent to research different contexts.

This chapter presents an experiment exploring user preferences of the direct-to-reverberant sound energy ratio (DRR) of such hybrid binaural rendering filters in relation to different types of reverberation and different categories of Ambisonic audio content. Participants were asked to find the preferred value of DRR for different Ambisonic scenes and artificial reverberation methods. After the experiment, each participant was asked to identify the perceived sound characteristics affected by the DRR adjustment.

8.1 Background

DRR is a single-value acoustic parameter conveniently describing the ratio of direct-to-reverberant sound energy within a signal. Based on a discrete-time room impulse response $h(n)$, DRR can be computed as follows:

$$\mathrm{DRR} = 10\log_{10}\left(\frac{\sum_{n=n_{pk}-\frac{1}{2}n_{win}}^{n_{pk}+\frac{1}{2}n_{win}} h^2(n)}{\sum_{n=n_{pk}+\frac{1}{2}n_{win}}^{\infty} h^2(n)}\right), \tag{8.1}$$

where n_{pk} is the sample index of the direct sound peak, and n_{win} is the length of the time window covering the direct sound impulse. Typically, the window length is set to ca. 2.5 ms (Møller, 1992; Zahorik, 2002b).

It is known to contribute to the perception of distance in reverberant environments (Bronkhorst and Houtgast, 1999). Adjusting the DRR artificially is also a standard sound engineering technique, which is typically achieved by sending a signal to an FX bus that feeds a reverb processor. Alternatively, the ratio can be controlled using specific processor controls, e.g. Dry/Wet, Ambience, Focus, etc.

The hybrid rendering concept was initially explored by Rudrich and Frank (2019) using state-of-the-art MagLS anechoic filters and simulated RIRs using an image source model (ISM). The authors focused on the perceived change in externalisation, timbre and localisation with different settings of the ISM while rendering dry Ambisonic signals. They have found an inverse relationship between perceived distance

[4] `https://vr.youtube.com`

and signal colouration. The trade-off between the two can be controlled by a number of simulated reflections, whereas satisfying results can be achieved with a small number of virtual loudspeakers and low ISM reflection orders.

8.2 Methods

This experiment uses a hybrid approach to the binaural rendering of Ambisonics, using two types of filters simultaneously. Direct sound is rendered using MagLS filters derived from the KU100 HRIR set (Bernschütz, 2013), which garnered the best ratings in Chapter 7 for higher-order rendering. Reverberant sound, which carries the acoustics of a virtual listening space, is rendered using the virtual loudspeaker approach. The DRR was chosen as the experimental dependent variable because its adjustment allows for a continuous and straightforward blending of the anechoic binaural filters with reverberant ones obtained through measurement or simulation. Based on the findings presented in previous chapters, 1st-order rendering is suboptimal and 5th-order would pose challenges in terms of content availability. Accordingly, and to limit the required time, this experiment used only 3rd-order Ambisonic rendering.

8.2.1 Experimental Procedure

Participants were asked to adjust the DRR while listening to binaural audio material using Sennheiser HD650 headphones and set the ratio to their preferred value. During this task, the participants were wearing a Quest 2 VR headset. The plain visuals displayed by the headset prevented participants from seeing visual cues present in the room where the experiment was taking place to avoid any interaction of the space with their auditory judgements.

Before the experiment, they were instructed on how to use the interface. They also participated in a short training session to familiarise themselves with the task and the VR environment. Both scene switching and DRR adjustment were made using one of the HMD's hand controllers. A vertical movement of the joystick caused a change in DRR. The rate of change was proportional to the displacement of the joystick and was indicated visually in VR. The actual DRR setting was not revealed to participants through visuals. Therefore, participants had to rely on what they were hearing to make their choice. The DRR adjustment covered a range from -24 dB to 36 dB, and the participant was notified of reaching the end of the scale by a controller vibration impulse.

After the listening task, participants were asked to complete a final questionnaire asking them to identify perceived audio quality attributes affected by the DRR adjustment. All participants gave their informed consent to include them in the study. The protocol was approved by the Physical Sciences Ethics Committee of the University of York (approval code: Rudzki20230131). A total of 17 participants took part in this experiment.

Table 8.1: Evaluated Ambisonic scenes.

Identifier	Original format	Reverberation type	Description
jazz	Ambisonics	club	The Vicente Magalhães Jazz Band - "Gaivota"; Recorded at the Abbey Road studios; Mixed by David Rivas. (Rivas Méndez et al., 2018)
piano	Ambisonics	hall	Eigenmike EM32 recording; Piano excerpt from the 3D-MARCo dataset (Lee and Johnson, 2019).
quartet	Ambisonics	hall	Eigenmike EM32 recording; String quartet excerpt from the 3D-MARCo dataset (Lee and Johnson, 2019).
rock	Ambisonics	room	Rock music track; Produced and mixed by Jacob Cooper.
hippie	Ambisonics	hall	Ambient track; Produced and mixed by Jacob Cooper.
tomfast	Ambisonics	club	Vocals and acoustic instruments; Tom McKenzie - "Give It a Go".
viola	Ambisonics	room	Anechoic viola recording encoded to Ambisonics using IEM RoomEncoder.
speech	Ambisonics	dry	Anechoic speech encoded to Ambisonics.
fight	Ambisonics	dry	Gaming sound design - Fight; Produced by Robert Hucknall. (Hucknall, 2023)
drone	Ambisonics	room	Gaming sound design - Drone in a warehouse; Produced by Robert Hucknall. (Hucknall, 2023)
pirates	Stereo	club	Music track encoded to Ambisonics at ±30° azimuth; Norah Jones - "Chasing Pirates". (Jones, 2009)
heart	Stereo	club	Music track encoded to Ambisonics at ±30° azimuth; Norah Jones - "Cold, Cold Heart". (Williams, 1950)
omar	5.1.2	hall	Music track encoded to Ambisonics; Omar Hakhim - "Listen Up!". (Levison, 2006)
forest	7.0.4	dry	ADM Radio drama rendered to 7.0.4 encoded to Ambisonics; S3A dataset (Woodcock et al., 2016) - "Forest".
protest	7.0.4	room	ADM Radio drama rendered to 7.0.4 encoded to Ambisonics; S3A dataset (Woodcock et al., 2016) - "Protest".
gravity	7.1.4	dry	Atmos rendered to 7.1.4 encoded to Ambisonics; Gravity movie. (Cuarón, 2013)

8.2.2 Experimental Stimuli

Considering the increasing use of spatial audio for music, this experiment emphasises the music content. The other audio scenes include sound design for games, radio dramas and a movie soundtrack. A total of 16 different scenes were used. Table 8.1 lists all Ambisonic scenes used in the experiment. All scenes were presented using 3rd-order Ambisonic rendering, although some of them were produced initially using 5th or 7th order Ambisonics.

8.2.3 Binaural Room Impulse Responses

The virtual listening spaces included in the experiment used the following types of BRIRs:

- **Listening Space A:** Measured KU100 BRIRs obtained from the SADIE II database (Armstrong et al., 2018b).
- **Listening Space B:** Simulated SH-domain RIRs obtained using the MCRoomSim shoebox room simulator (Wabnitz et al., 2010) convolved with anechoic SH-HRIR filters derived from KU100 HRIRs (Bernschütz, 2013).

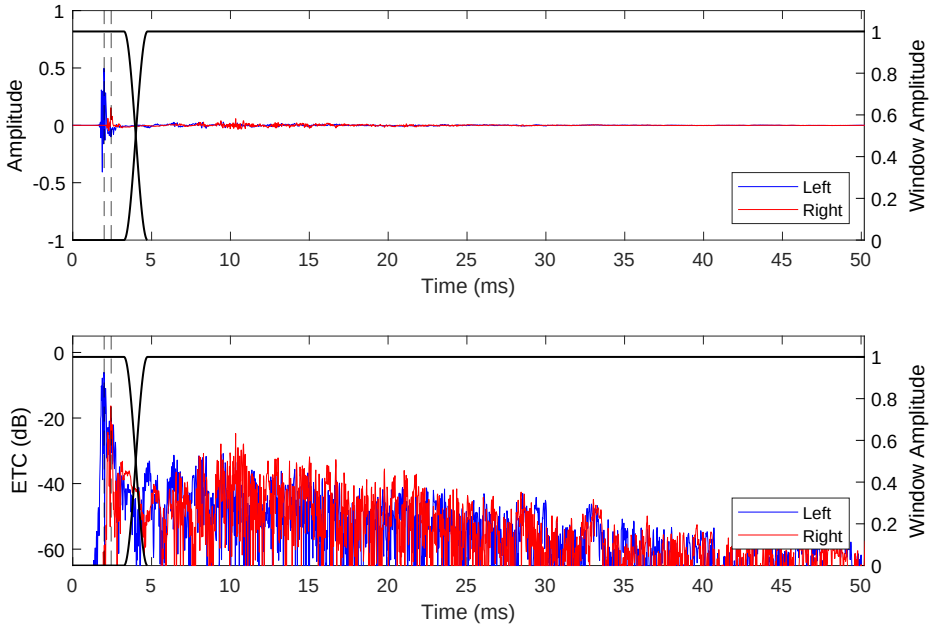


Figure 8.1: An example BRIR and the time window applied to separate direct and reverberant parts. Left and right ear impulse responses are plotted using blue and red colours, respectively.

The measured set of BRIRs (Space A) was obtained using the 50-channel loudspeaker array located at the AudioLab at the University of York. A detailed description of the measurement configuration has been provided by Armstrong et al. (2018b). For this experiment, the BRIRs were processed using a time window separating direct sound from the reverberant part, as shown in Figure 8.1. The mean DRR value for these 50 BRIRs was 7.54 dB originally.

The simulated set of 7th-order SH-domain RIRs (Space B) was obtained using the MCRoomSim shoebox room simulator (Wabnitz et al., 2010). Figure 8.2 shows the simulated room, the sound sources and the receiver. To match the source positions of the BRIR set measured in the laboratory space, the sources were arranged according to the 50-point Lebedev layout. The dimensions of the shoebox model and its acoustical properties were specified according to the size and reverberation guidelines of the ITU-R BS.1116 recommendation (ITU-R, 2015a). Table 8.2 lists the simulation parameters. BRIRs were created by convolving simulated SH-domain RIRs with 7th-order MagLS filters derived from KU100 HRTFs (Bernschütz, 2013).

Figure 8.3 shows the mean values and standard deviations of RT60 obtained from both measured and simulated BRIRs. The indicated ITU-R BS.1116 recommendation limits were calculated based on the volume of the simulated room.

8.2.4 Reverberant Ambisonic Rendering

Subsequently, SH-domain binaural filters were calculated using the virtual loudspeaker method for both direct and reverberant parts. 26-point Lebedev subset was used to calculate the binaural filters. The direct part of the filters was used to shift the

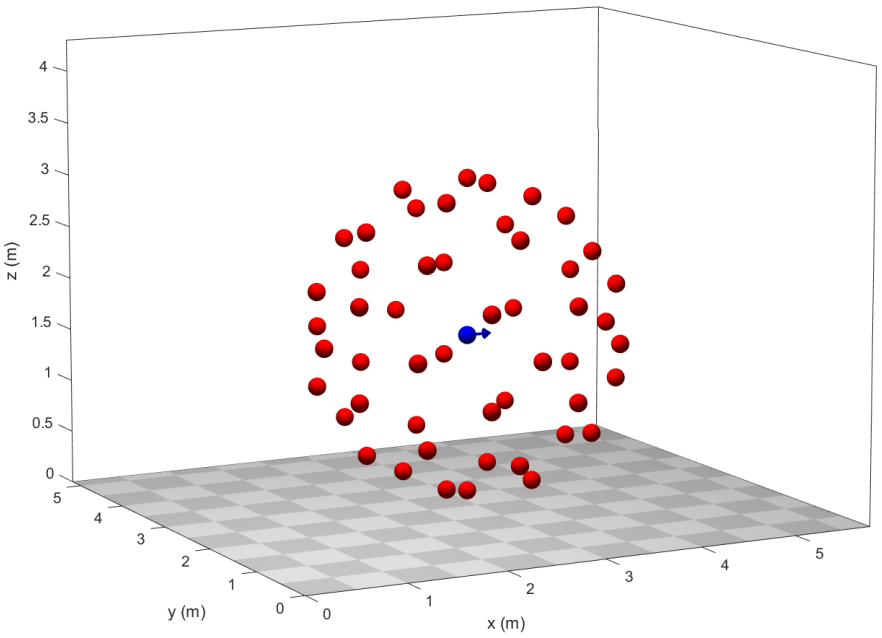


Figure 8.2: 3D view of the simulated shoebox room model (Space B). The red spheres represent sound sources. The blue arrow is the receiver.

Table 8.2: Simulation parameters.

Room dimensions	5.2 m x 5.8 m x 4.3 m
Volume	130 m^3
Temperature / Humidity	21°C / 62%
Number of sources	50
Source array origin	3.0 m; 2.7 m; 1.6 m;
Source array radius	1.5 m
Source directivity	omnidirectional
Receiver type	7th-order SH microphone
Simulation type	ISM and ray-tracing
Number of rays	6000
Diffuse reflection rate	45000
Time step	0.0015 s

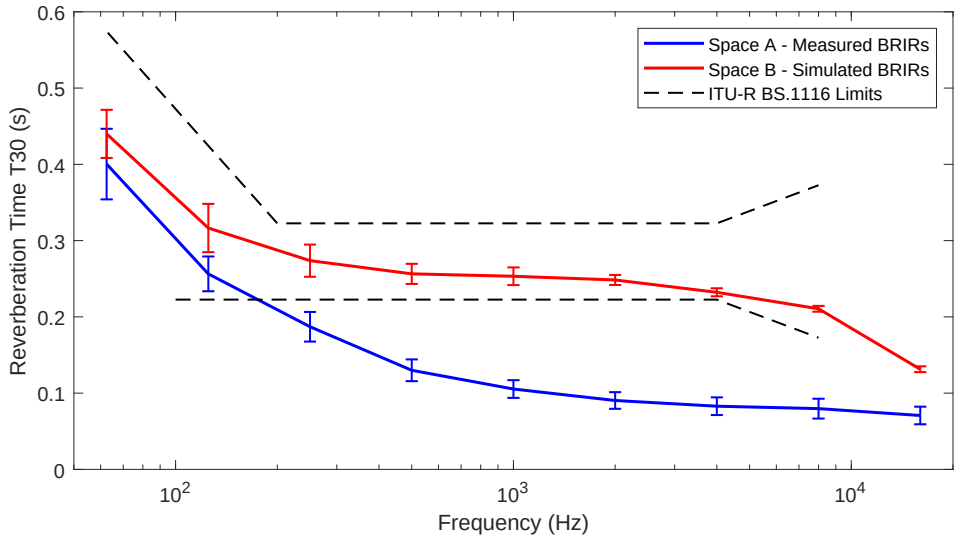


Figure 8.3: Estimated reverberation time based on the measured and simulated BRIRs.

SH-BRIR onsets, matching them with the anechoic binaural filter calculated using the MagLS method. The reverberant filters were then cut in length to 4096 samples at 48 kHz sampling rate (ca. 85 ms). The gain of the reverberant binaural filters was adjusted to match the energy of its diffuse-field response with the diffuse-field energy of the MagLS anechoic filters. Therefore, the initial DRR ratio between the two was 0 dB.

8.2.5 Real-time Rendering

Binaural rendering was done in Max, running two instances of the mcfx_convolver plugin. The VR HMD provided head-tracking data for audio rendering via OSC messages. The experiment was controlled by SALTE for VR[5] app running on the headset.

As noted by Zahorik (2002b), varying DRR by scaling the direct or reverberant energy components results in changes to the overall stimulus level. Therefore, loudness compensation is required to facilitate the continuous adjustment of DRR to avoid potential skew of the experimental results by inconsistent loudness. Prior to the experiment, the loudness of the binaural output was measured at varying DRRs, while the decoders were fed a spatially diffuse pink noise scene. The measurement was carried out using an ITU-R BS.1770-3 (ITU-R, 2012) compliant analysis. As seen in Figure 8.4, a non-linear relationship between loudness and DRR is observed for DRR values in the range from -10 dB to 20 dB. To model this change, a 5th-degree polynomial was fitted across the experimental DRR adjustment range for both types of reverberant filters. The polynomial coefficients and other test settings were then saved in the test control app. The level of the signal fed through the direct and

[5] `https://github.com/trsonic/SALTE4Quest-XRIT`

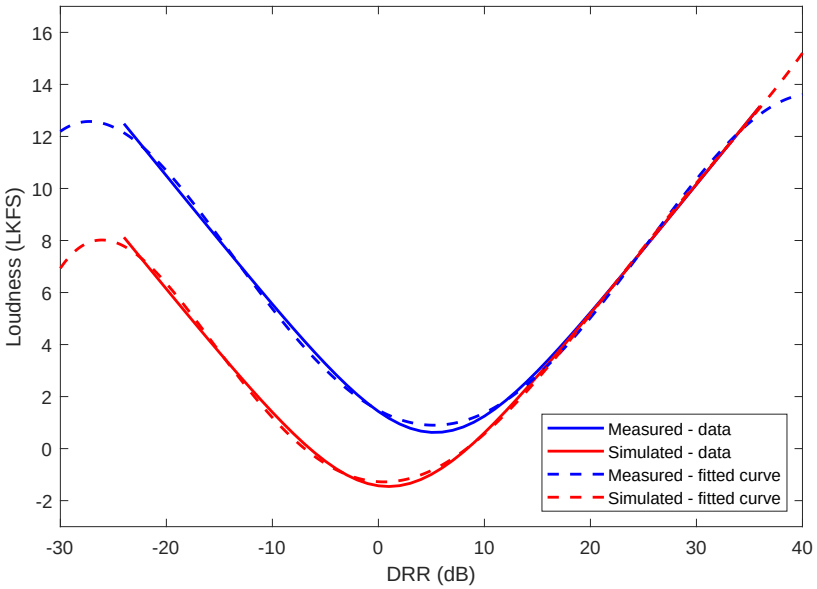


Figure 8.4: DRR adjustment and loudness compensation.

reverberant renderers was controlled during the experiment as follows:

$$L_{direct} = \frac{DRR}{2} - L_{comp}(DRR),$$
$$L_{rev} = -\frac{DRR}{2} - L_{comp}(DRR).$$

The initial DRR value for each experimental trial was randomised in the 0–12 dB range.

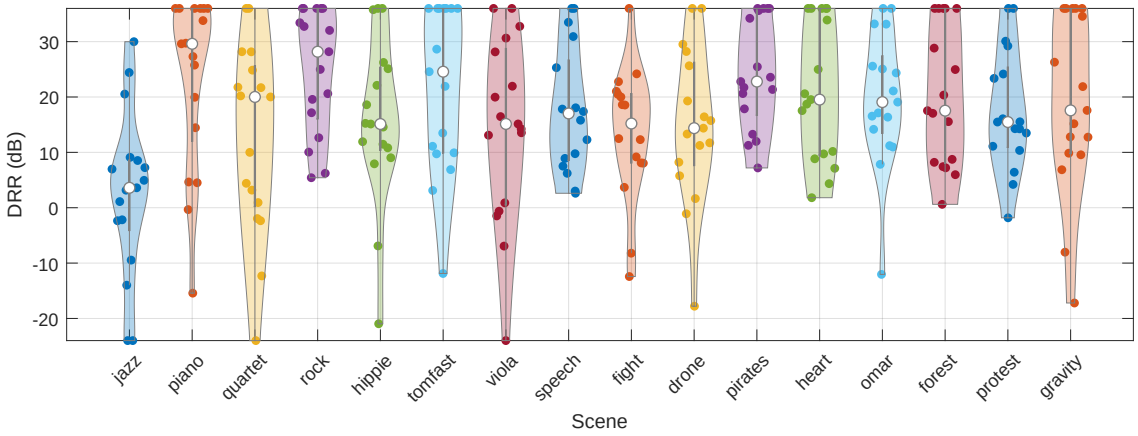
8.3 Results

The collected results consist of DRR adjustments made by participants during the experiment and their answers to questions included in the final questionnaire.

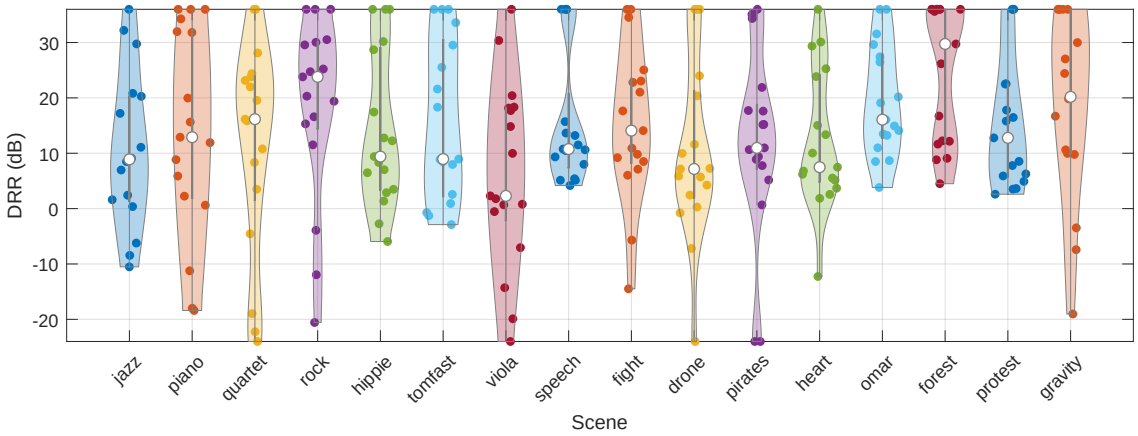
8.3.1 DRRs

Figure 8.5 shows the distributions of participants' responses for each audio scene and rendering condition. The responses are spread differently depending on the audio scene. It can be seen that for multiple conditions, the responses have bimodal distributions, with one mode located in the 0–20 dB range and the other towards the upper end of the DRR adjustment scale. The upper end corresponds to anechoic rendering.

The results combined across all audio contexts reveal differences in DRR preferences between Space A and Space B. Figure 8.6 shows probability density functions obtained using Gaussian kernels. The kernel bandwidth was set to 5 dB. Based on the shape of the distributions, participants were likely choosing between the balanced



(a) Listening Space A



(b) Listening Space B

Figure 8.5: Distributions of participants' responses for each audio scene and virtual listening space.

DRR value, providing a mix of anechoic and reverberant sound, and entirely anechoic rendering. When choosing the balanced condition, the preferred DDR value was set lower for Space B (9.2 dB), suggesting that its reverberation was more preferred than Space A (13.9 dB).

8.3.2 Questionnaire

This section provides a summary of participants' responses to the questions included in the final questionnaire.

Did you find the adjustment procedure challenging?
The mean response of the participants on a 5-point scale was 2.6, where *1* was *Not challenging at all*, and *5* was *Extremely challenging*.

Can you describe what change in the perception of audio scenes the adjustment caused?

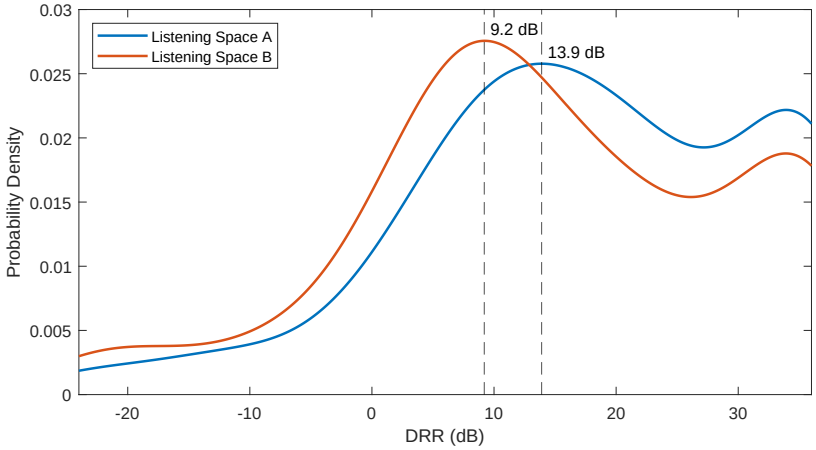


Figure 8.6: Probability density functions of DRR aggregated over all scenes for listening spaces A and B. The functions were estimated using Gaussian kernels of 5 dB bandwidth.

In general, the participants reported that the adjustment changed the perception of the audio scenes in a number of ways, including the amount of reverberation, the tonal balance, and the perceived distance between the listener and the sound scene. The specific changes that were reported varied depending on the scene and the individual participant. One participant said that there was a definite point at which the direct-to-reverberant ratio flipped from sounding good to very bad.

Can you describe what perceptual attributes did you focus on during the adjustment procedure?
The most common attributes that participants focussed on were: clarity, reverberation, externalization, distance and timbre. One participant said that they focused on how dry the signal appeared, and then they changed that depending on the nature of the stimuli. For example, they preferred much less reverb on anything that included speech or sound effects. A trade-off between externalisation and clarity was a typical response too.

Do you have any additional comments regarding the experiment?
Some participants commented that depending on the content, they could identify a sweet spot for the amount of reverberation. One participant noted that the "muddy" sounding reverb was preferred at lower levels relative to the direct sound. Another participant preferred direct sound rendering but found the effect of the reverb very realistic. The adjustment didn't make as much difference for some contexts as for others. The participants found the experiment interesting and liked the music selection.

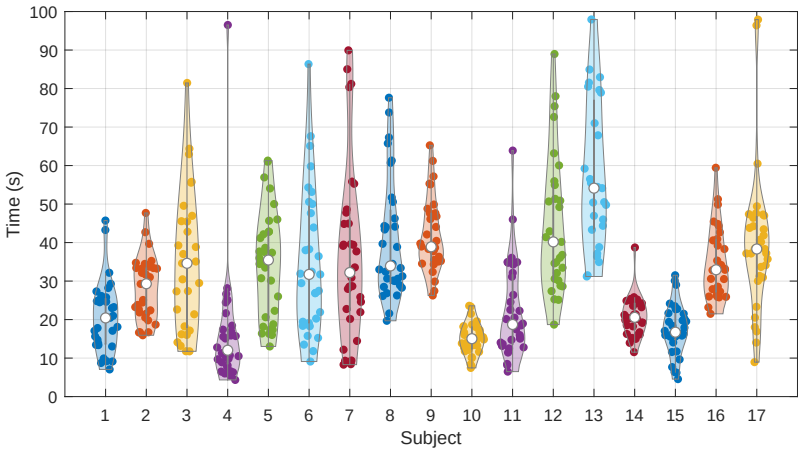


Figure 8.7: Distribution of time taken to decide on the preferred DRR value by each participant.

8.4 Discussion

This experiment provides an insight into which perceptual attributes matter to listeners when experiencing Ambisonic audio rendered over headphones, with clarity being identified the most often. The two virtual listening spaces employed provide different types of reverberation (as shown in Figure 8.3), which might have contributed to the difference in preferred DRR values. As one of the participants noted, the measured BRIRs were more "muddy" than the simulated ones, modelled according to standard guidelines. Seeing the bimodal distribution of user responses on Figure 8.6, two rendering strategies emerge, one hybrid with balanced DRR and the other using anechoic filters.

During the experiment, participants listened to audio scenes with different levels of reverberation already present. Scenes with a long reverb, e.g. 'piano' and 'quartet', had a larger spread in responses than the ones with less reverb ('speech', 'forest', 'protest').

Figure 8.7 shows how long each participant took to decide on their preferred DRR setting. It is interesting to observe that some participants took much more time than others.

It is worth noting that participants did not comment on any changes in loudness while adjusting DRR. This demonstrates that the method used for loudness compensation can be applied in further studies related to the problem of varying DRR in binaural reproduction.

8.5 Conclusion

This chapter explored the concept of hybrid binaural rendering combining anechoic filters with reverberant ones. The results show that such filters might provide an alternative to the established anechoic rendering. The reverberant filters can be obtained using available simulation software and the described workflow, while

anechoic filters can be calculated using the state-of-the-art MagLS method evaluated in the previous chapter.

A preferred DRR value for a virtual listening space with balanced reverberation characteristics (Space B) was approximately 9 dB. The results suggest that if such a hybrid approach is used, the listeners should be allowed to bypass the reverberant filters, as some prefer entirely dry rendering. Introducing a continuous DRR adjustment in the renderer could also be beneficial for some listeners.

Chapter 9

XR-based HRTF Measurements

Thus far in the thesis, the investigated components of the Ambinsonic delivery chain included optimal Ambisonic binaural decoders, low-bitrate coding and virtual listening room preference. A further important factor is the timbral and spatial distortions which can be introduced by using non-individualised Head Related Transfer Functions (HRTFs). The use of individual HRTFs is beneficial for the accurate and precise localisation of virtual sound sources rendered binaurally. The problem of obtaining individual HRTF filters is widely researched. However, such measurements typically require a dedicated laboratory space, extensive equipment and trained personnel to conduct the measurement session.

In a typical HRTF measurement setup, a human subject wearing binaural microphones is placed inside an anechoic or semi-anechoic room, while an arc consisting of multiple loudspeakers is being rotated around the subject in order to capture test signals emitted from a large number of directions (Algazi et al., 2001). Alternatively, the loudspeakers remain fixed and the subject is rotated in the horizontal plane using a motorized swivel chair (Armstrong et al., 2018b). While running such measurements, the subject sits or stands keeping their whole body still. Although such a scenario does not require the active participation of the subject in the procedure, it requires extensive infrastructure and some assistance from trained personnel.

Another way to measure HRTFs is by using a single speaker to minimise hardware requirements. A method where the loudspeaker is continuously rotated around a subject using a motorised boom was proposed by Pulkki et al. (2010). This scenario can be scaled down further if we assume the active participation of the subject who is asked to rotate their head and torso without any physical constraints and orient it at certain angles in relation to the single loudspeaker fixed on a stand. A system where the subject sits on a swivel chair and the loudspeaker height is accordingly adjusted was initially proposed by Brungart et al. (1998). More recently Li and Peissig (2017); He et al. (2018); Reijniers et al. (2020); Bau et al. (2021); Bau and Pörschmann (2022) proposed systems utilizing continuous arbitrary head movements performed by the subject while being tracked by a head tracking device. During such a procedure the head-above-torso orientation (HATO) varies, which has implications on the measured HRTFs (see Section 9.3). Modern XR headsets provide accurate six degrees of freedom (6DOF) spatial tracking, therefore it is feasible to use virtual or

augmented reality (VR/AR) headsets in order to track the subject's head orientation and provide visual feedback to the subject on the measurement progress. A system utilizing a VR headset was proposed before by Peksi et al. (2019), while Li et al. (2020) proposed the use of an AR headset for HRTF collection.

In this chapter, a system for easy and accessible single-speaker HRTF measurements using a mixed-reality headset is presented as well as the required post-processing of captured data. The system uses a novel virtual user interface in order to help the subject orient their head towards predefined directions in relation to the measurement loudspeaker. Contrary to the systems presented in the past, the software implementation of this system is freely available to the audio community, therefore it is possible to conduct independent validation experiments as well as allow other researchers to introduce improvements to the system.

9.1 Methods

9.1.1 Measurement Environment

A few factors need to be considered before attempting to measure HRIRs in non-anechoic conditions. In order to obtain far-field HRTFs, the distance between the head and the sound source should be greater than ca. 1.5 m. Another factor is the required duration of the time windowing function responsible for the attenuation of the room boundary reflections. A longer time window is beneficial, as it gives higher frequency resolution of the measured HRTFs (Kulkarni and Colburn, 1998). In a typical room, the ceiling is located at ca. 0.8 m above the head of a standing subject. Therefore it limits either the measurement distance or the duration of the windowing envelope. Equation 9.1 shows the relationship between the arrival time of the first acoustic reflection (known in acoustics as the initial time-delay gap (ITDG)), speed of sound c, the distance between the head and the source d_d and distance of the closest room boundary d_b, assuming that it is parallel to the head–source axis as seen in the Figure 9.1. Figure 9.2 shows this relation plotted for varying distances d_b.

$$ITDG = (2 \times \sqrt{\frac{{d_d}^2}{2} - d_b^2} - d_d) \times c^{-1} \tag{9.1}$$

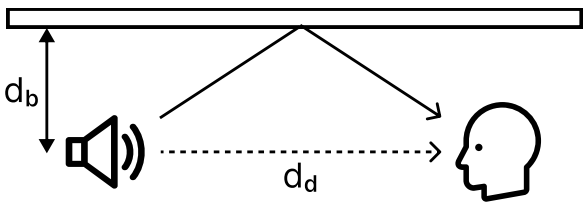


Figure 9.1: Direct and reflected sound paths.

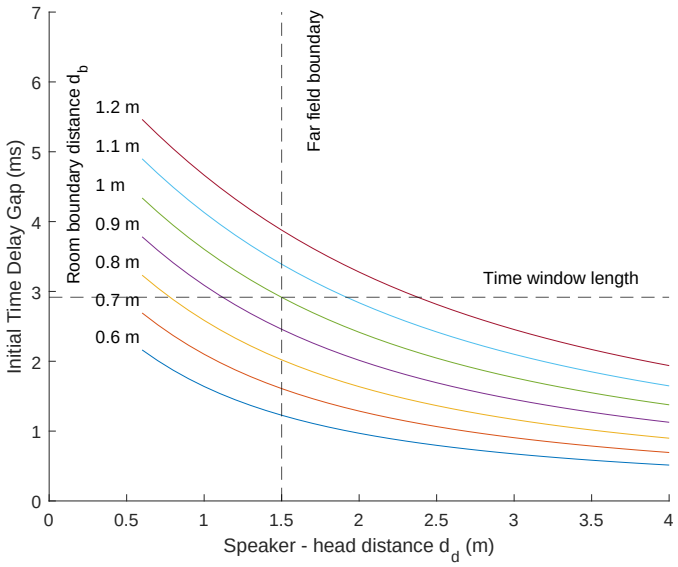


Figure 9.2: Relation between the first acoustic reflection time gap (ITDG) and subject's head distance from the source.

9.1.2 System Description

The proposed HRTF measurement procedure utilizes a minimal hardware setup consisting of the following devices:

- Mixed-reality headset (Oculus Quest 2 with a pass-through mode enabled)
- PC
- Audio interface (RME Fireface UC)
- In-ear microphones (Voice Technologies VT202)
- Loudspeaker and amplifier (custom 3D printed spherical loudspeaker enclosure, single Dayton Audio PS-95 driver, 50W digital amplifier)
- Wifi router

The system allows for the substitution of listed components (excluding the MR headset) with alternative devices available commercially. Figure 9.3 shows a block diagram of the system, while Figure 9.4 shows a subject during the measurement procedure. The use of a mixed-reality headset enables the subject taking part in the measurement to see 3D graphic cues overlaid on top of the monochromatic view of their surroundings (see Fig. 9.5). A custom measurement app run on the headset provides visual cues to help orient the subject's head in desired directions in relation to the fixed sound source position. The apparent direction and distance between the subject's head and the sound source are estimated in real-time based on the inside-out spatial tracking of the headset. The HRTF measurement process is controlled by another app running on a PC equipped with an audio interface (see Fig. 9.6). This

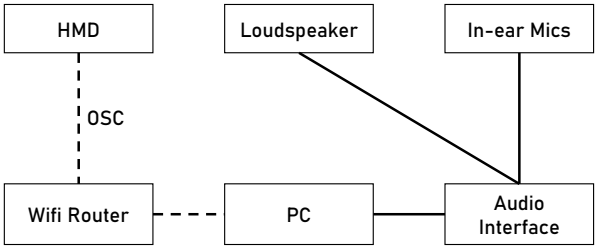


Figure 9.3: Block diagram illustrating components of the single-loudspeaker HRTF measurement system.

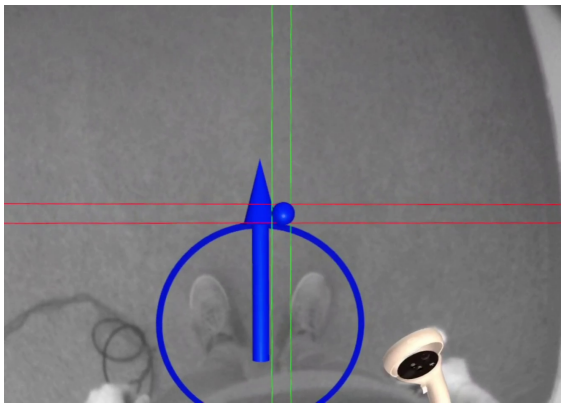
Figure 9.4: Subject during the measurement procedure.

app plays back the sweep signal as well as captures binaural microphone signals. The mixed-reality app was developed using Unity engine, Oculus SDK and Passthrough API. The measurement control app has been developed using C++ and the JUCE framework.

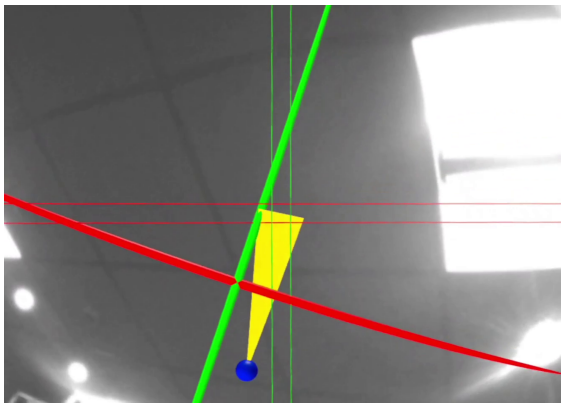
9.1.3 Measurement Procedure

First, the loudspeaker height should be adjusted to align it with the ears of a standing subject. Next, the speaker's position and orientation should be captured using one of the hand controllers. Then, the reference acoustic measurement with in-ear microphones placed at the predefined distance from the sound source should be taken. The reference measurement point is indicated with a virtual 3D marker displayed by the headset at the default distance of 1.5 m from the loudspeaker.

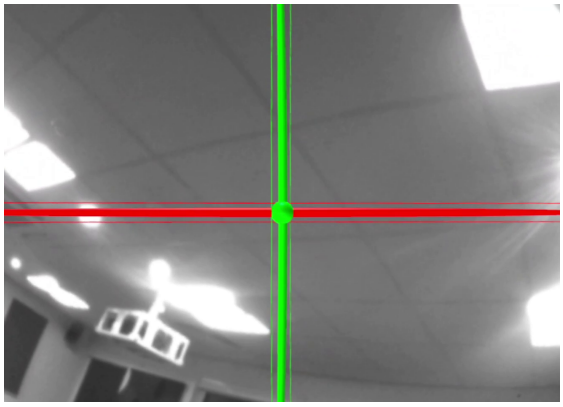
Before the start of the measurement procedure, the subject wears the microphones fitted into a soft, cylindrical foam fitting. The nominal dimensions of the microphone with the fitting are 9 mm in length and 8 mm in diameter. The fitting can be manually moulded before insertion into the ear canal similarly to standard foam-based earplugs. The wire is guided around the subject's ear. Subsequently, the subject wears the headset and is asked to move into the desired position and orient their head according to the visual cues displayed by the headset. During the



(a) Floor circle indicating desired standing position.



(b) Visual guiding towards the next measurement head orientation.



(c) Head orientation locked, measurement in progress.

Figure 9.5: First person view of the orientation guiding interface.

measurement, the surroundings are visible to the subject thanks to the pass-through mode of the VR headset. Their task is to follow the cues by standing inside a virtual circle displayed on the floor, orienting their torso towards the indicated direction and rolling their head in order to align the displayed guides. This allows for a controlled 'stop-measure-go' capture of binaural recordings at predefined directions and distances. The measurement starts automatically when the desired head position is reached. The subject is guided through a predefined list of measurement points. If the subject moves their head during the measurement, the procedure is stopped and repeated when the desired orientation is re-established. The measurement is concluded after capturing all directions. The time necessary to capture all Head Related Impulse Responses (the time domain equivalent of HRTFs) depends on the number of predefined directions, the length of the sweep signal and the subject's performance in following the visual cues. By default, the system uses 50 directions distributed regularly based on the Lebedev quadrature (Lebedev and Laikov, 1999) and a two-second exponentially-swept sine sweep (ESS) signal.

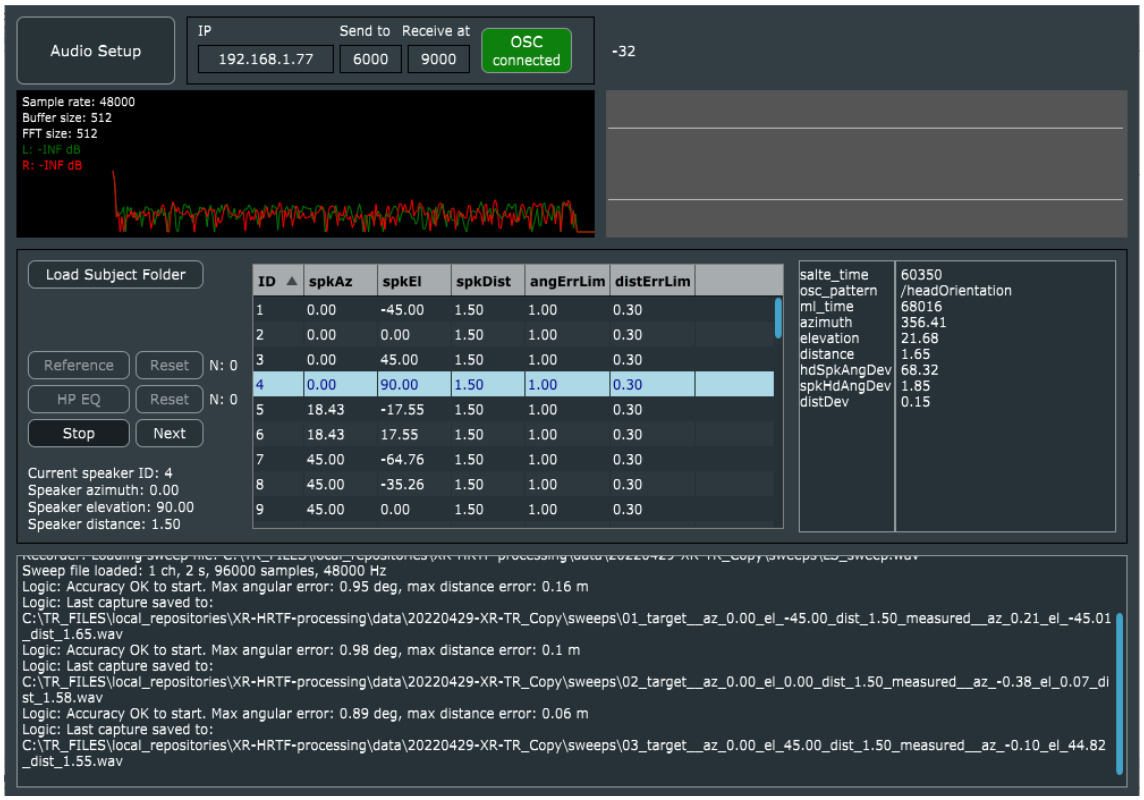


Figure 9.6: Measurement control and data acquisition app.

9.1.4 Data Post-Processing

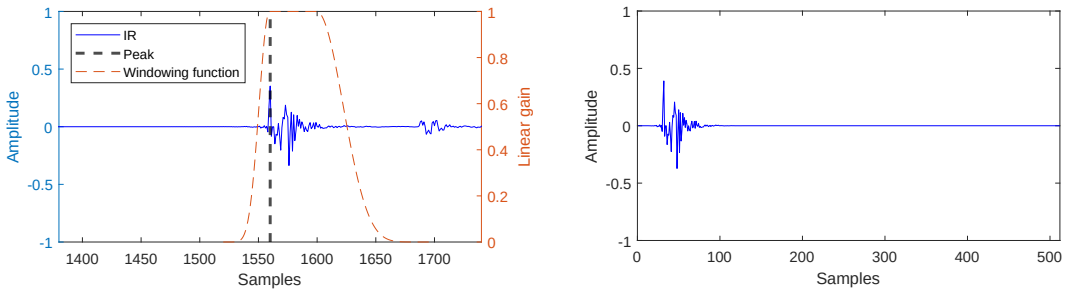
The measurement session results in a set of binaural recordings of the ESS signal reproduced using the loudspeaker. The post-processing of the recordings is done using MATLAB and consists of multiple steps necessary to produce binaural filters usable for spatial audio rendering.

Deconvolution

In order to transform binaural recordings into Binaural Room Impulse Responses (BRIRs), the recorded sweeps are convolved with an inverse filter. This filter is produced by time-reversing the original sine sweep and applying an amplitude attenuation envelope of 6 dB/oct (Farina, 2000). The obtained impulse responses characterize the linear component of the measurement system, the influence of the subject's morphology, and the acoustics of the room where the measurements were taken. These responses are also influenced by the acoustic shadowing of the XR headset used during the measurement process. The extent of these perturbations and a proposed HRTF correction filter are discussed in Section 9.2.

Time windowing

Subsequently, the relative time-of-arrival (TOA) of the test signal wavefront is identified for both left and right ear recordings. The TOA is indicated by the maximum



(a) Full BRIR (left ear).

(b) Windowed BRIR (left ear).

Figure 9.7: Time-windowing of BRIRs.

of the cross-correlation function between the original BRIR and its minimum-phase version (Nam et al., 2008). The accurate estimation of each wavefront arrival time is required in order to apply a custom time-windowing function to both BRIR signals. The purpose of this time windowing is to attenuate first and further reflections from the room boundaries, keeping only the part of the impulse response affected by the subject's morphology (ears and torso). According to (Xie, 2013) the main energy of HRIRs is concentrated in the first 1.4 ms. Figure 9.7 shows the used windowing function and time windowed response. The proposed window consists of a 40-sample rise followed by a 40-sample unity shelf and a 100-sample fall at 48 kHz sampling frequency. The rise and fall segments are modelled using a Hann function raised to the power of 4, in order to increase the attenuation curve steepness. The time-windowing function is aligned with each impulse response by matching the beginning of the unity shelf with the estimated wavefront arrival time.

Each measurement has gain correction applied to both signals according to the distance variation of the subject's head from the sound source and the inverse-square law. The reference distance is set by default to 1.5 m.

Normalization

The next step in obtaining HRIRs from measured BRIRs is to normalize the IRs captured at the entrance to the blocked ear canal to reference IRs measured in the middle of the head with subject absent (Møller, 1992). The reference measurement is deconvolved and time-windowed similarly to the binaural recordings and then used for the calculation of two minimum-phase inverse filters for left and right ear signals. These filters compensate for the magnitude response of the measurement sound source and microphones (see Figure 9.8).

The classical regularization method for inverse filtering proposed by Kirkeby and Nelson (1999) can be used to invert both minimum and non-minimum-phase filters within a specified frequency range while attenuating unwanted frequency components. This is done using a frequency-dependent regularization parameter, which must be carefully adjusted for each measurement system configuration to achieve the desired effect. In such cases, the non-inverted components of the measured spectrum can be either attenuated or left untreated. In order to avoid the multiplication

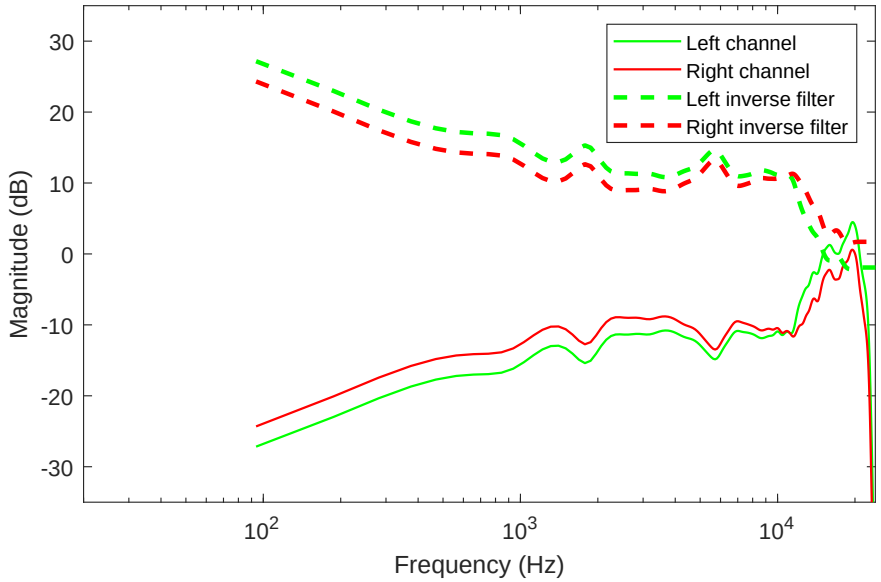


Figure 9.8: Magnitude characteristics of reference measurements and their respective inverse filters.

of frequency components of BRIR and reference measurement spectrum outside of the inversion range, we propose an alternative spectrum shaping approach to inverse filter regularization, provided that the original phase response of the reference measurement can be discarded in the HRIR normalization step.

First of all, a magnitude spectrum of the measured reference impulse response is obtained using an FFT. Secondly, a magnitude flattening at the high end of the spectrum is introduced. This is done by establishing the mean amplitude level across frequency bins in the 16–20 kHz range. This level is used to create a flat extension of the measured spectrum. The measured spectrum and the extension are merged using a simple cross-fade in the 18–20 kHz range. The cross-fade function is based on two halves of a Hann window applied as a linear gain at the specified frequency bins in order to provide a smooth transition between the actual response and the flat HF extension. Figure 9.9 shows the described procedure.

In order to avoid resonant peaks in the inverse filter response, the magnitude is smoothed out using Gaussian kernels of 1/12th octave standard deviation. This step, together with the fact that any room boundary reflections have already been removed from the measurement by time-windowing, provides a sufficient level of magnitude regularization for inverse filter calculation, which is done in the next step (see Equation 9.2).

$$iH = H^{-1} \tag{9.2}$$

A linear-phase FIR normalization filter is obtained using inverse FFT, and then its minimum-phase version is used to filter all time-windowed BRIRs.

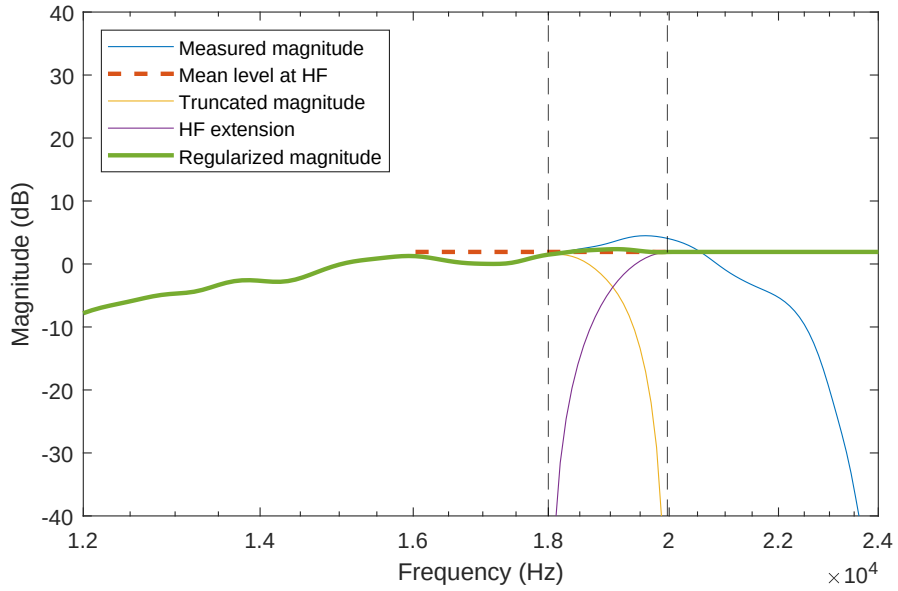


Figure 9.9: Reference measurement magnitude flattening at high frequencies.

Low-Frequency Extension

Measured HRIRs are prone to accuracy errors at low frequencies due to room resonances as well as the limited frequency response of the measurement source. The time windowing process introduces additional error, resulting in increased ILDs at low frequencies. However, in natural listening scenarios, head shadowing results in negligible ILD at low frequencies for HRTFs captured in the far field (Xie, 2013). Therefore a flat low-frequency response can be modelled using a Kronecker delta shifted in time to match the original measured HRIR onset. A similar approach was proposed by Bernschütz (2013). However, this work proposes different time and gain adjustments for the low-frequency extension pulse. The time-of-arrival estimation method described in Section 9.1.4 is used to time-align the LF extension pulse with the HRIR onset. The amplitude of the LF extension pulse for the left and right ear signal is adjusted to match the SPL difference between the ear locations in the free field according to the inverse-square law (see Equation 9.3, where d_{ee} is the interaural distance set to 0.16 m and d_{ref} is the reference sound source distance set to 1.5 m).

$$A_{LFE} = 1 \pm \frac{\frac{d_{ee}}{2} \sin\theta \cos\varphi}{d_{ref}} \tag{9.3}$$

The modelled response is blended with the original HRIR using a fourth-order Linkwitz–Riley crossover (24 dB/oct) set at 250 Hz. Figure 9.10 shows the magnitude response of the original HRIR, filtered components and the extended response.

Diffuse-Field Equalization

The DFE process starts with left and right ear HRTF magnitude averaging. Each direction is weighted by its corresponding solid angle value, as in Equation 9.4, where

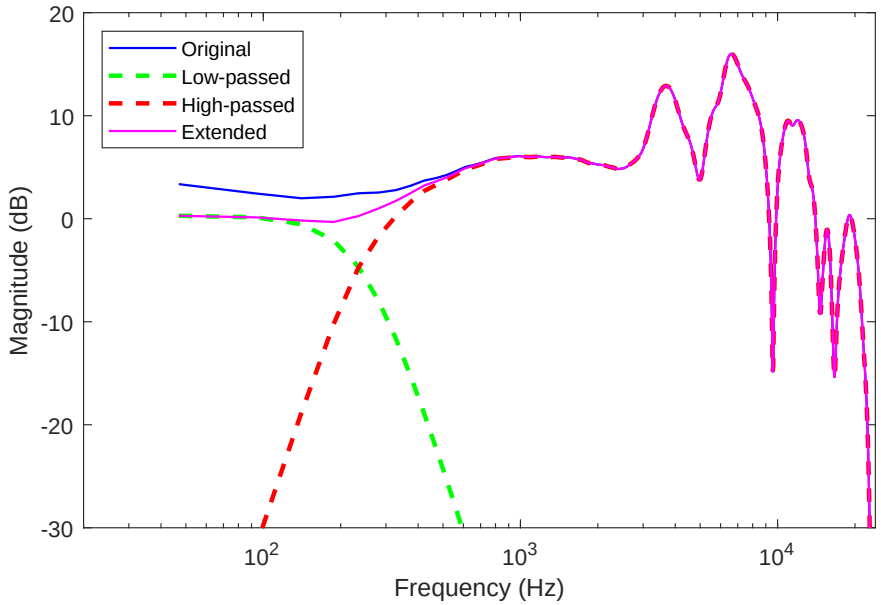


Figure 9.10: Low-frequency extension.

K represents the total number of HRTFs and sa is the solid angle expressed in steradians.

$$HRTF_{DFE} = \frac{HRTF_{RAW}}{\sqrt{\frac{1}{4\pi}\sum_{k=1}^{K}|HRTF(k)|^2\, sa(k)}} \tag{9.4}$$

The DFE filter is calculated based on the average of both left and right ear signals, as shown in Figure 9.11. Subsequently, the HRIRs are filtered with the minimum-phase version of the DFE filter.

HRIR Interpolation and Export

Measured HRIRs are exported as SOFA files (Majdak et al., 2013) in RAW (non-equalized) and DFE (equalized) versions. Additionally, both HRIR sets are interpolated at 2354-pt Lebedev grid vertices. The employed interpolation uses a barycentric coordinate system. Firstly, a convex hull is created based on the measurement grid. A ray-triangle intersection-seeking algorithm is used to find the three vertices needed to calculate each of the interpolated points. Subsequently, barycentric weights are calculated for the queried/interpolated directions. The interpolated HRIRs are created by a weighted summation of the chosen three time-aligned HRIRs. The same weights are used to obtain an interpolated value of ITD, which becomes a target value for the time alignment procedure employing fractional circular array shift. This interpolation method is yet to be evaluated against other methods, e.g. using spherical harmonics (Arend et al., 2021).

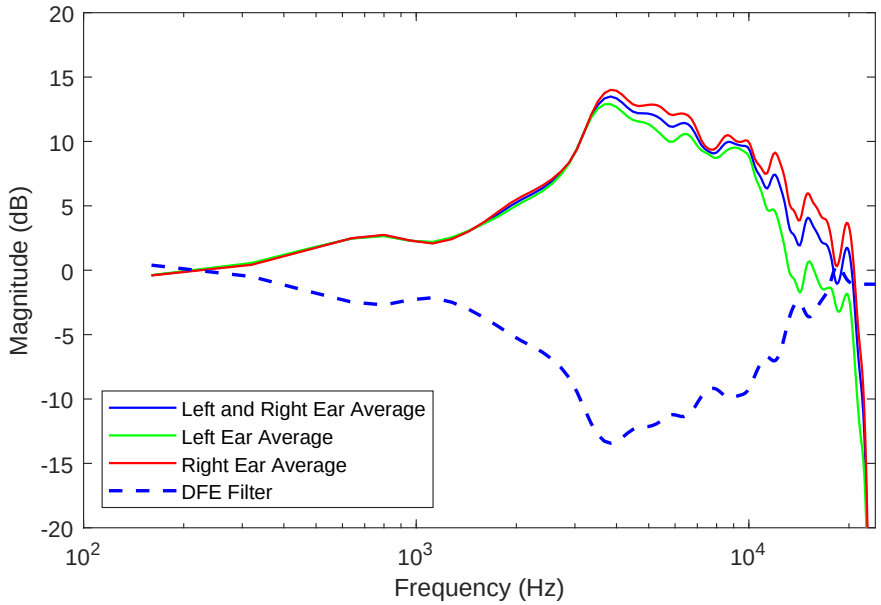


Figure 9.11: Averaged HRTF magnitudes and Diffuse-field equalization filters.

9.2 HMD Influence on HRTFs

Previous studies have shown that wearing a Head Mounted Device (HMD) affects the subject's HRTFs. Gupta et al. (2018) evaluated differences in the KEMAR manikin's HRTFs wearing different headsets measured in the horizontal plane. Spectral differences introduced by HMDs were observed at the middle and high frequencies. Genovese et al. (2018) used a KU100 dummy head and two AR headsets to conduct analysis at different azimuths and elevations. Ahrens et al. (2019) ran sound localization experiments in the context of using HMDs to display virtual environments while rendering audio using a multichannel loudspeaker array, maintaining the user's HRTFs. In that study, the measured auditory localization accuracy error slightly increased when the subjects were wearing an HMD. Another study conducted by Cuevas-Rodriguez et al. (2019) revealed the perceptual significance of the HMD influence on the measured individual HRTFs. However, this effect was smaller in comparison to using a generic HRTF set. Pörschmann et al. (2019) tested various types of headgear using two types of dummy heads. The Rift headset and a baseball cap provided the lowest spectral differences among all tested conditions. All studies indicated that the contralateral ear signal is affected more than the ipsilateral one due to the occluding effect of the headset. Although the results of previous objective studies seem to correlate well with each other, the extent of perceptually verified degradation of HRTFs by wearing an HMD remains unclear. In an attempt to compensate for the objectively measured distortions, a procedure of obtaining and applying direction-dependent correction filters is proposed.

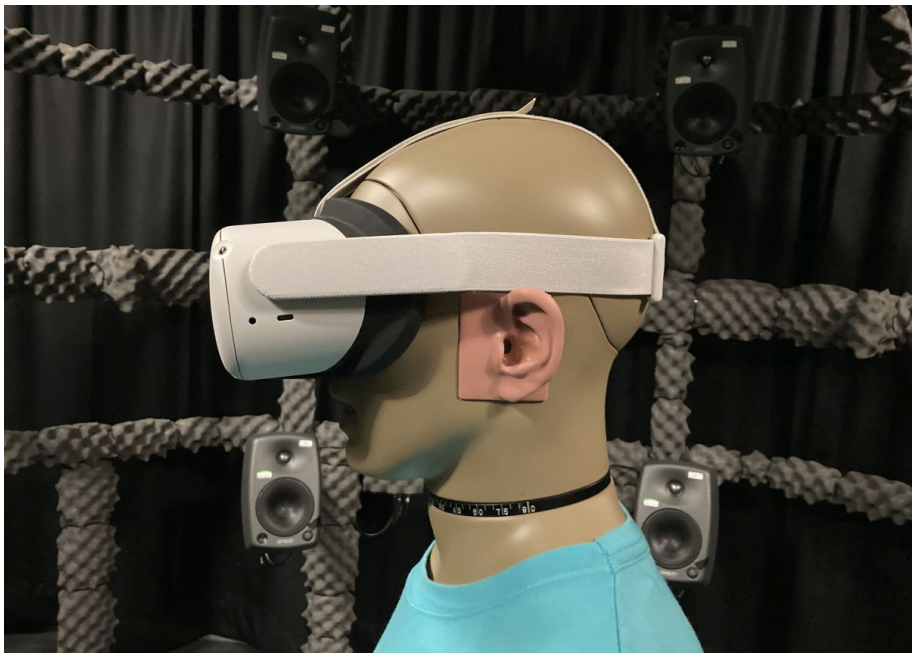


Figure 9.12: KEMAR manikin wearing Quest 2 headset.

9.2.1 KEMAR HRTF Measurements

The proposed HRTF measurement system uses a Quest 2 headset which has not been evaluated in previous studies. Thus it is desired to investigate the influence of this headset on measured HRTFs. For this purpose, a KEMAR manikin's HRTFs were captured inside a 50-pt Lebedev grid loudspeaker array with and without the headset on (see Figure 9.12). In order to increase the spatial resolution of the captured data, the measurements were repeated six times, while the KEMAR manikin was rotated every 15° around its vertical axis between the measurements. This resulted in 266 unique directions being analyzed. For arrival time and spectral difference analysis, data measured with both ears have been aggregated into a single-channel dataset by reversing the azimuthal coordinates of the right ear channel.

The analysis of Interaural Time Difference (ITD), Interaural Level Difference (ILD) errors and spectral difference was performed on non-normalized time-windowed binaural IRs. Figure 9.13 shows measured ITD error for the KEMAR manikin equipped with a Quest 2 headset referenced to the measurement without a headset. The measured error is higher in the frontal and upper hemispheres, which might be caused by the fact that XR headsets are typically located in the front, slightly above the user's ears, thus causing a significant obstruction for the sound coming from that direction. The maximum ITD error of 75 µs was measured at +/-45° azimuth and 0° elevation sound source incidence. This result suggests that HRTFs measured with a Quest 2 headset on will exhibit ITD inaccuracy exceeding JND values (Mills, 1958) at these particular directions.

Figure 9.14 shows measured ILD error. The ILD error can be observed in the frontal hemisphere with its maximum value of 3 dB at +/-27° azimuth and -18° elevation.

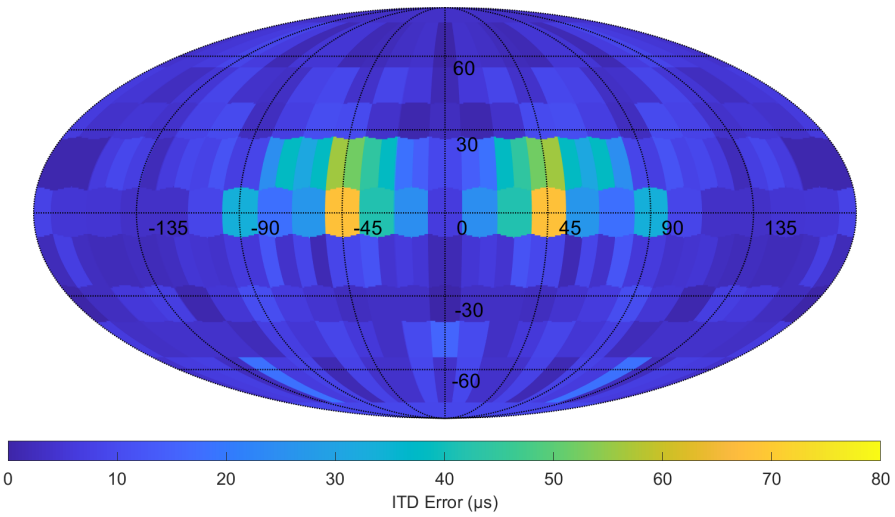


Figure 9.13: ITD error of KEMAR HRTFs introduced by the Quest 2 headset.

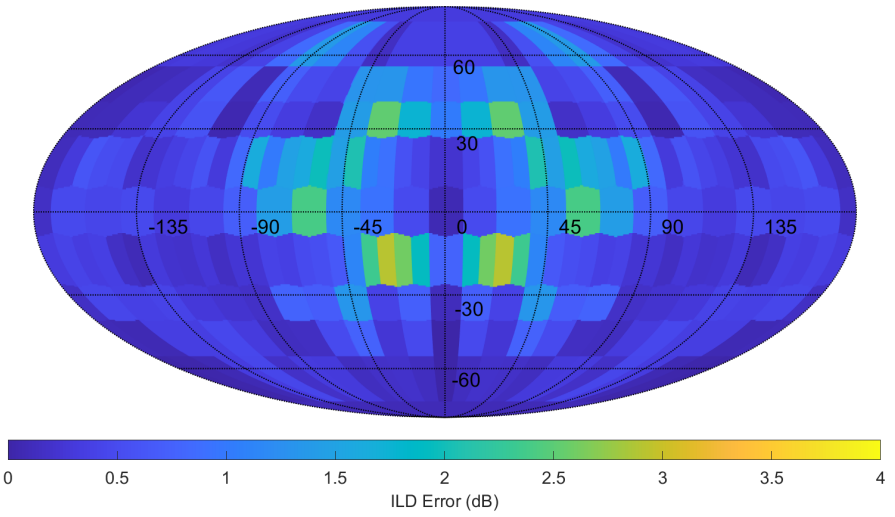


Figure 9.14: ILD error of KEMAR HRTFs introduced by the Quest 2 headset.

Spectral difference was analyzed using ERB bands following the methodology proposed by Gupta et al. (2018). Figure 9.15 shows the spectral difference at three frequency bands: 0.1–1 kHz, 1–5 kHz and 5–16 kHz. The acoustic shadowing effect of the headset can be observed at the second and third band for the contralateral sound source incidence.

9.2.2 Measured HRTF Correction

The conducted KEMAR measurements using Quest 2 headset suggest that the device affects the following spatial cues: ITD, ILD and contralateral ear HRTF spectrum. Assuming that the HMD affects HRTFs of real subjects similar to those of KEMAR, a method of delivering direction-dependent HRTF correction filters based on the set of KEMAR measurements is proposed. It consists of the following steps:

- Time-of-arrival difference extracted from measurements with and without HMD.
- Magnitude spectrum difference calculated.
- Both measures are interpolated at a very dense regular layout, eg. 4334-pt Lebedev grid.
- For each interpolated direction, a TOA correction is calculated, as well as a minimum phase inverse filter for the magnitude difference.
- Each HRTF measurement of the human subject is corrected using correction data of the nearest point from the set.

Interpolation is done using distance-based weights. A one-sided Gaussian window, defined over 0° to 180° distance, is used as a weighting function. The standard deviation of the Gaussian window can be adjusted to control the degree of spatial smoothing of the measured data, therefore providing a smooth transition of correction parameters across all directions. Figure 9.16 shows interpolated arrival time difference introduced by the headset.

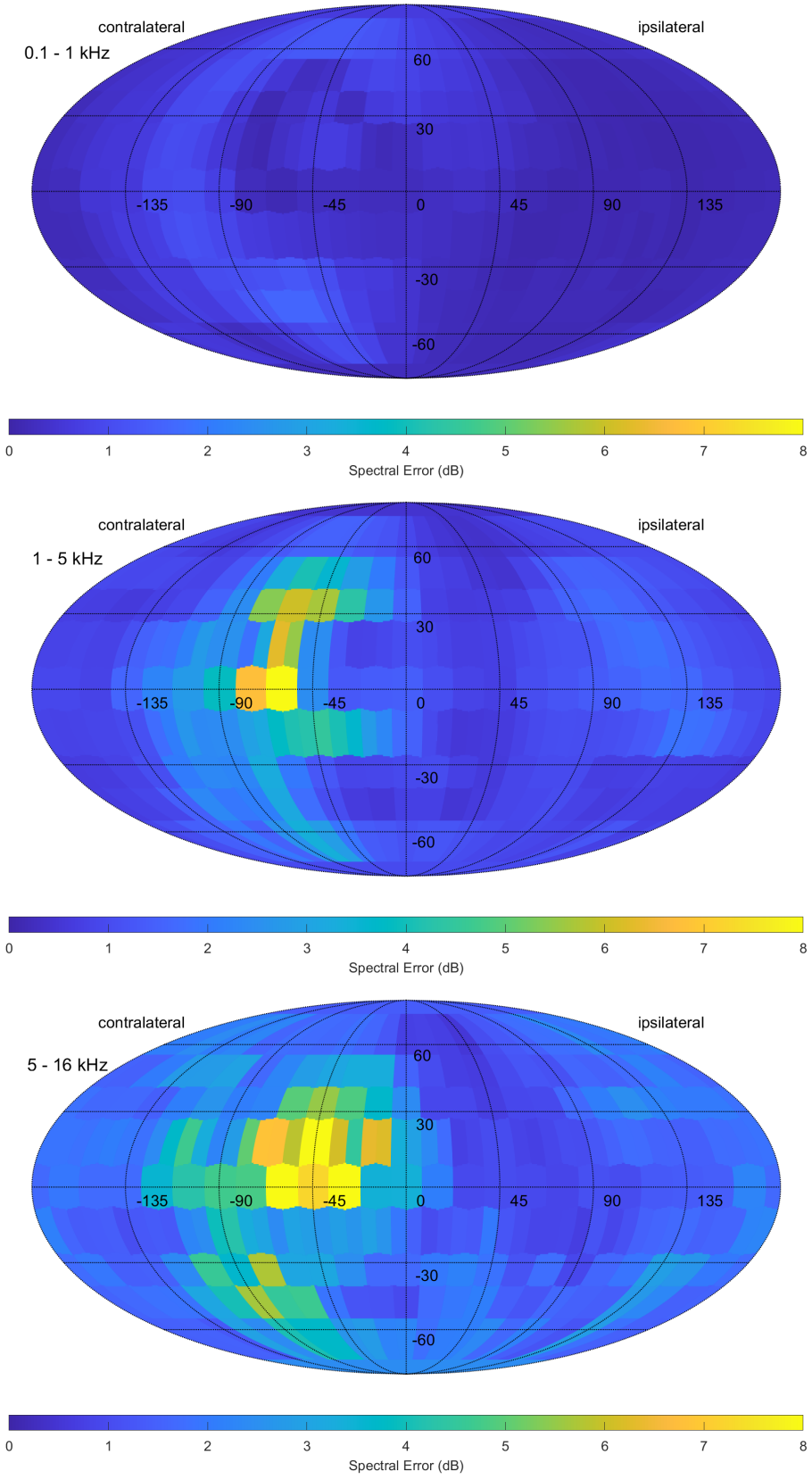


Figure 9.15: Spectral difference between KEMAR HRTFs equipped with Quest 2 headset and without a headset analyzed in three frequency bands.

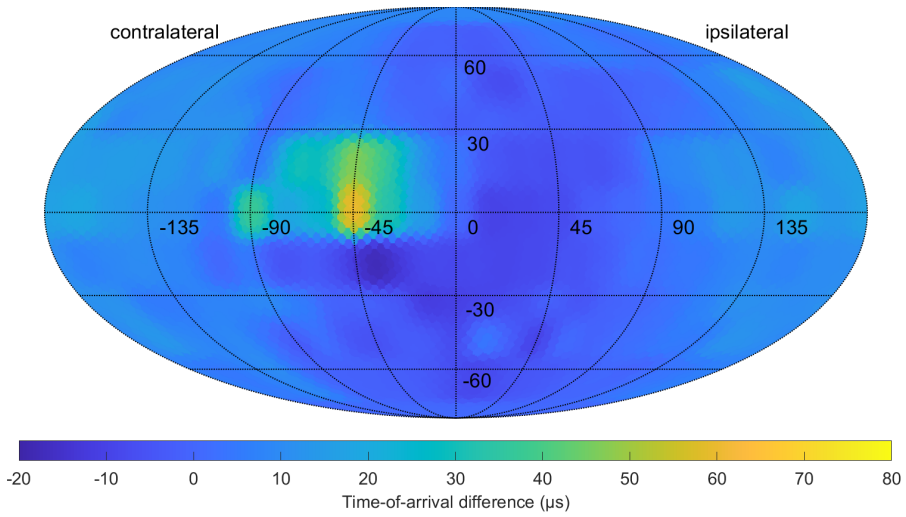


Figure 9.16: Interpolated difference between ear signal arrival time for KEMAR equipped with Quest 2 headset and without a headset.

9.3 Discussion

It remains unclear if the local HRTF perturbations caused by the Quest 2 headset lead to significant perceptual degradation of measured binaural filters. The proposed correction method is yet to be validated using measurements of real subjects and through subjective tests. Nevertheless, one may suspect that further generations of XR devices will exhibit smaller dimensions, therefore the problem will become less relevant for the next iterations of analogous measurement systems.

It is important to note that in the proposed measurement system the head-above-torso orientation (HATO) varies across the measurement session. While the subject is instructed to keep their torso oriented according to the visual cue, there is some degree of pitch and roll rotation of the head required to reach predefined directions. Additionally, the subject needs to slightly bend their torso to reach some points at extreme elevations (see Fig. 9.4). The implications of varying HATO on the measured HRTFs have been researched before (Brinkmann et al., 2014; Reijniers et al., 2020; Bau and Pörschmann, 2022), however, only the recent study by Bau and Pörschmann (2022) was conducted using analogous HATO variations and rather small impact on HRTF magnitudes was observed. This should be further researched in a perceptual experiment comparing individual HRTFs measured with this system against a standard multi-loudspeaker method.

Thanks to the visual component provided by the HMD, the proposed system offers a high level of control over the direction and distance of measured HRTFs. Moreover, the use of XR technology might enable virtual assistance to the user wanting to capture their HRTFs in a home environment. The system could be further scaled down if the XR headset would be able to record signals from the binaural microphones. The sweep signal could be then triggered wireless and played back from a dedicated loudspeaker. This would eliminate the need for an external PC responsible for audio signal capture.

The frequency response of single-driver loudspeakers typically varies with the direction of sound propagation. In order to compensate for off-axis head positions, correction filters could be measured and employed. The current implementation of the system stops the measurement when the head displacement from the reference point exceeds 0.3 m.

The proposed HRTF measurement method can be utilized easily to measure sparse HRTFs. Any layout can be measured, although the required measurement time will increase linearly with the number of measurement points. The time between measurements could be utilized to measure additional data points using adaptive filtering techniques (He et al., 2018) while the subject is rotating their head between the predefined orientations.

The postprocessing of the captured HRIRs could be easily extended by interpolation using an autoencoder with source position conditioning, as proposed by Ito et al. (2022), who have open-sourced their implementation code.

9.4 Summary

This chapter proposes an XR-based system for HRTF measurements using a minimal set of equipment. The virtual guiding interface displayed by the XR headset allows for measuring HRTFs at any predefined sound incidence direction. The procedure does not require an anechoic room thanks to the described post-processing of the measured data. A direction-dependent filter for HRTF perturbations caused by the XR headset has been proposed and will be evaluated in the future.

The described software is open source and can be obtained at: `https://trsonic.github.io/XR-HRTFs/`.

Chapter 10

Conclusions

A summary of the work presented in this thesis is as follows.

Chapter 2 discusses the physics of sound propagation and the human auditory system. Next, a review of the literature on spatial hearing is presented, which introduces the psychoacoustic concepts referred to in this dissertation.

Chapter 3 discusses sound reproduction methods for spatial audio, established formats and perceptual coding schemes, followed by a review of the literature on perceptual evaluation of spatial audio systems.

Chapter 4 presents a study focused on the evaluation of perceived timbral quality degradation introduced by the Opus audio codec at different bitrate settings and Ambisonic orders. Evaluations were conducted using three different reproduction methods: multichannel loudspeaker array, binaural using generic HRTFs and binaural using individual HRTFs. A strong relationship has been found between the codec bitrate, order truncation, and timbral fidelity. The results suggest that 3rd-order and 5th-order Ambisonics provide higher quality over 1st-order, and the 3rd-order streaming could be implemented using a relatively low total bitrate.

Chapter 5 presents a study focused on auditory localisation performance within the presented bitrate-compressed Ambisonic scenes. The impact of the employed reproduction method on the collected responses was also investigated, as the scenes were reproduced over loudspeakers and binaurally using generic and individually measured HRTF sets. The results show that auditory localisation in low-bitrate compressed Ambisonic scenes is not significantly affected by codec parameters. The key factors influencing localisation are the rendering method and Ambisonic order truncation. This suggests that efficient perceptual coding might be used successfully for spatial audio delivery. However, using higher-order Ambisonic content instead of 1st-order Ambisonics will improve localisation within the scenes, especially when using personalised binaural rendering or multi-loudspeaker reproduction.

Chapter 6 extends the evaluated set of codec parameters by testing different chan-

nel mappings and various audio stimuli contexts using a purposely developed VR listening test framework. Key findings were that in all cases, Ambisonic scenes compressed with Opus at 64 kbps/ch using Channel Mapping Family 3 garnered a median BAQ rating not significantly different than uncompressed audio. Channel Mapping Family 3 demonstrated the least variation in BAQ across evaluated contexts.

Chapter 7 focuses on the implementation of established as well as alternative methods for the binaural rendering of Ambisonics. The chapter also presents subsequent objective and subjective evaluations of these. Both objective and subjective evaluations clearly indicated that the MagLS method provides the most perceptually accurate reconstruction of the reference HRTFs at the 3rd and 5th Ambisonic orders. The results reinforce the findings of the experiments in Chapter 4 and Chapter 5, i.e., Ambisonics should be delivered using at least 3rd-order signals for improved perceived quality.

Chapter 8 presents an experiment exploring user preferences of direct-to-reverberant sound ratio (DRR) of virtual Ambisonic listening spaces in relation to different types of reverberation and different Ambisonic audio content. The results show that such a hybrid approach might provide an alternative to the established anechoic rendering. The reverberant filters can be obtained using available simulation software and the described workflow, while anechoic filters can be calculated using the state-of-the-art MagLS method evaluated in the previous chapter.

Chapter 9 discusses a Head Related Transfer Function (HRTF) measurement system that uses minimal hardware configuration.

10.1 Restatement of Hypothesis

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

Streaming and rendering of Ambisonics can be improved through perceptual evaluation and optimisation of the Ambisonic delivery chain.

The conducted perceptual evaluations identified key areas where the Ambisonic delivery chain could be improved to provide a more satisfactory user experience. Based on the results of the experiments, this thesis formulates a set of specific recommendations which could be implemented to optimise the delivery of Ambisonic audio in the existing platforms, e.g. YouTube. That is the use of minimum 3rd-order Ambisonics compressed using Opus with Channel Mapping Family 3 and rendered binaurally using MagLS filters.

10.2 Closing Remarks

The research presented in this thesis investigated the perceived quality of low-bitrate compression and binaural rendering of Ambisonics. The evaluated systems exhibited various degrees of spatial audio quality degradation. Based on the results, it is clear that the established Ambisonic delivery methods could be improved by introducing higher-order streaming, state-of-the-art binaural filters and rendering using virtual listening spaces.

It is hoped that the results of this work will inform the development of the next generations of media streaming platforms and will contribute to the popularisation of the Ambisonic audio technique in a broader context.

Appendix A

A DAW-based Interactive Tool for Perceptual Spatial Audio Evaluation

To provide a comprehensive test environment for spatial audio listening tests, the author has created a dedicated tool (Rudzki et al., 2018) utilising the spatial audio standard DAW software – Reaper[1] as a digital "tape-machine" controlled by a custom listening test application. The current version of the listening test tool has two types of tests implemented for perceptual evaluation of spatial audio codecs: timbral quality test based on MUSHRA recommendation and localisation accuracy test. Using a DAW as an audio playback engine allows easy implementation of the required signal routing and spatial audio processing software. Fully networked communication enables various test interfaces, including mobile touchscreens and physical controllers. The test controller software can trigger the playback of desired audio material and gather users' responses. The Reaper DAW's audio routing flexibility allows each test sample to be reproduced using the relevant spatial audio transform plugins, followed by headphone or speaker calibration and equalisation plugins. A head-tracking and low-latency HRIR convolution can be applied for the headphone-based tests. The resulting listening test data is saved into a standardised text file which can be easily imported into the statistical analysis software.

A software tool for subjective audio evaluation is presented. The tool helps to overcome the limits of the existing listening test tools by allowing DAW-based multichannel playback with required signal processing and enabling the use of novel test participant interfaces: mobile app, physical controller and VR interface. Test preparation is done by importing audio samples into the spatial audio standard DAW and setting up the required signal-processing plugins. The listening test tool triggers the playback of the desired audio samples inside the DAW, according to the participant's choice. The tool described in this paper can be used for various perceptual audio tests, including the evaluation of spatial audio codecs, virtual acoustics and binaural rendering engines.

[1] `https://www.reaper.fm/`

A.1 Introduction

One of the challenges spatial audio researchers face during the preparation of perceptual evaluation experiments is the requirement to play a large number of audio channels simultaneously alongside additional processing in the signal chain. The range of listening test tools currently available to the audio community allowing multichannel playback is very small. Current solutions are Web-based Schoeffler et al. (2018) or created using graphical audio programming environments like Max[2] Gribben and Lee (2015). Other tools include test interfaces created using Matlab Vazquez (2015); Ciba et al. (2009) which allow network-controlled audio playback performed by external applications.

Most of the available tools allow for conducting listening tests according to the ITU-R BS.1534 (MUSHRA) ITU-R (2015b) recommendation, where the assessor is asked to rate specific audio attributes on a Continuous Quality Scale in a relation to the reference audio sample. Alongside these tests, spatial audio researchers use other perceptual methods, like the Method of Adjustment for the evaluation of localization performance (Thresh et al., 2017). There is no available software supporting the latter test.

Another challenge is providing an optimal physical test interface to the participants, as traditional desktop and laptop computers can introduce acoustic shadowing and reflections while using loudspeaker playback systems. The use of compact and wireless interfaces like tablets, can minimize the influence of the controller and make the test more convenient for the assessor.

To provide a comprehensive test environment for spatial audio listening tests, the author created a tool utilizing the spatial audio standard DAW software – Reaper[3] as a digital "tape-machine" controlled by a custom listening test application. The proposed solution allows for straightforward listening test preparation by creating the DAW session with the test material. Using a DAW as an audio playback engine allows easy implementation of the required signal routing and spatial audio processing software. Fully networked communication enables the use of various test interfaces, including mobile touchscreens and physical participant controllers.

A.2 Proposed tool

The test controller software can be used to trigger the playback of desired audio material and to gather users' responses. The audio routing flexibility of the Reaper DAW allows for each test sample to be reproduced using the relevant spatial audio transform plugins, followed by headphone or speaker calibration and equalization plugins. For the headphone-based tests, a head-tracking and low-latency HRIR convolution can be applied. The resulting listening test data is saved into a standardized text file which can be easily imported into a statistical analysis software.

[2] `cycling74.com/products/max/`
[3] `www.reaper.fm/`

Figure A.1: Test material inside Reaper DAW software.

A.2.1 DAW configuration

In order to prepare the DAW session for the experiment, the user has to import audio samples containing the test stimuli into Reaper project according to the following scheme:

- Each sample variation of the assessed audio material (corresponding to the level of impairment) should be placed on a separate track. To indicate tracks containing the test material, their respective names should start with "##" symbols. All audio samples should begin at the same sequencer position for a corresponding trial.
- Audio samples belonging to different trials should be located at different sequencer positions, avoiding overlapping and preserving the track assignment corresponding to their level of impairment. To indicate the beginnings and endings of the test stimuli for each trial, single markers should be placed at the respective positions.

Figure A.1 shows a DAW session ready for conducting listening tests. At the first use of the tool, the user is required to configure Reaper's Open Sound Control[4] settings by adding a generic OSC control surface in the Preferences window.

A.2.2 Main application

The main application window is designed as a researcher's panel. The operation of the application starts by examining the DAW session through the OSC message exchange

[4] `opensoundcontrol.org/`

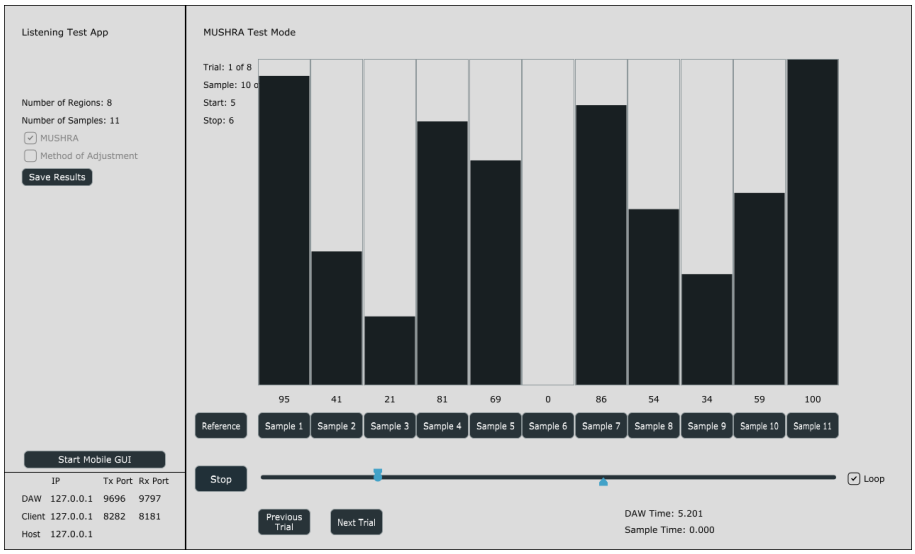


Figure A.2: Main application window with MUSHRA test loaded.

and establishing the number of samples and trials in the experiment. Subsequently, the user is asked to choose the required test interface. The current version of the listening test tool has two types of tests implemented for the purpose of perceptual evaluation of spatial audio codecs: the timbral quality test based on MUSHRA recommendation and the localization accuracy test using the Method of Adjustment. Figure A.2 shows the main application window with the MUSHRA-type test interface loaded.

At the bottom of the screen, the user can find an assessor's mobile application initialization button as well as networking preferences. When using the tool over the network, the IP addresses and port settings should be altered to match the user's hardware configuration before the initialization of the test. After the completion of the test by the participant the results can be saved into a formatted text file.

A.2.3 Remote interfaces

For both implemented tests, mobile-based touch-responsive interfaces have been developed. The mobile interface application can be run on iOS and Android devices. Additionally, for the localization test, a physical controller with azimuth and elevation encoders can be utilized. Both interfaces can be easily operated inside a controlled listening environment i.e. multichannel spherical loudspeaker array or anechoic chamber.

The third type of participant interface has been created as a graphical virtual controller which can be added to the existing VR experiences built using the Unity game engine. This configuration allows for a controlled spatial audio material playback and assessment in virtual environments, overcoming game engine-specific audio rendering software limitations, like Unity's maximum audio file channel count. This makes it possible to easily conduct listening tests in VR with Higher Order Ambisonic material.

A.3 Implementation

The main test application, as well as the mobile interface application, has been written in C++ and utilizes the versatile audio programming framework JUCE[5]. Communication between DAW, listening test application and chosen participant interface is done by OSC protocol. The physical controller has been built using an Arduino[6] microcontroller board which communicates with the main application via serial connection.

A.4 Summary

The proposed system has been tested with both loudspeaker and headphone-based listening experiments. Performed tests included evaluation of Ambisonic recordings coded with Opus audio codec and preference evaluation of virtual acoustic auralizations.

Utilizing a DAW-based listening test application simplifies the process of the listening test preparation significantly. It doesn't require any coding skills from the user. Test stimuli and signal processing management are straightforward to anyone who can operate the DAW Reaper.

[5] `github.com/WeAreROLI/JUCE/`
[6] `www.arduino.cc/`

Appendix B

On the Design of the SALTE Audio Rendering Engine for Spatial Audio Listening Tests in VR

Conducting listening tests with Ambisonic stimuli requires controlled playback of multichannel audio files alongside additional processing in the signal chain. This can be achieved by using the proposed audio rendering engine of the Spatial Audio Listening Environment (SALTE) framework. SALTE is a collaborative and open-source software project developed by members of the AudioLab at the University of York. It is a continuation of the DAW-based listening test software project described in Appendix A. The SALTE framework relies on its own rendering software in opposition to the DAW-based listening test software. The SALTE audio rendering engine differs from existing spatial audio renderers (e.g. Spat Jot and Warusfel (1995), SoundScape Renderer Geier et al. (2008), 3DTI Toolkit Cuevas-Rodríguez et al. (2019)) in that it offers a built-in listening test functionality. It is designed to work with a proprietary VR-based participant interface Johnston et al. (2019b), as well as standard desktop or mobile-based test interfaces.

SALTE can be used for the perceptual evaluation of Ambisonic scenes and different elements of spatial audio rendering systems. The current version of the renderer supports binaural reproduction allowing for the evaluation of individually measured or simulated HRTF sets. The proposed software allows for the investigation of the following experimental variables:

- Ambisonic scenes (e.g. virtual room acoustics, soundscapes, quality degradation introduced by low-bitrate coding),
- Ambisonic decoding algorithms (e.g. decoding matrices and dual-band processing),
- HRIR sets loaded as SOFA Majdak et al. (2013) files or individual WAV files,
- Headphone frequency response compensation.

B.1 The Rendering Engine

The rendering engine is a standalone application. It can be compiled on both Windows and MacOS operating systems. The deployment on Linux and mobile platforms has not been tested at the time of writing. The application has been programmed in C++ using the JUCE[1] audio programming framework. SALTE renderer has a multi-class structure (see Figure B.1). The following libraries have been used as dependencies:

- WDL FFT library[2],
- libSOFA[3],
- Steinberg ASIO SDK[4],
- SADIE II Database[5].

Ambisonic rotation DSP code is based on the SceneRotator VST plugin[6]. For optimal performance on Windows machines, it is recommended to use ASIO driver-enabled audio interfaces. The audio rendering engine can be controlled using a standardized Open Sound Control communication, which allows for head-tracked binaural playback, as well as full control of the application settings by remote listening test interfaces. The main window of the application (see Figure B.2) hosts the following components: audio device manager and OSC settings, listening test interface, console output and the audio processing block which is structured as follows: stimulus player, binaural rendering and headphone frequency response compensation.

B.1.1 Stimulus Player

The stimulus player allows for a controlled playback of multichannel WAV files: binaural (2 channels) or Ambisonic up to 7th order (64 channels). The graphical interface of the player hosts playback controls and information associated with the loaded audio file, including the first channel audio waveform. The waveform presentation area of the GUI can be used to manually set the looped playback region. The stimulus player has three slider controls which allow for the controlled rotation of the Ambisonic scene.

B.1.2 Binaural Rendering

If an Ambisonic audio file is loaded into the player, the audio data buffer is passed to the binaural rendering component. In order to accommodate head tracking, Ambisonic scenes are rotated in real-time based on the received OSC data and

[1] https://github.com/WeAreROLI/JUCE/
[2] https://github.com/justinfrankel/WDL/tree/master/WDL/
[3] https://github.com/sofacoustics/API_Cpp/
[4] https://www.steinberg.net/en/company/developers.html
[5] https://www.york.ac.uk/sadie-project/database.html
[6] https://git.iem.at/audioplugins/IEMPluginSuite/tree/master/SceneRotator/

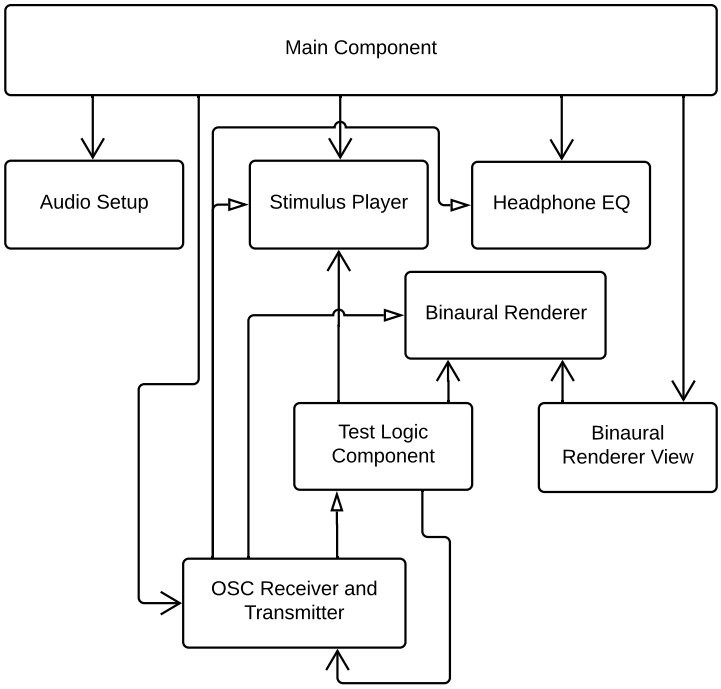


Figure B.1: Unified Modeling Language diagram of the SALTE audio rendering engine.

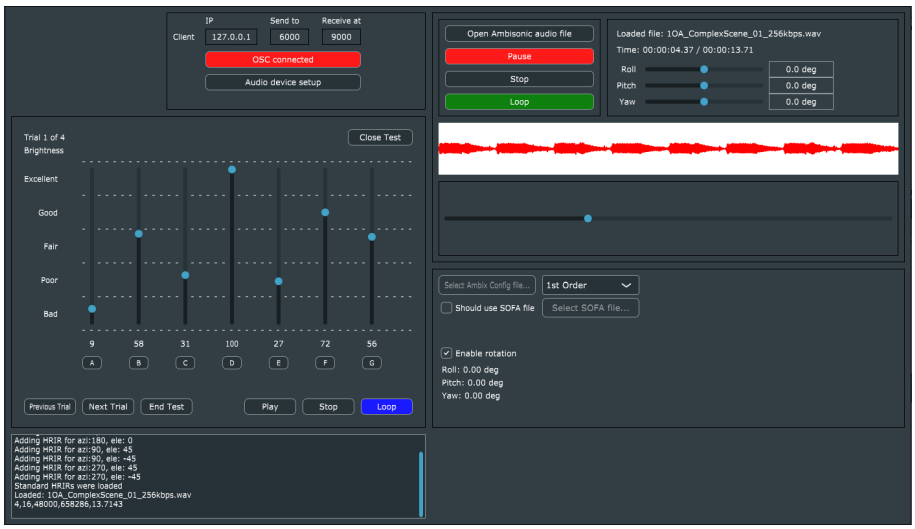


Figure B.2: Graphical User Interface of the SALTE audio rendering engine.

subsequently convolved with HRIRs. Computation of the binaural filters is based on the Ambisonic decoding matrices and SADIE II KU100 HRIRs or custom ones which can be specified by ambiX binaural decoder plugin presets Kronlachner (2014a). The HRIRs can be loaded as individual WAV files specified in ambiX preset or can be read from the selected SOFA file.

Two options for convolution have been coded within the renderer. One option is a standard approach of convolving each virtual loudspeaker feed with its corresponding HRIR in order to obtain the binaural output. The second option is achieved by encoding the HRIR for each virtual loudspeaker position into the spherical harmonic domain, and then convolving the input signal with the spherical harmonic encoded HRIRs Noisternig et al. (2003). The computational savings of the latter approach are significant.

The binaural renderer features dual-band processing capabilities that are required for a more faithful reproduction of the Ambisonic sound scene. Basic decoding matrices are chosen for the lower band for optimized interaural time difference cues and MaxRe decoding matrices are used for the higher band for optimized interaural level difference cues. Crossover frequencies for the different bands are picked depending on the chosen Ambisonic order Moreau et al. (2006).

Convolution has been facilitated by the WDL library set of audio tools, specifically the FFT routines and convolution engine. The implemented convolution engine uses a uniform partitioned convolution, which improves latency over the standard brute force approach.

B.1.3 Headphone Compensation

Headphone frequency response compensation can be performed using two-channel FIR filters or a 31-band graphical equalizer available to the user. Component settings can be loaded manually from the user interface or supplied by the listening test configuration file.

B.1.4 Listening Test Logic

The built-in test components allow for conducting perceptual evaluations using different methods. The direct assessment component is based on the flexible class which allows creation of multiple test trials in accordance with ITU-R BS.1116 ITU-R (2015a) and BS.1534 ITU-R (2015b) recommendations. In these tests, participants are asked to rate a single audio attribute using quality or impairment scales based on the multiple stimulus presentations. Additionally, the direct assessment component allows for conducting multiple stimulus - multiple attribute tests (e.g. the TS26.259 3GPP (2018) test recommended by 3GPP). The choice of methods and experimental variables depends on the user. The direct assessment class can be configured manually or using a JSON file prepared using the VR listening test builder. Examples of different test configuration files are included within the source code repository.

Another test component focuses on auditory localization performance. The proposed methods include the method of adjustment (utilizing a virtual acoustic

pointer) and head pointing.

B.2 Summary

A flexible software tool for conducting listening tests using Ambisonic stimuli has been presented. This tool will significantly aid future research and development of spatial audio systems. The development roadmap of the SALTE framework includes multi-loudspeaker and cross-talk cancellation-based reproduction and the development of indirect evaluation methods utilizing additional sensors, e.g. eye tracking.

The source code of the audio renderer can be obtained from the AudioLab's GitHub page[7].

[7] `https://github.com/AudioLabYork/`

List of Acronyms

SPL	Sound Pressure Level
ITD	Interaural Time Difference
ILD	Interaural Level Difference
HRTF	Head-Related Transfer Function
HRIR	Head-Related Impulse Response
FIR	Finite Impulse Response
RIR	Room Impulse Response
BRIR	Binaural Room Impulse Response
MagLS	Magnitude Least Squares
VL	Virtual Loudspeakers
AllRAD	All Round Ambisonic Panning
MAA	Minimum Audible Angle
FOA	1st-Order Ambisonics
HOA	Higher-Order Ambisonics
VR	Virtual Reality
AR	Augmented Reality
3DOF	Three Degrees of Freedom
6DOF	Six Degrees of Freedom
DAW	Digital Audio Workstation
SH	Spherical Harmonics
SH-HRIRs	SH-domain HRIRs
SH-BRIRs	SH-domain BRIRs
PCM	Pulse-code modulation
ERB	Equivalent rectangular bandwidth
RMS	Root-mean-square
FFT	Fast Fourier transform
MUSHRA	Multiple stimulus test with hidden reference and anchor
RT	Reverberation Time

References

3GPP. Ts 26.259: Subjective test methodologies for the evaluation of immersive audio systems. *3GPP Specifications*, 2018.

S. Adams and F. Boland. On the distortion of binaural localization cues using headphones. In *IET Irish Signals and Systems Conference (ISSC 2010)*, pages 82–87, 2010. doi: 10.1049/cp.2010.0492.

A. Ahrens, K. D. Lund, M. Marschall, and T. Dau. Sound source localization with varying amount of visual information in virtual reality. *PloS one*, 14(3):e0214603, 2019.

V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. The cipic hrtf database. In *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pages 99–102. IEEE, 2001.

J. M. Arend, F. Brinkmann, and C. Pörschmann. Assessing spherical harmonics interpolation of time-aligned head-related transfer functions. *Journal of the Audio Engineering Society*, 69(1/2):104–117, 2021.

C. Armstrong and G. Kearney. Ambisonics understood. In *3D Audio*, pages 99–129. Routledge London and New York, 2021.

C. Armstrong, D. Murphy, and G. Kearney. A bi-radial approach to ambisonics. In *Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2018a.

C. Armstrong, L. Thresh, D. Murphy, and G. Kearney. A perceptual evaluation of individual and non-individual hrtfs: A case study of the sadie ii database. *Applied Sciences*, 8(11):2029, 2018b.

H. Bahu, T. Carpentier, M. Noisternig, and O. Warusfel. Comparison of different egocentric pointing methods for 3d sound localization experiments. *Acta acustica united with Acustica*, 102(1):107–118, 2016.

D. W. Batteau. The role of the pinna in human localization. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 168(1011):158–180, 1967.

D. Bau and C. Pörschmann. Technical evaluation of an easy-to-use head-related transfer function measurement system. In *Proceedings of the 48th DAGA*, pages 1–4, 03 2022.

D. Bau, T. Lübeck, J. M. Arend, D. Dziwis, and C. Pörschmann. Simplifying head-related transfer function measurements: A system for use in regular rooms based on free head movements. In *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pages 1–6. IEEE, 2021.

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

Z. Ben-Hur, F. Brinkmann, J. Sheaffer, S. Weinzierl, and B. Rafaely. Spectral equalization in binaural signals represented by order-truncated spherical harmonics. *The Journal of the Acoustical Society of America*, 141(6):4087–4096, 2017.

Z. Ben-Hur, D. Alon, P. W. Robinson, and R. Mehra. Localization of virtual sounds in dynamic listening using sparse hrtfs. In *Audio Engineering Society Conference: 2020 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2020.

Z. Ben-Hur, D. L. Alon, R. Mehra, and B. Rafaely. Binaural reproduction based on bilateral ambisonics and ear-aligned hrtfs. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:901–913, 2021.

E. Benjamin and T. Chen. The native b-format microphone. In *Audio Engineering Society Convention 119*. Audio Engineering Society, 2005.

E. Benjamin, A. Heller, and R. Lee. Design of ambisonic decoders for irregular arrays of loudspeakers by non-linear optimization. In *Audio Engineering Society Convention 129*. Audio Engineering Society, 2010.

A. J. Berkhout, D. de Vries, and P. Vogel. Acoustic control by wave field synthesis. *The Journal of the Acoustical Society of America*, 93(5):2764–2778, 1993.

B. Bernschütz. A spherical far field hrir/hrtf compilation of the neumann ku 100. In *Proceedings of the 40th Italian (AIA) annual conference on acoustics and the 39th German annual conference on acoustics (DAGA) conference on acoustics*, page 29. AIA/DAGA Merano, 2013.

B. Bernschütz. Adaptation of hrtfs to plane waves with reduced modal order. In *Proceedings of the German DAGA Conference, DEGA*, 2014.

S. Bertet, J. Daniel, E. Parizet, and O. Warusfel. Investigation on localisation accuracy for first and higher order ambisonics reproduced sound sources. *Acta Acustica united with Acustica*, 99(4):642–657, 2013.

L. M. Biga, S. Dawson, A. Harwell, R. Hopkins, J. Kaufmann, M. LeMaster, P. Matern, K. Morrison-Graham, D. Quick, and J. Runyeon. *Anatomy & physiology*. OpenStax/Oregon State University, 2020.

J. Blauert and J. Allen. *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press, 1997. ISBN 9780262024136. URL `https://books.google.co.uk/books?id=wBiEKPhw7r0C`.

K. Brandenburg and R. Henke. Near-lossless coding of high quality digital audio: first results. In *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 193–196. IEEE, 1993.

S. Braun and M. Frank. Localization of 3d ambisonic recordings and ambisonic virtual sources. In *1st International Conference on Spatial Audio,(Detmold)*, 2011.

J. Brettle and J. Skoglund. Open-source spatial audio compression for vr content. In *SMPTE 2016 Annual Technical Conference and Exhibition*, pages 1–9, Oct. 2016. doi: 10.5594/M001712.

W. O. Brimijoin, A. W. Boyd, and M. A. Akeroyd. The contribution of head movement to the externalization and internalization of sounds. *PloS one*, 8(12): e83068, 2013.

F. Brinkmann, R. Roden, A. Lindau, and S. Weinzierl. Audibility of head-above-torso orientation in head-related transfer functions. In *Forum Acusticum*, pages 1–6, 2014.

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

A. W. Bronkhorst and T. Houtgast. Auditory distance perception in rooms. *Nature*, 397(6719):517–520, 1999.

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

D. S. Brungart, W. Nelson, R. Bolia, and R. Tannen. Evaluation of the snapshot 3d head-related transfer functions measurement system. Technical report, United States Air Force Research Laboratory, 1998.

J. Catic, S. Santurette, J. M. Buchholz, F. Gran, and T. Dau. The effect of interaural-level-difference fluctuations on the externalization of sound. *The Journal of the Acoustical Society of America*, 134(2):1232–1241, 2013.

J. Catic, S. Santurette, and T. Dau. The role of reverberation-related binaural cues in the externalization of speech. *The Journal of the Acoustical Society of America*, 138(2):1154–1167, 2015.

S. Ciba, A. Wlodarski, and H.-J. Maempel. Whisper–a new tool for performing listening tests. In *Audio Engineering Society Convention 126*. Audio Engineering Society, 2009.

J. Cooper, S. Carlile, and D. Alais. Distortions of auditory space during rapid head turns. *Experimental brain research*, 191(2):209–219, 2008.

P. G. Craven. Continuous surround panning for 5-speaker reproduction. In *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality*. Audio Engineering Society, 2003.

P. G. Craven and M. A. Gerzon. Coincident microphone simulation covering three dimensional space and yielding various directional outputs, Aug. 16 1977. US Patent 4,042,779.

R. Crawford-Emery and H. Lee. The subjective effect of brir length on perceived headphone sound externalization and tonal coloration. In *Audio Engineering Society Convention 136*. Audio Engineering Society, 2014.

A. Cuarón. Gravity. Universal Pictures, 2013.

M. Cuevas-Rodriguez, D. L. Alon, S. Clapp, P. W. Robinson, and R. Mehra. *Evaluation of the effect of head-mounted display on individualized head-related transfer functions*. Universitätsbibliothek der RWTH Aachen, 2019.

M. Cuevas-Rodríguez, L. Picinali, D. González-Toledo, C. Garre, E. de la Rubia-Cuestas, L. Molina-Tanco, and A. Reyes-Lecuona. 3d tune-in toolkit: An open-source library for real-time binaural spatialisation. *PloS one*, 14(3):e0211899, 2019.

J. Daniel, J.-B. Rault, and J.-D. Polack. Ambisonics encoding of other audio formats for multiple listening conditions. In *Audio Engineering Society Convention 105*. Audio Engineering Society, 1998.

J. Daniel, S. Moreau, and R. Nicol. Further investigations of high-order ambisonics and wavefield synthesis for holophonic sound imaging. In *Audio Engineering Society Convention 114*. Audio Engineering Society, 2003.

E. Deleflie and S. Goodwin. Interview with simon goodwin of codemasters on the ps3 game dirt and ambisonics, Aug 2007. URL `https://ambisonicbootlegs.wordpress.com/2007/08/30/interview-with-simon-goodwin-of-codemasters-on-the-ps3-game-dirt-and-ambisonics/`.

T. Deppisch, N. Meyer-Kahlen, B. Hofer, T. Latka, and T. Zernicki. Hoast: A higher-order ambisonics streaming platform. In *Audio Engineering Society Convention 148*. Audio Engineering Society, 2020.

T. Deppisch, H. Helmholz, and J. Ahrens. End-to-end magnitude least squares binaural rendering of spherical microphone array signals. In *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pages 1–7. IEEE, 2021.

I. Engel, D. F. Goodman, and L. Picinali. Assessing hrtf preprocessing methods for ambisonics rendering through perceptual models. *Acta Acustica*, 6:4, 2022.

A. Farina. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *Audio engineering society convention 108*. Audio Engineering Society, 2000.

R. F. Fela, A. Pastor, P. Le Callet, N. Zacharov, T. Vigier, and S. Forchhammer. Perceptual evaluation on audio-visual dataset of 360 content. In *2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2022.

R. Fino, M. J. Lin, A. Caballero, and F. F. Balahadia. Disaster awareness simulation for children with autism spectrum disorder using android virtual reality. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, 9(2-6):59–62, 2017.

N. I. Fisher, T. Lewis, and B. J. Embleton. *Statistical analysis of spherical data*. Cambridge university press, 1993.

M. B. Gardner and R. S. Gardner. Problem of localization in the median plane: effect of pinnae cavity occlusion. *The Journal of the Acoustical Society of America*, 53(2):400–408, 1973.

M. Geier, J. Ahrens, and S. Spors. The soundscape renderer: A unified spatial audio reproduction framework for arbitrary rendering methods. In *In 124 th AES Conv.* Citeseer, 2008.

M. Geier, J. Ahrens, and S. Spors. Object-based audio reproduction and the audio scene description format. *Organised Sound*, 15(3):219–227, 2010.

A. Genovese, G. Zalles, G. Reardon, and A. Roginska. Acoustic perturbations in hrtfs measured on mixed reality headsets. In *Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2018.

M. A. Gerzon. Periphony: With-height sound reproduction. *J. Audio Eng. Soc*, 21 (1):2–10, 1973. URL `http://www.aes.org/e-lib/browse.cfm?elib=2012`.

M. A. Gerzon. Practical periphony: The reproduction of full-sphere sound. In *Audio Engineering Society Convention 65*. Audio Engineering Society, 1980.

M. A. Gerzon. General metatheory of auditory localisation. In *Audio Engineering Society Convention 92*. Audio Engineering Society, 1992.

T. G. Ghirardelli and A. A. Scharine. Auditory-visual interactions. *Helmet-mounted displays: Sensation, perception and cognition issues*, pages 599–618, 2009.

R. H. Gilkey, M. D. Good, M. A. Ericson, J. Brinkman, and J. M. Stewart. A pointing technique for rapidly collecting localization responses in auditory research. *Behavior Research Methods, Instruments, & Computers*, 27(1):1–11, 1995.

P. M. Giller, F. Wendt, and R. Höldrich. The influence of different brir modification techniques on externalization and sound quality. In *EAA Spatial Audio Signal Processing Symposium*, pages 61–66, 2019.

B. R. Glasberg and B. C. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing research*, 47(1-2):103–138, 1990.

M. Gorzel, A. Allen, I. Kelly, J. Kammerl, A. Gungormusler, H. Yeh, and F. Boland. Efficient encoding and decoding of binaural sound with resonance audio. In *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society, 2019.

D. W. Grantham, B. W. Hornsby, and E. A. Erpenbeck. Auditory spatial resolution in horizontal, vertical, and diagonal planes. *The Journal of the Acoustical Society of America*, 114(2):1009–1022, 2003.

M. C. Green and D. Murphy. Eigenscape: A database of spatial acoustic scene recordings. *Applied Sciences*, 7(11):1204, 2017.

C. Gribben and H. Lee. Toward the development of a universal listening test interface generator in max. In *Audio Engineering Society Convention 138*. Audio Engineering Society, 2015.

R. Gupta, R. Ranjan, J. He, and G. Woon-Seng. Investigation of effect of vr/ar headgear on head related transfer functions for natural listening. In *Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2018.

E. H. Langendijk. Collecting localization response with a virtual acoustic pointer. *Journal of The Acoustical Society of America - J ACOUST SOC AMER*, 101, 05 1997. doi: 10.1121/1.418867.

W. M. Hartmann and A. Wittenberg. On the externalization of sound images. *The Journal of the Acoustical Society of America*, 99(6):3678–3688, 1996.

H. G. Hassager, F. Gran, and T. Dau. The role of spectral detail in the binaural transfer function on perceived externalization in a reverberant environment. *The Journal of the Acoustical Society of America*, 139(5):2992–3000, 2016.

J. He, R. Ranjan, W.-S. Gan, N. K. Chaudhary, N. D. Hai, and R. Gupta. Fast continuous measurement of hrtfs with unconstrained head movements for 3d audio. *Journal of the Audio Engineering Society*, 66(11):884–900, 2018.

H. L. Helmholtz. *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. Cambridge University Press, 2009.

G. B. Henning. Detectability of interaural delay in high-frequency complex waveforms. *The Journal of the Acoustical Society of America*, 55(1):84–90, 1974.

J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties. Mpeg-h audio—the new standard for universal spatial/3d audio coding. *Journal of the Audio Engineering Society*, 62(12):821–830, 2015.

C. Hold, H. Gamper, V. Pulkki, N. Raghuvanshi, and I. J. Tashev. Improving binaural ambisonics decoding by spherical harmonics domain tapering and coloration compensation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 261–265. IEEE, 2019.

D. Howard and J. Angus. *Acoustics and psychoacoustics*. Routledge, 2013.

R. Hucknall. Contextual use of reverberation in the externalisation of higher-order ambisonic binaural audio. Beng report, University of York, 2023.

Y. Ito, T. Nakamura, S. Koyama, and H. Saruwatari. Head-related transfer function interpolation from spatially sparse measurements using autoencoder with source position conditioning. In *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 1–5. IEEE, 2022.

ITU-R. Bs.1770-3: Algorithms to measure audio programme loudness and true-peak audio level. *ITU-R Recommendations and Reports*, 2012.

ITU-R. Bs.1116-3: Methods for the subjective assessment of small impairments in audio systems. *ITU-R Recommendations and Reports*, 2015a.

ITU-R. Bs.1534-3: Method for the subjective assessment of intermediate quality levels of coding systems. *ITU-R Recommendations and Reports*, 2015b.

ITU-R. Bs.2399-0: Methods for selecting and describing attributes and terms, in the preparation of subjective tests. *ITU-R Recommendations and Reports*, 2017.

ITU-R. Bs.2076-2: Audio definition model. *ITU-R Recommendations and Reports*, 2019.

ITU-R. Bs.2051-3: Advanced sound system for programme production. *ITU-R Recommendations and Reports*, 2022a.

ITU-R. Bs.775-4: Multichannel stereophonic sound system with and without accompanying picture. *ITU-R Recommendations and Reports*, 2022b.

L. A. Jeffress and R. W. Taylor. Lateralization vs localization. *The Journal of the Acoustical Society of America*, 33(4):482–483, 1961.

S. G. Johnson. The nlopt nonlinear-optimization package, 2014.

D. Johnston, H. Egermann, and G. Kearney. Measuring the behavioral response to spatial audio within a multi-modal virtual reality environment in children with autism spectrum disorder. *Applied Sciences*, 9(15):3152, 2019a.

D. Johnston, B. Tsui, and G. Kearney. Salte pt. 1: A virtual reality tool for streamlined and standardized spatial audio listening tests. In *Audio Engineering Society Convention 147*. Audio Engineering Society, 2019b.

N. Jones. Chasing pirates. EMI, 2009.

J.-M. Jot and O. Warusfel. Spat: A spatial processor for musicians and sound engineers. In *CIARM: International Conference on Acoustics and Musical Research*, 1995.

J.-M. Jot, S. Wardle, and V. Larcher. Approaches to binaural synthesis. In *Audio Engineering Society Convention 105*, Sept. 1998. URL `http://www.aes.org/e-lib/browse.cfm?elib=8319`.

K. A. Karl, J. V. Peluchette, and N. Aghakhani. Virtual work meetings during the covid-19 pandemic: The good, bad, and ugly. *Small Group Research*, 53(3): 343–365, 2022.

G. Kearney. *Auditory scene synthesis using virtual acoustic recording and reproduction*. PhD thesis, Trinity College Dublin, 2010.

G. Kearney and T. Doyle. An hrtf database for virtual loudspeaker rendering. In *Audio Engineering Society Convention 139*. Audio Engineering Society, 2015.

O. Kirkeby and P. A. Nelson. Digital filter design for inversion problems in sound reproduction. *Journal of the Audio Engineering Society*, 47(7/8):583–595, 1999.

J. Klug and M. Dietz. Frequency dependence of sensitivity to interaural phase differences in pure tones. *The Journal of the Acoustical Society of America*, 152 (6):3130–3141, 2022.

M. Kronlachner. Plug-in suite for mastering the production and playback in surround sound and ambisonics. *Gold-Awarded Contribution to AES Student Design Competition*, 2014a.

M. Kronlachner. Spatial transformations for the alteration of ambisonic recordings. *M. Thesis, University of Music and Performing Arts, Graz, Institute of Electronic Music and Acoustics*, 2014b.

W. H. Kruskal and W. A. Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

G. F. Kuhn. Model for the interaural time differences in the azimuthal plane. *The Journal of the Acoustical Society of America*, 62(1):157–167, 1977. doi: 10.1121/1.381498. URL `https://doi.org/10.1121/1.381498`.

A. Kulkarni and H. S. Colburn. Role of spectral detail in sound-source localization. *Nature*, 396(6713):747–749, 1998.

E. H. Langendijk, D. J. Kistler, and F. L. Wightman. Sound localization in the presence of one or two distracters. *The Journal of the Acoustical Society of America*, 109(5):2123–2134, 2001.

V. Larcher, J.-M. Jot, and G. Vandernoot. Equalization methods in binaural technology. In *Audio Engineering Society Convention 105*. Audio Engineering Society, 1998.

V. I. Lebedev and D. Laikov. A quadrature formula for the sphere of the 131st algebraic order of accuracy. *Doklady Mathematics*, 59(3):477–481, 1999.

P. Lecomte, P.-A. Gauthier, C. Langrenne, A. Berry, and A. Garcia. A fifty-node lebedev grid and its applications to ambisonics. *J. Audio Eng. Soc*, 64(11):868–881, 2016. URL `http://www.aes.org/e-lib/browse.cfm?elib=18524`.

B. Lee. Test Stimuli for Context Based Evaluation of the OPUS Audio Codec, July 2022. URL `https://doi.org/10.5281/zenodo.6906836`.

B. Lee, T. Rudzki, J. Skoglund, and G. Kearney. Context-Based Evaluation of the Opus Audio Codec for Spatial Audio Content in Virtual Reality. *Journal of the Audio Engineering Society*, 71(4):145–154, 2023.

H. Lee and D. Johnson. An open-access database of 3d microphone array recordings. In *Audio Engineering Society Convention 147*. Audio Engineering Society, 2019.

H. Lee, M. Frank, and F. Zotter. Spatial and timbral fidelities of binaural ambisonics decoders for main microphone array recordings. In *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society, 2019.

J. Levison. Alive. DTS Demo Disc, 2006.

S. Li and J. Peissig. Fast estimation of 2d individual hrtfs with arbitrary head movements. In *2017 22nd International Conference on Digital Signal Processing (DSP)*, pages 1–5. IEEE, 2017.

S. Li and J. Peissig. Measurement of head-related transfer functions: A review. *Applied Sciences*, 10(14):5014, 2020.

S. Li, A. Tobbala, and J. Peissig. Towards mobile 3d hrtf measurement. In *Audio Engineering Society Convention 148*. Audio Engineering Society, 2020.

A. Lindau, T. Hohn, and S. Weinzierl. Binaural resynthesis for comparative studies of acoustical environments. In *Audio Engineering Society Convention 122*. Audio Engineering Society, 2007.

A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkmann, and S. Weinzierl. A spatial audio quality inventory for virtual acoustic environments (saqi). *Acta Acustica united with Acustica*, 100(5):984–994, 2014.

T. Lübeck, H. Helmholz, J. M. Arend, C. Pörschmann, and J. Ahrens. Perceptual evaluation of mitigation approaches of impairments due to spatial undersampling in binaural rendering of spherical microphone array data. *Journal of the Audio Engineering Society*, 68(6):428–440, 2020.

P. Majdak, M. J. Goupell, and B. Laback. 3-d localization of virtual sound sources: effects of visual environment, pointing method, and training. *Attention, perception, & psychophysics*, 72(2):454–469, 2010.

P. Majdak, Y. Iwaya, T. Carpentier, R. Nicol, M. Parmentier, A. Roginska, Y. Suzuki, K. Watanabe, H. Wierstorf, H. Ziegelwanger, et al. Spatially oriented format for acoustics: A data exchange format representing head-related transfer functions. In *Audio Engineering Society Convention 134*. Audio Engineering Society, 2013.

P. Majdak, C. Hollomey, and R. Baumgartner. Amt 1. x: A toolbox for reproducible research in auditory modeling. *Acta Acustica*, 6:19, 2022.

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

P. Marins, F. Rumsey, and S. Zielinski. Unravelling the relationship between basic audio quality and fidelity attributes in low bit-rate multi-channel audio codecs. In *Audio Engineering Society Convention 124*. Audio Engineering Society, 2008.

M. L. Max and J. C. Burke. Virtual reality for autism communication and education, with lessons for medical training simulators. In *Medicine Meets Virtual Reality*, pages 46–53. IOS Press, 1997.

R. McGill, J. W. Tukey, and W. A. Larsen. Variations of box plots. *The American Statistician*, 32(1):12–16, 1978.

A. McKeag and D. S. McGrath. Sound field format to binaural decoder with head tracking. In *Audio engineering society convention 6r*. Audio Engineering Society, 1996.

T. McKenzie, D. Murphy, and G. Kearney. Diffuse-field equalisation of binaural ambisonic rendering. *Applied Sciences*, 8(10):1956, 2018.

T. McKenzie, C. Armstrong, L. Ward, D. T. Murphy, and G. Kearney. Predicting the colouration between binaural signals. *Applied Sciences*, 12(5):2441, 2022.

J. C. Middlebrooks. Individual differences in external-ear transfer functions reduced by scaling in frequency. *The Journal of the Acoustical Society of America*, 106(3): 1480–1492, 1999.

A. W. Mills. On the minimum audible angle. *The Journal of the Acoustical Society of America*, 30(4):237–246, 1958.

H. Moller. Cancellation of crosstalk in artificial head recordings, reproduced through loudspeakers. In *Audio Engineering Society Convention 84*. Audio Engineering Society, 1988.

H. Møller. Fundamentals of binaural technology. *Applied acoustics*, 36(3-4):171–218, 1992.

H. Møller, C. B. Jensen, D. Hammershøi, and M. F. Sørensen. Design criteria for headphones. *Journal of the Audio Engineering Society*, 43(4):218–232, 1995.

S. Moreau, J. Daniel, and S. Bertet. 3d sound field recording with higher order ambisonics–objective measurements and validation of a 4th order spherical microphone. In *120th Convention of the AES*, pages 20–23, 2006.

C. Nachbar, F. Zotter, E. Deleflie, and A. Sontacchi. Ambix-a suggested ambisonics format. In *Ambisonics Symposium*, volume 2011, 2011.

J. Nam, J. S. Abel, and J. O. Smith III. A method for estimating interaural time difference for binaural synthesis. In *Audio Engineering Society Convention 125*. Audio Engineering Society, 2008.

M. Narbutt, S. O'Leary, A. Allen, J. Skoglund, and A. Hines. Streaming vr for immersion: Quality aspects of compressed spatial audio. In *2017 23rd International Conference on Virtual System & Multimedia (VSMM)*, pages 1–6. IEEE, 2017.

M. Narbutt, A. Allen, J. Skoglund, M. Chinen, and A. Hines. Ambiqual-a full reference objective quality metric for ambisonic spatial audio. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2018.

M. Noisternig, A. Sontacchi, T. Musil, and R. Holdrich. A 3d ambisonic based binaural sound reproduction system. In *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality*, June 2003. URL `http://www.aes.org/e-lib/browse.cfm?elib=12314`.

T. H. Pedersen and N. Zacharov. The development of a sound wheel for reproduced sound. In *Audio Engineering Society Convention 138*. Audio Engineering Society, 2015.

S. Peksi, N. D. Hai, R. Ranjan, R. Gupta, J. He, and W. S. Gan. A unity based platform for individualized hrtf research and development: From on-the-fly fast acquisition to spatial audio renderer. In *Audio Engineering Society Conference: 2019 AES International Conference on Headphone Technology*. Audio Engineering Society, 2019.

D. R. Perrott and K. Saberi. Minimum audible angle thresholds for sources varying in both elevation and azimuth. *The Journal of the Acoustical Society of America*, 87(4):1728–1731, 1990.

A. Politis, S. Tervo, and V. Pulkki. Compass: Coding and multidirectional parameterization of ambisonic sound scenes. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6802–6806. IEEE, 2018.

C. Pörschmann, J. M. Arend, and R. Gillioz. How wearing headgear affects measured head-related transfer functions. In *EAA Spatial Audio Signal Processing Symposium*, pages 49–54, 2019.

P. Power, W. Davies, J. Hirst, C. Dunn, et al. Localisation of elevated virtual sources in higher order ambisonic sound fields. *Proceedings of the Institute of Acoustics*, 2012.

D. Pralong and S. Carlile. The role of individualized headphone calibration for the generation of high fidelity virtual auditory space. *The Journal of the Acoustical Society of America*, 100(6):3785–3793, 1996.

V. Pulkki. Virtual sound source positioning using vector base amplitude panning. *Journal of the audio engineering society*, 45(6):456–466, 1997.

V. Pulkki, M.-V. Laitinen, and V. Sivonen. Hrtf measurements with a continuously moving loudspeaker and swept sines. In *Audio Engineering Society Convention 128*. Audio Engineering Society, 2010.

V. Pulkki, S. Delikaris-Manias, and A. Politis. *Parametric time-frequency domain spatial audio*. John Wiley & Sons, 2017.

L. Rayleigh. Xii. on our perception of sound direction. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 13(74):214–232, 1907.

J. Reijniers, B. Partoens, J. Steckel, and H. Peremans. Hrtf measurement by means of unsupervised head movements with respect to a single fixed speaker. *IEEE Access*, 8:92287–92300, 2020.

D. Rivas Méndez, C. Armstrong, J. Stubbs, M. Stiles, and G. Kearney. Practical recording techniques for music production with six-degrees of freedom virtual reality. In *Audio Engineering Society Convention 145*. Audio Engineering Society, 2018.

A. Roginska and P. Geluso. *Immersive sound: the art and science of binaural and multi-channel audio*. Taylor & Francis, 2017.

D. Rudrich and M. Frank. Improving externalization in ambisonic binaural decoding. In *Proceedings of the DAGA German Annual Conference on Acoustics*, pages 1466–1469, 2019.

T. Rudzki, D. Murphy, and G. Kearney. A DAW-Based Interactive Tool for Perceptual Spatial Audio Evaluation. In *Audio Engineering Society Convention 145*. Audio Engineering Society, 2018.

T. Rudzki, C. Earnshaw, D. Murphy, and G. Kearney. SALTE Pt. 2: On the Design of the SALTE Audio Rendering Engine for Spatial Audio Listening Tests in VR. In *Audio Engineering Society Convention 147*. Audio Engineering Society, 2019a.

T. Rudzki, I. Gomez-Lanzaco, P. Hening, J. Skoglund, T. McKenzie, J. Stubbs, D. Murphy, and G. Kearney. Perceptual Evaluation of Bitrate Compressed Ambisonic Scenes in Loudspeaker Based Reproduction. In *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society, 2019b.

T. Rudzki, I. Gomez-Lanzaco, J. Stubbs, J. Skoglund, D. T. Murphy, and G. Kearney. Auditory Localization in Low-Bitrate Compressed Ambisonic Scenes. *Applied Sciences*, 9(13):2618, 2019c.

T. Rudzki, D. Murphy, and G. Kearney. XR-based HRTF Measurements. In *Audio Engineering Society Conference: AES 2022 International Audio for Virtual and Augmented Reality Conference*. Audio Engineering Society, 2022.

T. Rudzki, D. Murphy, and G. Kearney. User Preference Evaluation of Direct-to-Reverberant Ratio of Virtual Ambisonic Listening Spaces. In *Audio Engineering Society Conference: 2023 AES International Conference on Spatial and Immersive Audio*. Audio Engineering Society, 2023.

F. Rumsey, S. Zieliński, R. Kassier, and S. Bech. On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality. *The Journal of the Acoustical Society of America*, 118(2):968–976, 2005.

N. H. Salminen. Human cortical sensitivity to interaural level differences in low-and high-frequency sounds. *The Journal of the Acoustical Society of America*, 137(2): EL190–EL193, 2015.

M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre. webmushra–a comprehensive framework for web-based listening tests. *Journal of Open Research Software*, 6(1), 2018.

C. Schörkhuber, M. Zaunschirm, and R. Höldrich. Binaural rendering of ambisonic signals via magnitude least squares. In *Proceedings of the DAGA*, volume 44, pages 339–342, 2018.

H. Shimazaki and S. Shinomoto. Kernel bandwidth optimization in spike rate estimation. *Journal of computational neuroscience*, 29(1-2):171–182, 2010.

B. G. Shinn-Cunningham, S. Santarelli, and N. Kopco. Tori of confusion: Binaural localization cues for sources within reach of a listener. *The Journal of the Acoustical Society of America*, 107(3):1627–1636, 2000.

J. Skoglund and M. Graczyk. Ambisonics in an Ogg Opus Container. RFC 8486, Oct. 2018. URL `https://rfc-editor.org/rfc/rfc8486.txt`.

T. Takeuchi and P. A. Nelson. Optimal source distribution for binaural synthesis over loudspeakers. *The Journal of the Acoustical Society of America*, 112(6):2786–2797, 2002.

S. P. Thompson. Xliii. the pseudophone. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 8(50):385–390, 1879.

L. Thresh, C. Armstrong, and G. Kearney. A direct comparison of localization performance when using first, third, and fifth ambisonics order for real loudspeaker and virtual loudspeaker rendering. In *Audio Engineering Society Convention 143*, 2017.

J.-M. Valin, G. Maxwell, T. B. Terriberry, and K. Vos. High-quality, low-delay music coding in the opus codec. In *Audio Engineering Society Convention 135*, Oct. 2013. URL `http://www.aes.org/e-lib/browse.cfm?elib=16992`.

J. Van Opstal. *The auditory system and human sound-localization behavior*. Academic Press, 2016.

A. Vazquez. Scale-conduction psychacoustic experiments with dynamic binaural synthesis. In *Proc. of the DAGA*, 2015.

T. Verdebout. On some validity-robust tests for the homogeneity of concentrations on spheres. *Journal of Nonparametric Statistics*, 27(3):372–383, 2015.

J. Vilkamo, T. Bäckström, and A. Kuntz. Optimized covariance domain framework for time–frequency processing of spatial audio. *Journal of the Audio Engineering Society*, 61(6):403–411, 2013.

A. Wabnitz, N. Epain, C. Jin, and A. Van Schaik. Room acoustics simulation for multichannel microphone arrays. In *Proceedings of the International Symposium on Room Acoustics*, pages 1–6. Citeseer, 2010.

N. J. Wade and D. Deutsch. Binaural hearing—before and after the stethophone. *Acoustics Today*, 4(3):16–27, 2008.

H. Wallach. On sound localization. *The Journal of the Acoustical Society of America*, 10(4):270–274, 1939.

M. Wang and E. Anagnostou. Virtual reality as treatment tool for children with autism. *Comprehensive guide to autism*, pages 2125–2141, 2014.

R. M. Warren. Measurement of sensory intensity. *Behavioral and Brain Sciences*, 4 (2):175–189, 1981.

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

B. Wiggins. The generation of panning laws for irregular speaker arrays using heuristic methods. In *Audio Engineering Society Conference: 31st International Conference: New Directions in High Resolution Audio*. Audio Engineering Society, 2007.

B. Wiggins, I. Paterson-Stephens, and P. Schillebeeckx. The analysis of multi-channel sound reproduction algorithms using hrtf data. In *Audio Engineering Society Conference: 19th International Conference: Surround Sound-Techniques, Technology, and Perception*. Audio Engineering Society, 2001.

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

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

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

H. Williams. Cold cold heart (performed by norah jones, emi, 2002). MGM, 1950.

J. Woodcock, C. Pike, F. Melchior, P. Coleman, A. Franck, and A. Hilton. Presenting the s3a object-based audio drama dataset. In *Audio Engineering Society Convention 140*. Audio Engineering Society, 2016.

B. Xie. *Head-related transfer function and virtual auditory display*. J. Ross Publishing, 2013.

W. A. Yost. Sound source localization identification accuracy: Envelope dependencies. *The Journal of the Acoustical Society of America*, 142(1):173–185, 2017.

P. T. Young. The rôle of head movements in auditory localization. *Journal of experimental psychology*, 14(2):95, 1931.

N. Zacharov. *Sensory Evaluation of Sound*. Taylor & Francis Group, 2018. ISBN 9781498751360. URL `https://books.google.co.uk/books?id=8znbswEACAAJ`.

N. Zacharov and K. Koivuniemi. Unravelling the perception of spatial sound reproduction: Language development, verbal protocol analysis and listener training. In *Audio Engineering Society Convention 111*, Nov. 2001. URL `http://www.aes.org/e-lib/browse.cfm?elib=9815`.

P. Zahorik. Assessing auditory distance perception using virtual acoustics. *The Journal of the Acoustical Society of America*, 111(4):1832–1846, 2002a.

P. Zahorik. Direct-to-reverberant energy ratio sensitivity. *The Journal of the Acoustical Society of America*, 112(5):2110–2117, 2002b.

M. Zaunschirm, C. Schörkhuber, and R. Höldrich. Binaural rendering of ambisonic signals by head-related impulse response time alignment and a diffuseness constraint. *The Journal of the Acoustical Society of America*, 143(6):3616–3627, 2018.

D. N. Zotkin, R. Duraiswami, and N. A. Gumerov. Regularized hrtf fitting using spherical harmonics. In *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 257–260. IEEE, 2009.

F. Zotter and M. Frank. All-round ambisonic panning and decoding. *Journal of the audio engineering society*, 60(10):807–820, 2012.

F. Zotter and M. Frank. *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*. Springer Topics in Signal Processing. Springer International Publishing, 2019. ISBN 9783030172060. URL `https://books.google.co.uk/books?id=v7rewgEACAAJ`.

E. Zwicker and H. Fastl. *Psychoacoustics: Facts and models*, volume 22. Springer Science & Business Media, 2013.

T. Łętowski. Sound quality assessment: concepts and criteria. In *Audio Engineering Society Convention 87*. Audio Engineering Society, 1989.

T. R. Łętowski and S. T. Łętowski. Auditory spatial perception: Auditory localization. Technical report, ARMY RESEARCH LAB ABERDEEN PROVING GROUND MD HUMAN RESEARCH AND ENGINEERING ..., 2012.

Figure Permissions

The following figures were reproduced under the Creative Commons licenses:

Fig. 2.3: Author: OpenStax, License: CC-BY-4.0, URL: `https://commons.wikimedia.org/wiki/File:1404_The_Structures_of_the_Ear.jpg`