A Simplified User-Assisted Binaural Filtering for Virtual Reality

Hongbo Hu

Master in Science by Research

The University of York

School of Physics, Engineering and Technology

February 2025

Abstract

Hearing is a fundamental channel through which humans perceive and interact with the external environment, with spatial auditory perception playing a key role in this process. Binaural audio systems, enabled by Head-Related Transfer Functions (HRTFs), allow virtual sound sources to be reproduced with spatial attributes, forming a core component of immersive virtual reality (VR) audio experiences.

This study investigates a simplified approach to HRTF-based binaural rendering, focusing on two aspects: perceptual evaluation of generalized HRTFs using a MUSHRA-based listening test, and the generation of personalised HRTFs through Interaural Time Difference (ITD) simplification based on anthropometric parameters.

Results from the subjective listening test indicate that certain generalized HRTFs can achieve competitive perceptual performance in terms of timbral quality and listener preference. In the VR-based localisation test, the proposed ITD-simplified approach demonstrated horizontal localisation accuracy and confusion rates comparable to a standard KU100 reference, although it did not consistently outperform fully personalised HRTFs across all spatial conditions.

These findings suggest that simplified HRTF manipulation can provide a computationally efficient and perceptually acceptable alternative for spatial audio rendering in VR applications, while highlighting the continued importance of spectral and individualised cues for precise localisation.

Acknowledgements

As this decades-long odyssey of learning nears its twilight—whether crowned with grace or etched with stumbles—I stand at the precipice between an ending and a fragile new beginning. To those who lit lanterns in my darkest academic nights, I offer these trembling words from the depths of my imperfect soul.

To my mentor, Prof. Gavin Kearney: You saw potential in my chaos. Through every delayed deadline and half-baked draft, you met my anxieties with the quiet strength of a redwood—rooted, enduring, endlessly giving. Your belief in me became the compass I lacked. This thesis carries fingerprints of your wisdom that no acknowledgement page could ever hold.

To my starlight, Yaxin Liu: Four revolutions around the sun since our souls collided, yet your love remains my first sunrise. In the silent hours when equations blurred and hope thinned, your hand in mine whispered, "We can weather this." You are my eternal equation—the proof that love outlasts all theorems.

To my parents, Guoyun Hu and Zhihong Zhang: Your sacrifices were the invisible ink staining every page I wrote. When roads diverged, your voices became my north star—never commanding, always illuminating. These degrees belong to your sleepless nights, your swallowed worries, your immigrant dreams.

To Prof. Damian Murphy: Your surgical critiques cut through my intellectual vanity, leaving scars that bloomed into clarity. Beneath the razor-sharp feedback flowed an unspoken creed: "You can do better—because you must." For that merciless kindness, I am forever altered.

To Dr. Tomasz Rudski, my brother-in-arms: In the trenches of code and circuitry, your patience turned my "stupid questions" into sacred learning. Our shared coffee-stained whiteboards hold more wisdom than any textbook. You taught me that true collaboration is the oxygen of innovation.

To Huan Mi, my anchor in life's tempest: The brother fate chose for

me. May our laughter forever drown out life's dissonance, and may our bond outlast every algorithm we'll ever write.

To this fractured, beautiful world: May the equations we solve tomorrow heal what we break today. Let knowledge be our peace treaty.

Hongbo Hu,
York,
2024.

Declaration

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

Contents

List of Tables

List of Figures

1 Introduction

1.1 Background

The convergence of Virtual and Augmented Reality (VAR) technologies is transforming numerous sectors, enhancing user experiences in fields ranging from gaming to healthcare and e-learning. Despite these advancements, the audio components within these technologies often lag behind their visual counterparts, particularly in terms of delivering personalized and immersive auditory experiences. This discrepancy primarily arises from limitations in current Head-Related Transfer Functions (HRTFs), which are crucial for accurate spatial audio rendering in three-dimensional environments.

HRTFs, akin to acoustic fingerprints, vary significantly among individuals due to differences in anatomical structures such as the head and ears. Traditional HRTF measurement techniques, while precise, are impractical for widespread use due to their complexity and the extensive resource requirements, including time and specialized equipment. Consequently, there is a pressing need for a more accessible approach that accommodates the variability in human anatomy while simplifying the customization process.

1.2 Hypothesis

Hypothesis 1: Generalized HRTFs still possess considerable value in specific aspects. It is posited that certain generalized HRTFs may exhibit superior performance from a non-localization perspective, demonstrating their utility beyond traditional spatial audio accuracy.

Hypothesis 2: Modifying some specific generalized HRTFs with individual physiological parameters to create new personalized HRTFs can achieve satisfactory outcomes. This approach not only potentially improves the efficacy of the HRTFs but also simplifies the measurement procedures and reduces the time required for HRTF customization.

1.3 Statement of Ethical Approval

Ethical approval was gained for all the listening tests undertaken via the Physical Sciences Ethics Committee of the University of York with Reference Numbers: Hu111120 and Hu2960721.

2 Literature Background Review

This section provides a foundational introduction to the concepts relevant to this study, as well as a review of the pertinent literature.

2.1 Sound Source Localisation

Single sound-source localization includes two aspects: direction and distance. Psychoacoustic research reveals that in the free field, directional localization cues for a single sound-source include Interaural Time Difference (ITD), Interaural Level Difference (ILD) and spectral cues. These cues also play an important role in distance perception. [7]

2.1.1 Interaural Time Difference(ITD)

The difference in the arrival time of a sound wave at each ear, known as the Interaural Time Difference (ITD), serves as a crucial cue for source localization. When the sound source is positioned in the median plane, the ITD is approximately zero, as the distances from the source to each ear are equal. However, when the source deviates from the median plane, the ITD becomes nonzero.

$$\Delta t = \frac{d \sin \theta}{c} \tag{2.1}$$

The equation 2.1 describes the rough ITD calculation, where Δt is the time difference between the ears, d is the distance between the ears, ϑ is the angle of arrival of the sound from the median and c is the speed of sound. However, the equation is not precise, as sound must travel around the head in order to arrive at different ears. Thus, the proper equation was introduced by Woodworth in 1938 and named as Woodworth's Formula. [8]

$$ITD = \frac{r(\theta + \sin\theta)}{c} \tag{2.2}$$

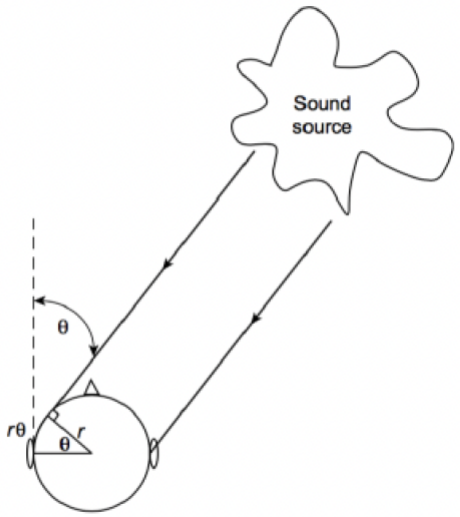

Figure 2.1: Path length around the head on the ITD

Generally, a typical human head radius is considered to be 0.09 meters, resulting in a maximum ITD of 673 microseconds ($\theta = 0.5\pi$). Although this time delay might appear negligible, it is critical for determining the direction of low-frequency sounds. Importantly, the human ear discerns time

delays through phase differences, imposing a frequency limit for effective ITD utilization. When the head dimension corresponds to half of the wavelength (equivalent to a frequency of 0.7 kHz), sources become out of phase, introducing potential ambiguity. At frequencies above 1.5 kHz, ambiguity becomes complete. Movement of the head or sound source can resolve some ambiguities at frequencies below 1.5 kHz; above this threshold, other auditory cues are necessary for sound localization. [9]

2.1.2 Interaural Level Difference(ILD)

Interaural Level Difference (ILD) serves as another critical cue for sound-source localization. ILD refers to variations in the sound pressure levels received by the left and right ears. When the sound source deviates from the median plane, the head's shadowing effect and diffraction cause attenuation of the sound pressure level at the ear further from the source, particularly at high frequencies. Conversely, the ear closer to the sound source experiences increased pressure levels due to its proximity to the sound-source.

$$ILD(r, \theta, \phi, f) = 20 \log_{10} \left[\frac{P_R(r, \theta, \phi, f)}{P_L(r, \theta, \phi, f)} \right] (dB) \tag{2.3}$$

The equation above describes the ILD, where $P_R(r, \theta, \phi, f)$ and $P_L(r, \theta, \phi, f)$ represent the sound pressures in the frequency domain from the source at coordinates (r, θ, ϕ), where r is the distance from the source to the head, and θ and ϕ are the azimuth and elevation of the head, respectively. However, when r is much greater than the head radius (in the far field), the ILD does not depend on r. [9]

It is challenging to quantify the shadowing effect; however, research indicates that the intensity ratio between the two ears changes sinusoidally, as illustrated in the figure below. The figure demonstrates that the lower the frequency, the smaller the variation. This occurs because the shadowing effect and scattering only manifest when the head diameter exceeds one-third of the wavelength. Consequently, there exists a minimum frequency below which ILD becomes ineffective. Given that the typical head diameter is 18 cm, we derive the following calculation:

$$f_{\min(\theta=\frac{\pi}{2})} = \frac{1}{3}\left(\frac{c}{d}\right) = \frac{1}{3}\left(\frac{344ms^{-1}}{0.18m}\right) = 637 \text{ Hz} \tag{2.4}$$

From the equation we can conclude that when the source frequency is less than 600–700 Hz, the ILD is usually no longer a significant cue for the sound-source localization. Thus, ILD is only an important localization cue for high frequencies of a sound-source.

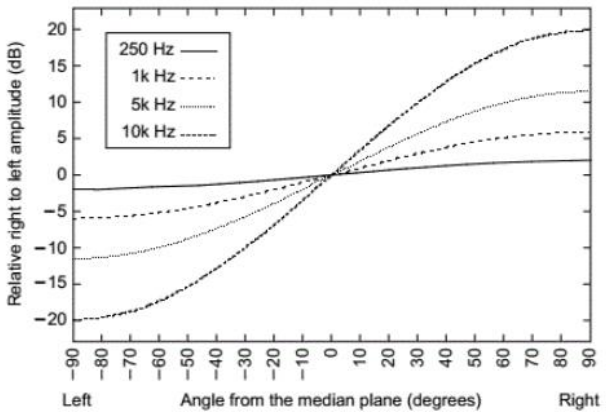

Figure 2.2: Interaural Level Difference - Function of angle and frequency

2.1.3 Cone of Confusion and Head Movement

ITD and ILD are critical cues for localization at low and high frequencies, respectively. However, as previously mentioned, relying solely on ITD and/or ILD can lead to front-back confusion when the sound source is located in the median plane. Moreover, research has shown that sound sources at the same azimuth but at different elevations may exhibit identical ITD and ILD values. This phenomenon is known as the 'cone of confusion,' as illustrated in the figure below:

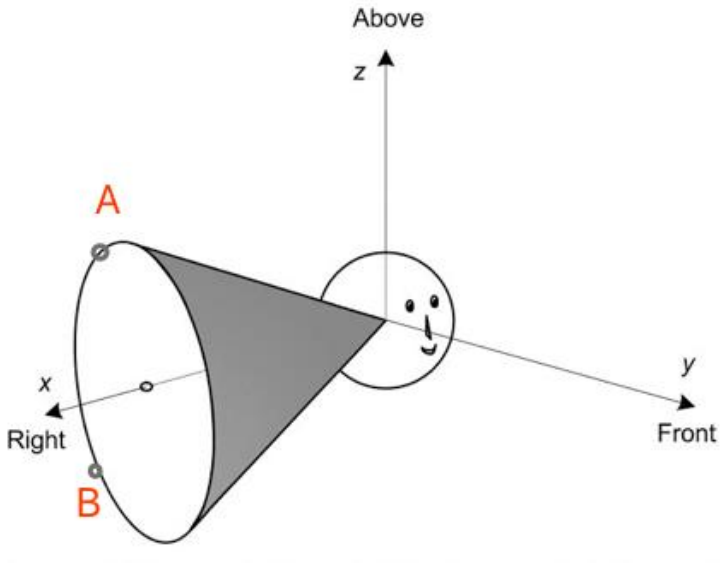

Figure 2.3: Cone of confusion in sound-source localization

The figure 2.3 illustrates that any sound sources located within this cone-shaped space will exhibit similar ITD and ILD values when the shape of the human head is approximated as a sphere (points A and B, despite their different positions, share the same ITD and ILD). Thus, while ITD and ILD can effectively indicate the direction of the cone, they cannot precisely localize the sound source. To address this issue, Wallach in 1940 [10] proposed

that head movements alter the ITD and ILD, significantly aiding in sound localization. Specifically, ITD remains positive when the source is in front of the listener and becomes negative as the source moves to the back of the head, as depicted in the subsequent figures.

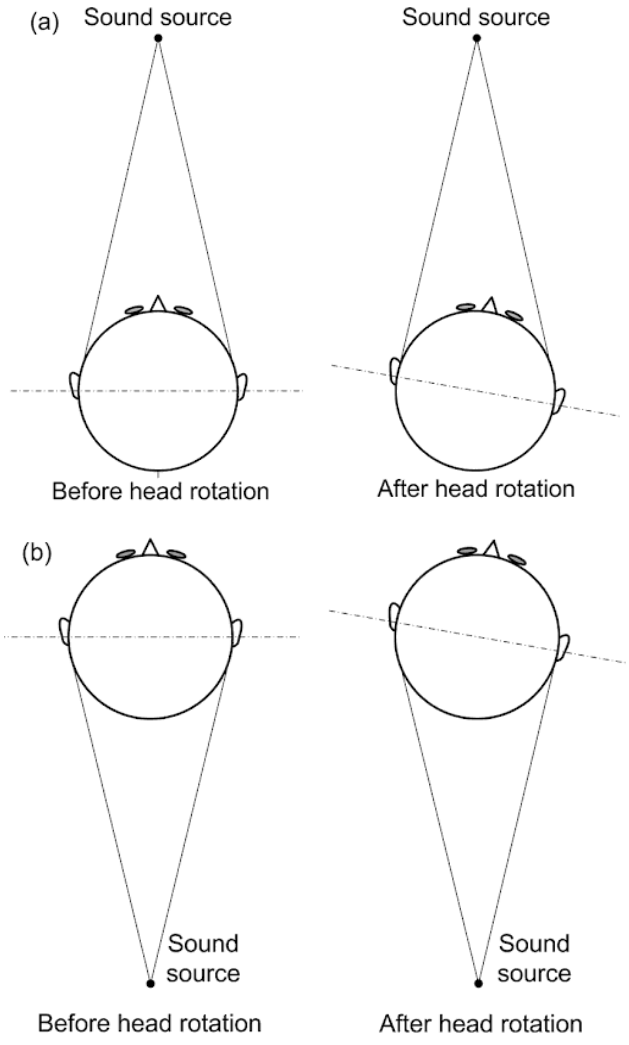

Figure 2.4: Changes in ITD caused by head movement

The theory discussed above was definitively proven by Perrett and Noble in 1995. [11] Furthermore, Rao and Xie in 2005 demonstrated that head

movements also provide cues for vertical localization, assisting humans in pinpointing low-frequency sounds in the median plane. [9]

2.1.4 Spectral Cues

Unlike ITD and ILD, which are binaural cues, the spectral cue (also known as a pinna cue) is a monaural cue. The reflection and diffraction of sound within the pinna alter the spectral characteristics of sound pressure, providing critical information for vertical localization and front-back disambiguation.

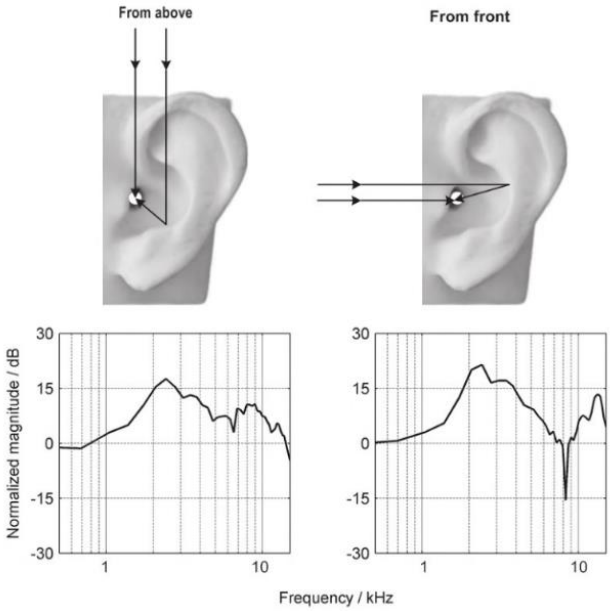

Figure 2.5: Interaction of sound and pinna in different directions

The pinna dimension is around 65mm, so the pinna cue is not useful for those soundwaves whose wavelength is shorter than 65mm. In short, when the frequency is higher than 2–3 kHz, the pinna cue starts working; when the frequency is greater than 5–6 kHz, the pinna cue plays a significant role [9].

Studies conducted by Lopez and Meddis in 1996 [12] reveal that reflective and diffractive sound will interfere with direct sound in the outer ear canal, which will change the spectral shape and produce a series of different crests and troughs. The change of spectrum is closely related to the variation in the direction of the incident wave. Similar to a fingerprint, the shape of the pinna is unique for everyone: Each person has their own different pinnae structure, and therefore, the spectral information is a highly individualized localization cue. [9] In addition to the pinna, some studies have also shown that the torso and hair also reflect and diffract the sound: the torso (especially the shoulders) changes the spectrum below 3 kHz, which can act as a localization cue in the median plane [6, 13, 14].

2.2 Binaural Audio and HRTFs

Listening to stereo audio over headphones provides a different auditory experience compared to loudspeakers. Speakers, positioned directly outside the ears, naturally affect the audio presentation and create a spatial image that differs from regular loudspeaker playback. One method that leverages headphone playback to create 3D-like audio is the binaural recording technique. This involves placing two microphones in the ears of a dummy head, which, when played back over headphones, recreates the soundfield as if the listener were physically present. However, for a more effective and dynamic immersive audio experience, binaural filters must be introduced. Head-Related Transfer Functions (HRTFs), which are a pair of transfer functions, describe the frequency differences between the original sound source and the sound measured at the entrance of a person's (or a dummy head's) left and right

ear canals. Ideally, filtering the binaural audio signal with the correct HRTF allows listeners to precisely perceive the spatial~~origin~~[Hongbo] of the sound.

2.2.1 Calculation of HRTFs

Typically~~In nature~~[Hongbo], the transmission process from a point source to each of the two ears on a fixed head can be regarded as a linear time-invariant process. HRTFs describe the overall filtering effect imposed by anatomical structures [13]:

$$\begin{aligned} H_L &= H_L(r,\theta,\phi,f,a) = \frac{P_L(r,\theta,\phi,f,a)}{P_0(r,f)}, \\ H_R &= H_R(r,\theta,\phi,f,a) = \frac{P_R(r,\theta,\phi,f,a)}{P_0(r,f)}, \end{aligned} \tag{2.5}$$

where P_L and P_R represent the complex-valued sound pressures in the frequency domain at the left and right ears; P_0 represents the complex-valued free field sound pressure in the frequency domain at the centre of the head with the head absent, r denotes the source distance, θ and ϕ represent azimuth and elevation angles respectively, f denotes frequency, and a represents additional parameters related to the acoustic configuration.[Hongbo]

2.2.2 HRTFs in time domain and frequency domain

Head-Related Impulse Responses (HRIRs) are the time-domain counterparts of HRTFs. An impulse response characterises how a linear time-invariant system, such as an acoustic path, responds to a broadband impulse signal [13]. In binaural audio, the HRIR represents the time-domain description of the acoustic filtering introduced by the listener's anatomy (e.g.,

head, pinna, and torso), capturing spatial cues such as interaural time differences (ITDs) and direction-dependent spectral shaping [15]. This relationship underpins binaural recording and synthesis: by convolving a dry audio signal with measured HRIRs, spatial attributes of a sound field can be reproduced over headphones, enabling the perceptual impression of externalised sound sources.~~Head Related Impulse Response (HRIRs) are the time-domain counterpart of HRTFs.~~[Hongbo]

Figure 2.6: KEMAR far-field HRIRs for several different source azimuth in the horizontal plane

Figure 2.6 and 2.7~~Above figure~~ [Hongbo]illustrates the part of the far-field HRIRs measurement of the binaural dummy head KEMAR (Knowles Electronic Mannequin for Acoustic Research) by the MIT Media Lab,. This example contains different azimuths from 0^{o} to 180^{o} with increments of 30^{o} (only shown 0^{o}, 30^{o}, 60^{o} above) [9]. From the figure 2.6~~From the Figure 9~~[Hongbo], the HRIR amplitude at the preceding 30-50 samples is approximately zero, which is corresponding to the propagation delay from sound source to ear while the main part of the HRIRs reflects the complicated interactions between incident sound waves and the pinna, head and torso. Most of the

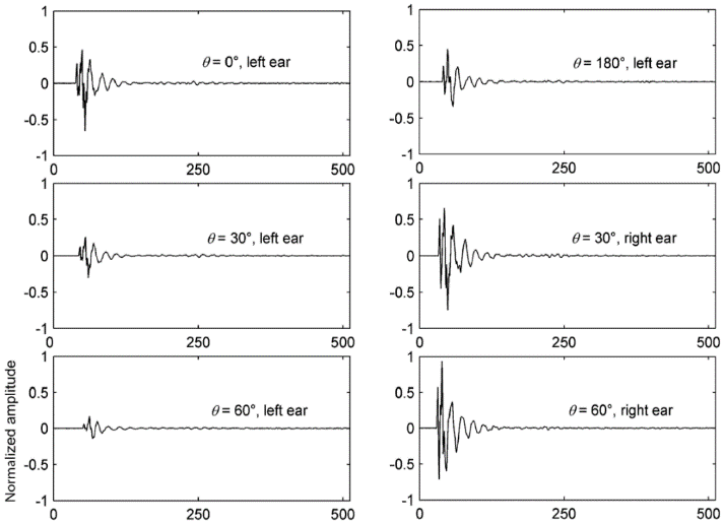

Figure 2.7: KEMAR far-field HRIRs for several different source azimuth in the horizontal plane

energy is within 50 samples (with a sample rate of 44.1kHz). Subsequently, the HRIR amplitudes return to nearly zero.

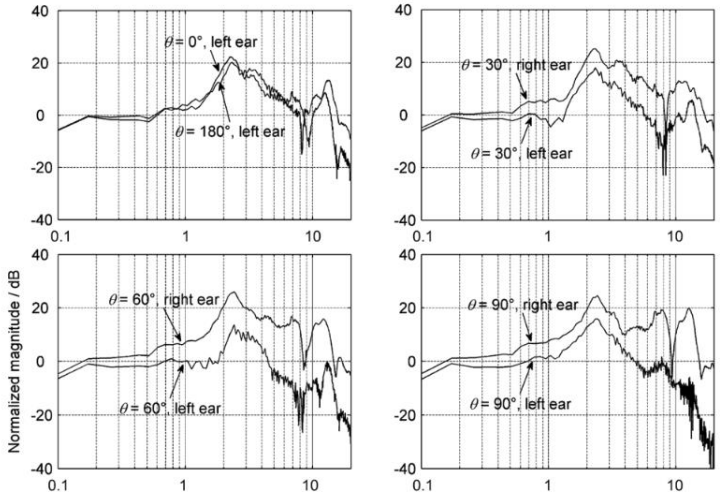

Figure 2.8: Magnitudes of KEMAR HRTFs at various azimuths in the horizontal plane

Figure above presents the normalized (logarithmic) HRTF magnitudes of

KEMAR at some different azimuths. The decrease in magnitude below approximately 150 Hz is caused by the low-frequency limit of the loudspeaker response used in this measurement. At frequencies below 0.4-0.5 kHz, the normalized magnitudes of HRTFs approach 0 dB, and are roughly independent of frequency because the scattering and shadowing effects of the head are negligible. As frequency increases, the normalized magnitudes of the HRTFs vary with frequency and azimuth in a complex manner. This complexity is attributed to the overall filtering effects of the head, pinna, torso, and ear canal.

2.2.3 Excitation Signals for HRTF Measurement

A HRTF is a transfer function in the frequency domain (as mentioned in last section, HRIR is the time domain equivalent). There are many methods for measuring Head-Related Impulse Responses (HRIRs), from which HRTFs can subsequently be obtained via Fourier transformation.~~There are many methods for measuring a HRTF.~~[Hongbo] The typical measurement process is shown in the block diagram below:

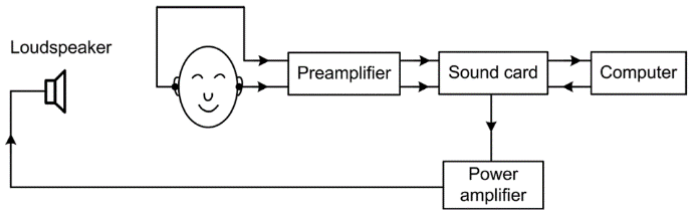

Figure 2.9: Block diagram of HRIR measurement processing

The measurement is usually undertaken in an anechoic room to minimize ~~the~~[Hongbo] room reverb. There is usually a loudspeaker array on a spheri-

cal surface providing different directional excitation signals, all of them are aiming to create the perfect impulse signal. A good excitation signal should have a wide and equally spread frequency range, also with a good signal to noise ratio (SNR) to separate the signal and the unwanted background noise. Some of the typical signals are introduced below:

- Impulse Signal

An ideal impulse signal is modelled as a Dirac delta function, which is a mathematical construct representing an infinitely short signal with unit area. Although it cannot be physically realised, the Dirac delta is theoretically significant because its Fourier transform has constant magnitude across all frequencies and zero phase shift, meaning it excites all frequency components of a linear time-invariant system equally. When such an impulse is applied to an acoustic system, the resulting output directly corresponds to the system's impulse response.[Hongbo]

In practical acoustic measurements, the impulse signal is only an approximation of the ideal Dirac delta function. Traditional impulse sources used in room acoustic measurements include starting guns, electrical sparks, and bursting air balloons [16]. However, the physical properties of these sources are difficult to control, and the excessive transient sound pressure levels may introduce nonlinear propagation effects in air [17].~~The ideal impulse signal is a Dirac delta function, a deterministic signal with a flat magnitude spectrum and linear phase. The impulse signal used in actual measurement is only an approximation of the ideal Dirac delta function. Starting guns, sparks, and popping air balloons have been used as traditional impulse sources in~~

~~room acoustic measurements [18]. However, the physical properties of these sound sources are difficult to control. Moreover, the excessive transient sound pressure is likely to cause a nonlinear effect in the air. [17]~~[Hongbo]

- Linear sine sweep

A linear sine sweep is usually considered a 'white' excitation signal with a certain frequency range. The term 'linear' refers to the fact that the frequency changes linearly with time. By deconvolving the linear sine sweep, the linear impulse response of a system can be retrieved. Deconvolving a linear sine sweep uses a time-reverse filter technique, and the result is an impulse response in the time domain, an HRIR in this case. The equation of this signal is shown below, where the signal lasts for T seconds and contains a sine with linearly increasing frequency from ω_1 to ω_2 . [17]

$$x_{\text{linsweep}}(t) = \sin\left(\omega_1 t + \frac{\omega_2 - \omega_1}{T}\frac{t^2}{2}\right) \tag{2.6}$$

However, in the time domain signal, linear sine sweep usually causes some strange oddities, also called 'wraparound' effects [18]. Moreover, a linear sine sweep does not ensure the same accuracy for all frequency components unless the spectrum of the background noise is also flat, but the background noise is usually more prominent at low frequencies. The diagram of the linear sine sweep waveform is shown in Figure 2.10~~Figure~~[Hongbo] below:

- Logarithmic Sine Sweap

Although all frequencies are equally excited on linear sine sweep, less time is spent on excitation of low frequencies. Therefore, a logarithmic sine sweep (also known as Exponential Sine Sweep (ESS) in some literature) is designed

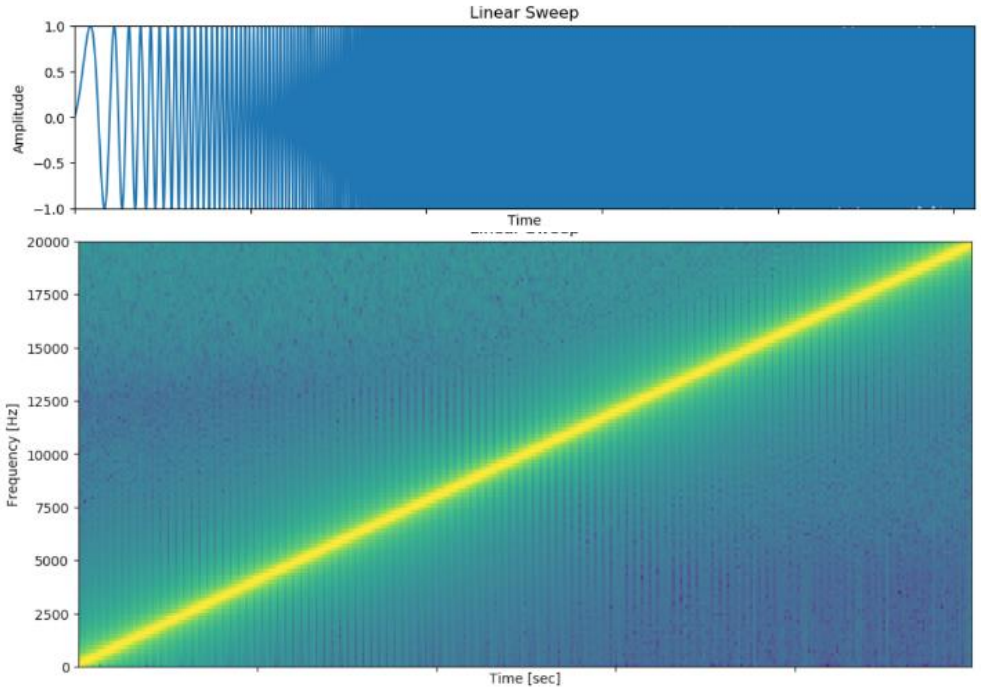

Figure 2.10: Linear sine sweep in both time and frequency domain

to emphasise the low-frequency signal [17]. The equation of the logarithmic sine sweep is shown below:

$$x_{\text{logsweep}}(t) = \sin\left[\frac{\omega_1 T}{\ln\left(\frac{\omega_2}{\omega_1}\right)}\left(e^{\frac{t}{T}\ln\left(\frac{\omega_2}{\omega_1}\right)}1\right)\right] \tag{2.7}$$

The sweep rate of the logarithmic sine sweep is not constant. The signal grows slowly in the beginning, low frequency and rises rapidly in the later high frequency, to give the low frequencies more time to evolve. [18] The diagram of the logarithmic sine sweep waveform is shown in Figure below

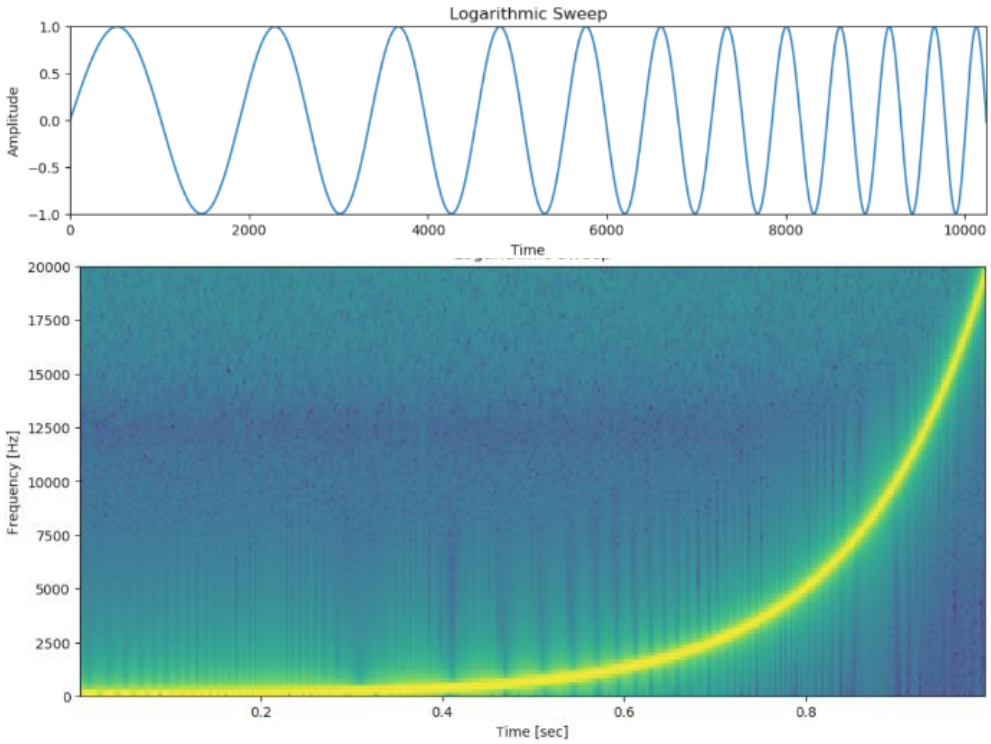

Figure 2.11: Logarithmic sine sweep in both time and frequency domain

- Random Noise Signal

A stationary noise signal such as white noise or pink noise, is a continuous signal with wide frequency range and can be a good excitation signal. These noise signals have a predictable frequency response that is even powered in a linear or logarithmic scale of the spectrum. [9] However, it is difficult to replicate the noise signal exactly the same every time, because of that the signal processing of the measurement can be quite complicated, so some of the HRTF measurements choose pseudorandom signal as the excitation signal, which is introduced in the next coming section.

- Maximal length sequence (MLS)

Maximal Length Sequence(MLS) is one of the popular pseudorandom noise commonly used in different acoustic measurements. It is a binary sequence of a series of integers and periodic and has a good signal-to-noise ratio which can be improved 3dB by doubling the MLS period length. [18] The major problem of the MLS method resides in the appearance of distortion artefacts that introduce the characteristic crackling noise when the impulse response is convolved with some other signals from an anechoic environment. [19]

2.3 Overview of HRTF Measurement and Evaluation

This section introduces common issues associated with measuring individualized HRTFs, and some improvements in HRTF measurement and synthesis. These topics will be discussed in detail, combined with listening tests, in later chapters.

2.3.1 Recording in the eardrum

This is the direct method that was initially adopted, of recording the sound pressure at the position of the eardrum. This method is always considered inconvenient for human subjects (see Figure 2.12 below): the probe microphone placement is uncomfortable for humans, and the frequency response is also poor (the recording signal from a typical probe microphone always needs to be equalized or boosted in some frequency range) and less sensitive. On the other hand, if the microphone is too close to the eardrum (1mm to 2mm), it is possible to physically damage the eardrum when setting up the microphone, and the damage is irreversible. [9] [19]

2.3.2 Recording at the entrance to the blocked ear canal

The blocked ear canal method was first introduced by MøllerMoller[Hongbo]. It is safer and more convenient than the eardrum method. [2] A pair of miniature microphones with high sensitivity and wide frequency response range is used for this recording. The sample position is shown in the Figure

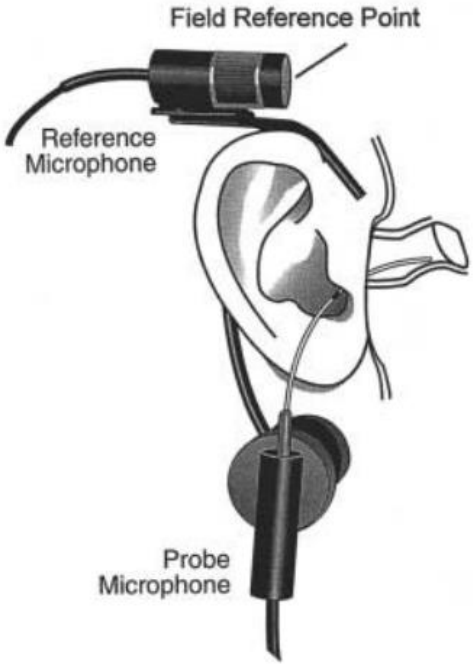

Figure 2.12: Typical arrangement of probe and reference microphone [1]

2.13:~~following figure:~~[Hongbo]

Figure 2.13: The position of the microphone inside a participant's ear [2]

This method is the most commonly applied in the latest binaural recording for human subjects. [6] [20] [21] Although many of the foundational studies on the blocked ear canal method were published between the mid-1990s and early 2000s, the technique remains widely adopted in contemporary binaural recording research and practice. This is largely because the methodological framework and measurement principles established in these earlier works continue to underpin modern implementations, with subsequent developments focusing primarily on refinements in hardware, calibration procedures, and post-processing techniques rather than fundamental changes to

the measurement concept itself.Hongbo

2.3.3 Inverse Method

Different from the traditional method, an inverse method was proposed by Zotkin et al in 2006, [3] which put a microspeaker in ear canals and placed a microphone array around the head (see Figures 2.14 and 2.15).~~(see the following Figures).~~Hongbo

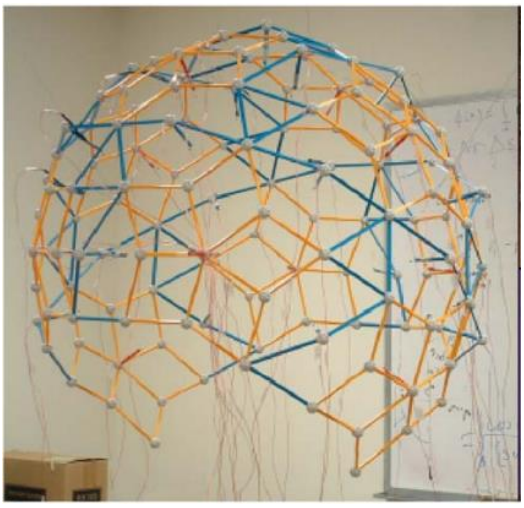

Figure 2.14: The microphone array [3]

Figure 2.15: The size of microspeaker [3]

However, there are two major problems when using this method:First, although the frequency response of the microspeaker is approximately flat

between 1 kHz and 16 kHz, its performance at low frequencies is limited, making it insufficient for accurate broadband HRTF measurements.Second, there are safety concerns related to sound pressure levels. To improve the signal-to-noise ratio (SNR), the output amplitude of the microspeaker must be increased. However, due to the close proximity of the speaker to the eardrum, elevated sound levels may cause discomfort and pose potential risks to the listener [3].~~First, the result shows that the frequency spectrum for the microspeaker is roughly flat from 1 kHz to 16 kHz which performed poorly on low frequency, this is not sufficient enough. Another problem is the safety concern on sound level, in order to improve the SNR, the amplitude of output signal from the speaker needs to be higher, but this is unsafe and uncomfortable for human especially when the speaker is so close to the eardrum. [3]~~Hongbo

2.3.4 Discussion of the HRTF Measuring Method

Both the eardrum and inverse methods carry a high risk of damaging the eardrum, and the inverse method lacks sufficient precision. Consequently, most researchers today prefer to use the blocked ear canal method to record signals. However, when employing the blocked ear canal recording technique, the absence of natural ear canal resonance at the eardrum may compromise the accuracy of frontal source localization. This is particularly relevant for median-plane perception, where subtle spectral cues introduced by the ear canal resonance contribute to accurate front localisation. To enhance results, some studies recommend equalizing headphones to more accurately reproduce the timbre of a frontal sound source at the

eardrums [22].~~However, while employing the blocked canal recording, the absence of resonance at the eardrum may compromise the accuracy of frontal source localization. To enhance results, some studies recommend equalizing headphones to more accurately reproduce the timbre of a frontal sound-source at the eardrums. [22]~~[Hongbo] In summary, these measurement techniques for human subjects are time-consuming, costly, and technically complex. Although commercial efforts towards individualized audio personalisation have emerged in recent years—such as self-calibrating headphone systems exemplified by Nura's personalised sound technology [23] and software-based solutions such as Dolby Personalised Audio [24]—these approaches typically rely on perceptual calibration or simplified modelling rather than full individualized HRTF measurement. Academic research has also explored perceptual calibration strategies for spatial audio personalisation [25]. Nevertheless, while the industry is actively pursuing personalised spatial audio, large-scale implementation of precise, measurement-based individual HRTF acquisition remains impractical for widespread commercial deployment due to financial, technical, and time constraints, as such systems generally prioritise usability and scalability over laboratory-level precision.~~In summary, these measurement techniques for human subjects are time-consuming, costly, and complex. Currently, it is not feasible to conduct individualized measurements for commercial purposes on every person.~~[Hongbo]

2.4 Hardware Requirement for HRTF Measurement

HRTF measurements are typically conducted in an anechoic or a well-designed semi-anechoic environment to minimize room reverberation. The output for the excitation signal generally involves a loudspeaker array. For measuring human subjects, equipment such as a motor-controlled chair, head tracker, and laser calibrator are required to stabilize the body and adjust the azimuth and elevation. [26] This section discusses the environmental setup from typical measurements published by various institutes.

2.4.1 Loudspeaker Setup

The HRTF for a specified ear at a certain sound-source position is obtained using a pair of transfer functions: the transfer function of the sound propagation path from the sound source in a specific direction to the entrance of the subject's ear canal, and from the sound source to the center of the head when no subject is present. The more measurement points that are provided, the higher the spatial resolution will be. Typically, loudspeakers move around the test subject to capture different angles of measurement, or alternatively, the subject is rotated while the loudspeakers remain stationary. All of these methods adhere to the stop-measure-go process, as movement inherently produces noise. However, the stop-measure-go method is highly time-consuming, with a typical measurement for a human subject requiring at least 1.5 hours, and often about 3 hours. [9] Studies conducted by Ville Pulkki et al. introduced a method involving a continuously moving

loudspeaker and swept sine waves to decrease the time cost and improve efficiency, [27] although the results have not met expectations.

2.4.2 Other Supportive Equipment

Microphones and loudspeakers are at the core~~cores~~[Hongbo] of the HRTF measurement, but other supportive equipment still plays an important role. The motor-controlled rotating chair is used for rotating the test subject to obtain a different angle. [28] To track the test subject and make sure the test subject is in the correct position with the expected angle, the tracking device is always used. Head position could be tracked in real-time via a head tracker. The SADIE II database measurement method~~SADIE II database measurement~~[Hongbo] used a multi-purpose restraint on the head of test subjects. [29] Multiple reflective markers on this device can be captured by Infra-Red motion capture cameras and related software. The resolution is down to 0.1°. [1]

2.5 Critical Evaluation of Different HRTF Measurement

Quality and error in HRTFs have been extensively evaluated by many researchers, [9, 20, 30, 31] including assessments of signal-to-noise ratio measurement errors. However, since HRTFs are typically applied in virtual reality or immersive audio environments, where the user experience regarding sound quality and localization accuracy is highly subjective, perceptual evaluation becomes crucial. Some listening tests have indicated that generic

HRTFs can sometimes yield better results than individualized HRTFs. Consequently, this section focuses on the perceptual evaluation of selected HRTF databases.

2.5.1 Localisation Accuracy for Real Sound-Source

Localization accuracy is one of the most crucial features used to evaluate the suitability of a HRTF for a user. Localization blur refers to the smallest perceptible change in the direction of a sound source [32]. Researchers typically employ two methods to test localization accuracy. The first method, known as absolute measurement, requires participants to specify the exact position of the source. This method is commonly used during loudspeaker playback [28]. The second method involves searching for the Minimum Audible Angle (MAA) or the Just Noticeable Difference (JND), wherein the tester compares the reference audio with the test audio to identify the smallest change in source direction [33].

Numerous tests on localization accuracy have been conducted in both free-field and virtual acoustic environments. An overview of experimental results prior to 1970 was provided by Blauert [34], who found that the most precise spatial hearing occurs in the forward direction. The absolute lower limit for localization blur is about 1 degree with broadband signals in free-field listening. [35] Research by Makous and Middlebrooks has also demonstrated that discrimination error is minimal in the forward direction on the horizontal plane, approximately 2° frontal horizontal and 5° vertical. From the sides, the error increases to about 20° [34].

More recent work has extended these classic paradigms to headphone-

based and virtual sound reproduction contexts, where additional factors such as HRTF individualisation, head movements (dynamic rendering), and multimodal cues can substantially influence localisation accuracy. For example, perceptual evaluations comparing individual and non-individual HRTFs highlight that localisation is only one component of overall spatial quality and may be accompanied by changes in timbre and externalisation [29]. In virtual and head-tracked conditions, dynamic binaural synthesis has been reported to improve localisation performance and reduce reversal errors (e.g., front–back confusions) compared with static rendering [36, 37]. Threshold-based measures have also been adapted to virtual sound synthesis, supporting the use of MAA-style procedures in headphone reproduction settings [38]. Furthermore, studies on training and adaptation suggest that listeners can partially accommodate poorly matched HRTFs over time, which is relevant when assessing "suitability" beyond one-off accuracy measurements [39, 40].
Hongbo

2.5.2 Localisation Accuracy in the Horizontal Direction Using Individulased and Generalised HRTFs

As previously mentioned, direct personalized HRTF measurement for applications targeting the general public seems challenging to implement. Many studies have investigated the subjective selection of HRTFs from established databases. [41] Among the most widely used databases for such work are the CIPIC HRTF Database [6], which provides high spatial-resolution measurements for 45 subjects along with detailed anthropometric data, and the RIEC HRTF Database [42], developed by Tohoku University, which offers

densely sampled individualized HRTFs measured under controlled anechoic conditions. These databases are frequently adopted in perceptual evaluation studies due to their accessibility, standardized measurement procedures, and relatively high angular resolution.[Hongbo] H. Su [4] conducted an experiment to assess localization perception and perceived width in a VR environment, using subjectively selected HRTFs with the assistance of a head-tracking device. Before conducting listening tests, participants were required to rate HRTFs provided from the CIPIC and RIEC databases. The highest-scored HRTF was treated as the individualized one, while the lowest-scored was considered a non-individual HRTF. Both scenarios were tested, with and without the head tracker. The sound sources for the localization tests were noise pulses, and for the width evaluation test, anechoic cello recordings and pink noise, each with a duration of 10 seconds. [4]

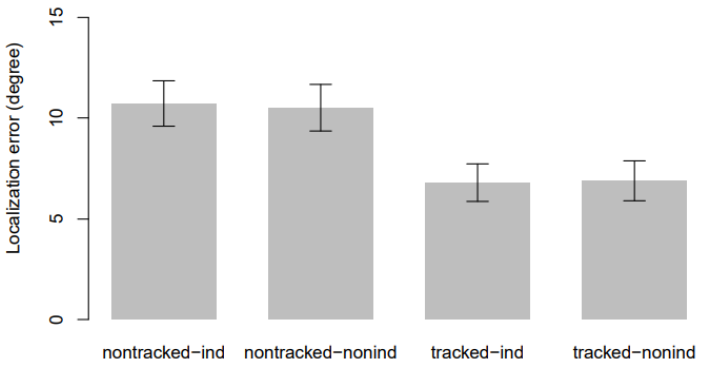

Figure 2.16: Average absolute angular errors of localization from [4]

Test results are shown in Figures 2.16 and 2.17, which suggest that head tracker improved the localization accuracy significantly. However, the result did not show whether the localization accuracy is improved by using

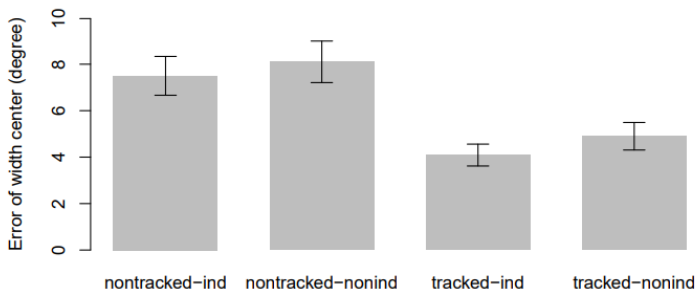

Figure 2.17: Average absolute angular errors of centers of source widths from [4]

individualized HRTFs. The researcher claims that this is probably due to the directions of the[Hongbo]source being limited to the horizontal plane in the study. A localization test based in different elevation plane is therefore introduced in next section.

2.6 Personal HRTF Matching

HRTF matching aims to determine the user's preference from different HRTF datasets or standardized HRTFs. As discussed in previous sections, personalized HRTFs are tedious to acquire and currently not feasible for public and commercial applications. Traditionally, users are required to rate various HRTFs from a database and select the best fit based on their own preferences—a process known as subjective matching, [43] which can still be time-consuming. Consequently, recent studies have begun exploring alternative methods, such as anthropometric measurements, [5] and the use of deep learning neural networks to assist in HRTF matching. This section provides an overview of these HRTF matching methods.

2.6.1 Anthropometric measurements based HRTF matching

Research conducted by Zotkin et al. proposed a fast HRTF matching method by using camera capture the anthropometric structure of user's ears, each ear is providing different features. [5]

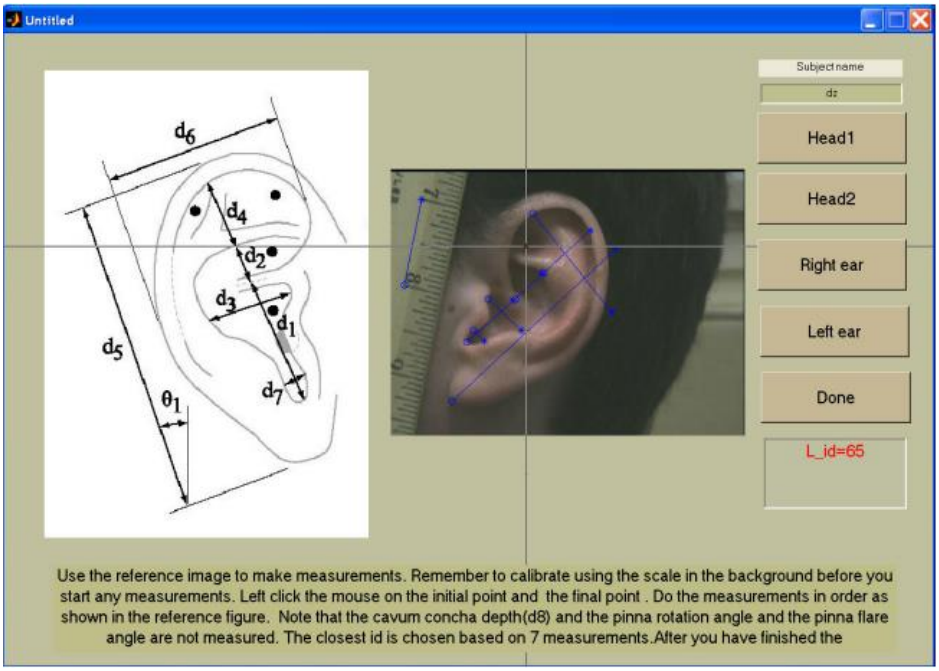

Figure 2.18: Screenshot of the HRTF customization software from [5]

Figure 2.18 displays features labeled d1 to d7 captured by the camera, which include cavum concha height, cymba concha height, cavum concha width, fossa height, pinna height, pinna width, and intertragal incisure width. [5] The best-fit personalized HRTF from the CIPIC database is selected by capturing these anthropometric measurements as one of the test models, referred to as a personalized HRTF. Additionally, HRTFs for KEMAR (with small pinna) serve as a generic HRTF test model. The CIPIC Interface

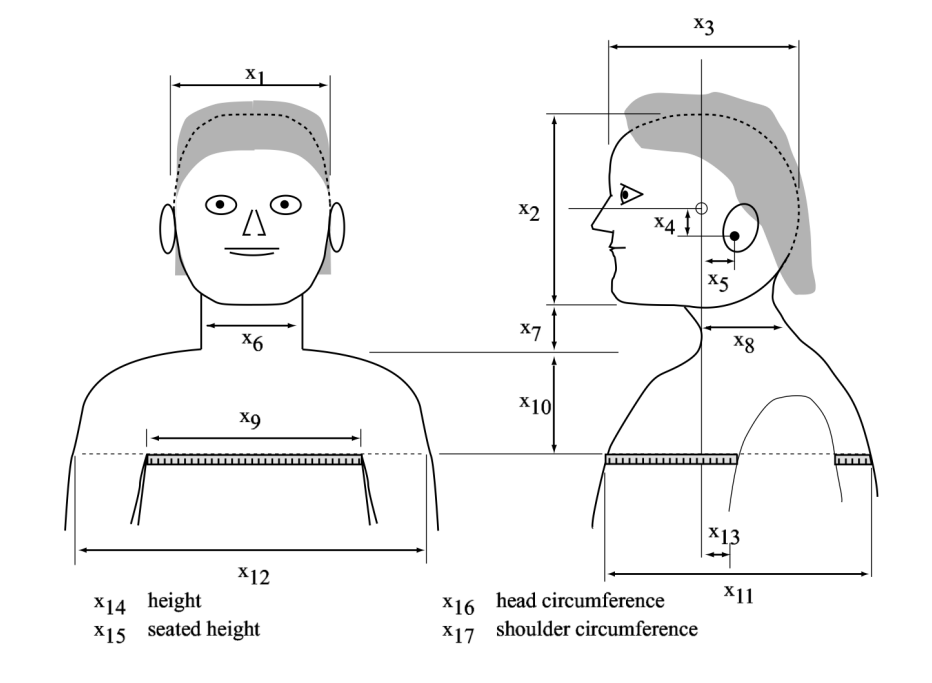

Figure 2.19: Head and torso measurements illustration from [6]

Laboratory has introduced a simple head-and-torso (HAT) model (snowman model), which is combined with the anthropometric method to deliver a personalized-plus-snowman HRTF and applied as another test model(see Figure 2.19). KEMAR + HAT is used as the fourth test model. Similar to other listening tests that use the MAA method, participants were required to pinpoint the exact position of the virtual sound source. The results showed that the Generic HAT model improves localization accuracy, whereas the personalized HRTF did not perform as well as expected. [44]

3 Listening Test for HRTF Comparison

3.1 Background of the Listening Test

Recent developments in multimedia technology have further increased the demand for headphone-based 3D audio reproduction. Although binaural and HRTF-based rendering techniques have been used in video games and multimedia applications for decades, their relevance has expanded significantly with the rise of immersive media platforms. For example, in First Person Shooter (FPS) games, headphone-based 3D audio enhances players' localisation of sound sources and spatial awareness, contributing to a more realistic and competitive gaming experience [45]. Similarly, spatial audio rendering is widely employed in film and music streaming platforms (e.g., Dolby Atmos for headphones), enabling immersive listening experiences even without dedicated VR/AR hardware [46, 47]. The emergence of virtual and augmented reality has further intensified this trend, as accurate spatial audio reproduction becomes essential for perceptual realism and user immersion.~~Recent developments in multimedia technology such as virtual or augmented reality have popularised the use of 3D audio over headphones. For example, in First Person Shooter (FPS) games headphone based 3D audio can enhance player's localisation of sound sources and sense of acoustic environment, and therefore~~

~~creates a more realistic gaming experience [45]. It can also be adopted to provide an immersive experience when watching movies and listening to music on portable multimedia devices.~~ Hongbo

Binaural surround sound is a method of reproducing headphone based 3D audio that can give the end user a convincing experience that sound sources are externalised outside of the head and located at a specific angle and distance. Whilst it is ultimately a stereo reproduction technique, the method superimposes psychoacoustic cues onto the reproduced sound, making the user believe that the sound played is a convincing render of a real world stimulus. Binaural recording can be achieved using a dummy head with microphones in its ears and when recordings are played back, the sound field is recreated as if the listener were there [13]. Binaural sound can be synthesized using filters known as Head-Related Transfer Functions (HRTFs) which are capable of describing the frequency difference between an original sound-source and the sound measured at the ear canal entrance of a person's (or a dummy head's) left and right ears. Ideally, when an audio signal is filtered with well-matched HRTFs to the listener, they can precisely localise the Hongbo source of the sound at the intended source position [9].

The significance of HRTF measurement cannot be underestimated. As with fingerprints, the shape of the ear and torso changes from individual to individual and so the corresponding individualized HRTF is also unique. If we can measure an individual's HRTF, the quality of binaural reproduction will be noticeably improved, and virtual sources will start to be indistinguishable from real sources [48]. However, conventional HRTF measurement methods are generally time consuming, usually requiring participants to remain as

motionless as possible during the measurement, with the process ~~is~~Hongbo likely to take several hours. There has been some experimental work on innovative or rapid personalised HRTF measurement methods [49–51], although results are not indicated to be as satisfactory as conventional methods.

Conversely, artificial head models can be employed for HRTF measurements, instead of real human subjects. The Neumann KU100 and the Knowles' Electronic Manikin for Acoustic Research (KEMAR) are artificial heads commonly applied in research and commercial applications [52, 53]. Replacing a real human with an artificial head model for HRTF measurements is considered cost-effective. However, the defects of using an artificial head model are also apparent. As reported by some earlier studies, generic HRTFs are lacking in spatial accuracy and perception of externalization [54, 55]. Moreover, Zamir et al. performed a series of listening test to investigate the difference of the perceptual localisation accuracy when using personalised HRTF, Generic HRTF(KEMAR) and a real sound source. The results show that the use of generic HRTFs leads to relatively more perceptual errors than personalised HRTFs, either in azimuth, elevation or front-back confusion [56].

Nevertheless, some researchers argued that besides the localisation accuracy in virtual sound sources, the differences in timbre and spatial characteristics that arise from the use of different HRTFs are worth studying in depth [57, 58]. In 2018, Armstrong et al. presented an insight into the perceptual evaluation of individualized HRTF and non-individualized HRTF by constructing a HRTF database for 20 subjects (e.g., KU100 and KEMAR) and subjectively assessing this database in accordance with the brightness and richness of the timbre, externalization, as well as the overall prefer-

ence [29]. As indicated by their above results, the KU100's HRTF was rated as the most preferred, instead of the subject's own individualized HRTFs, and there was a correlation between the participants' preference and the brightness, richness of the timbre and externalisation. This paper postulates that body movement from the human subjects might have an effect on the quality of the measurements as well as the overall performance of the individualized HRTF. The above findings confirmed that the use of generic HRTFs is of high significance.

Over the past few years, universities and research institutes have constructed their own HRTF databases containing a varying number of individualized HRTFs based on real human subjects, as well as generic HRTFs based on artificial manikins (e.g., Neumann KU100 and KEMAR). The spatial resolution of recent databases has also significantly increased. For instance, the SADIE II project contains up to 8802 measurement points [29], while the typical 'CIPIC' database contain 2500 points and the UMD-University of Maryland database cantain only 823 points.There are also differences in methodology and hardware utilised during the measurement, all leading to greater variability between different HRTF databases.

Katz et al. conducted a public project called 'Club Fritz', containing more than 60 KU100-based HRTF databases from different research laboratories worldwide, and their study drew a full comparison of the physical properties of the different HRTFs (e.g., Interaural Time Difference (ITD) Variation and spectral magnitude differences). Although their study was extraordinarily detailed, it essentially focused on the comparison of objective technical data, instead of on a subjective evaluation between the above databases [59]. Prior

to their study, Katz et al. also conducted a subjective evaluation of six different HRTF sets from IRCAM's LISTEN HRTF database [60], in which a listening test was performed to evaluate sense of direction, sense of distance and front image quality [61]. As revealed by the results, the variability across trials was significant for all subjects.

Existing evaluations of HRTF databases primarily focused on the perception of individualized HRTFs, and evaluations of generic HRTFs largely focused on objective analyses of databases.Thus, in this study, a subjective listening test is outlined based on evaluation of different generic HRTF databases.

3.2 Experimental Design

3.2.1 Selected HRTF Databases Review

After an initial screening and comparison procedure, six generic HRTF databases measured using the Neumann KU100 artificial head were selected for evaluation. The screening criteria were defined prior to the listening tests to minimise experimenter bias and ensure methodological consistency. First, only databases measured with the same artificial head model (KU100) were considered, in order to control for structural variations across different dummy heads. Second, databases were required to provide relatively high spatial sampling density (generally exceeding 1500 measurement points) to ensure sufficient angular resolution for perceptual comparison. Third, complete documentation of measurement conditions (e.g., anechoic environment and angular spacing) was required to ensure comparability.Based on these

criteria, six databases were retained: three from the "Club Fritz" project by Katz et al. (CF1-IRCAM, CF4-IRCAM XMod, CF5-ITA), two from the SADIE project (SADIE I and SADIE II), and one from THK. These databases represent well-documented research-grade datasets with relatively dense spatial sampling compared to many earlier HRTF collections. The specific parameters of the selected databases are summarised in Table 3.1. ~~After the preliminary screening and comparison, six different generic HRTF databases containing KU100 were selected for evaluation, which consisted of three from the "ClubFritz" project by Katz et al., two from the author's research institute and one from the University of Cologne. They all have relatively more measurement points than most other databases in the field. The specific parameters of the above databases are presented below and listed in Table 1:~~[Hongbo]

(1) Club Fritz IRCAM: 1-Institut de Recherche et Coordination Acoustique/Musique

(2) Club Fritz IRCAM XMod: 4-Institut de Recherche et Coordination Acoustique/Musique with 3 Loudspeakers

(3) Club Fritz ITA: 5-Institute of Technical Acoustics, RWTH Aachen

(4) Spatial Audio for Domestic Interactive Entertainment: SADIE I, University of York

(5) Spatial Audio for Domestic Interactive Entertainment: SADIE II, Version 1.4 University of York

(6) Technische Hochschule Köln (shorten for THK)

The figures 3.1-3.6 illustrate the geometry used for measurements across the six databases, while Table 3.1 presents an overview of the datasets included in the listening test.

Table 3.1: Overview of Database Selected in Listening Test

Database	Country	Year	Number of Points	Elevation	Azimuth
CF1-IRCAM	France	2004	2016	-45° : 5° : 90°	Δ5°
CF4-IRCAM XMod	France	2007	1944	-40° : 5° : 90°	Δ5°
CF5-ITA	Germany	2009	2016	-80° : 5° : 90°	Δ5° - Δ10°
SADIE I	UK	2018	1550	-90° : 5° : 90°	Δ5°
SADIE II	UK	2018	8802	-90° : 5° : 90°	
THK	Germany	2013	2702	Equidistant spherical Lebedev grids	

3.2.2 Author's Contribution

All experimental procedures presented in this chapter were designed and implemented by the author. The listening test protocol, including stimulus preparation, database selection criteria, experimental structure, and evaluation methodology, was independently developed by the author.[Hongbo]

The HRTF datasets used in this study were obtained from previously published research databases measured with the Neumann KU100 artificial head(see introduction in section 3.2.1). No new HRTF measurements were conducted as part of this work.[Hongbo]

Participant management, data collection, and all statistical analyses were

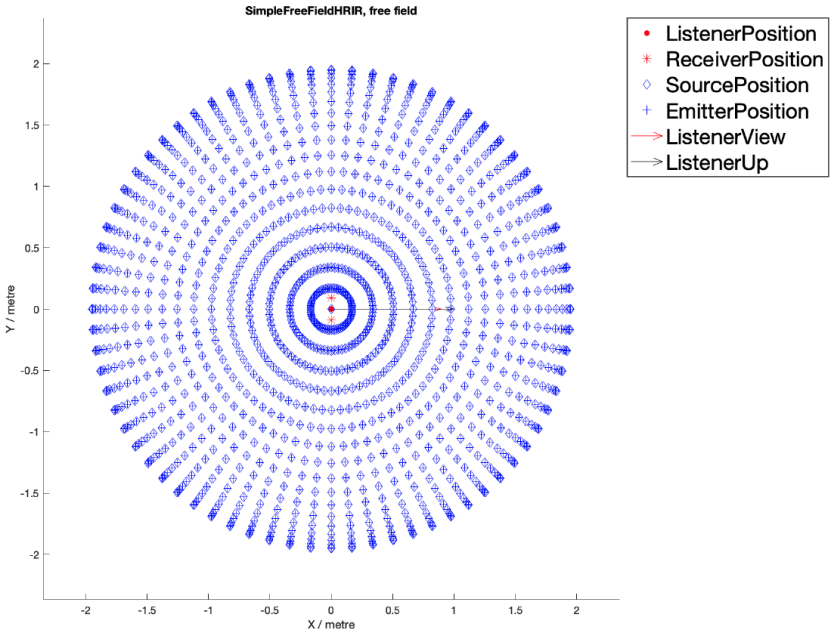

Figure 3.1: CF1 IRCAM

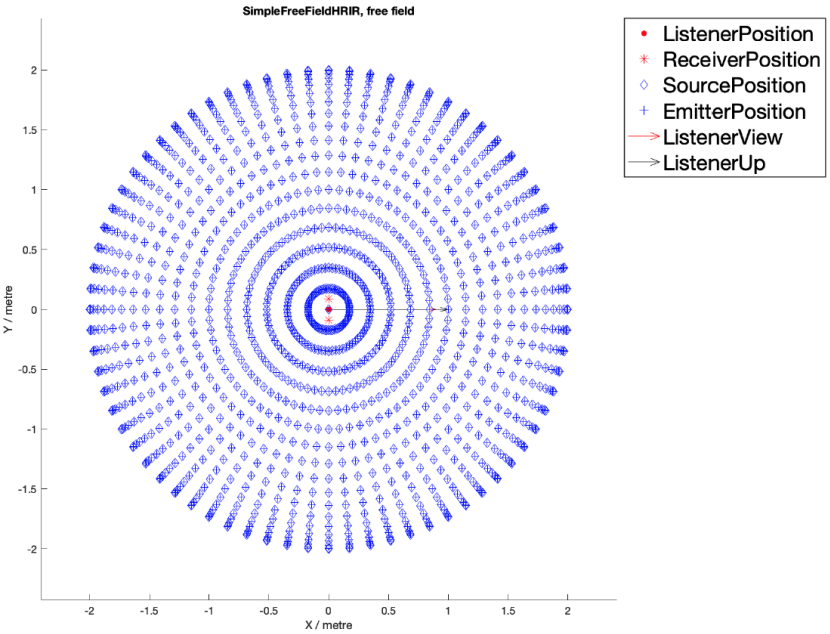

Figure 3.2: CF4 IRCAM Cross Mode

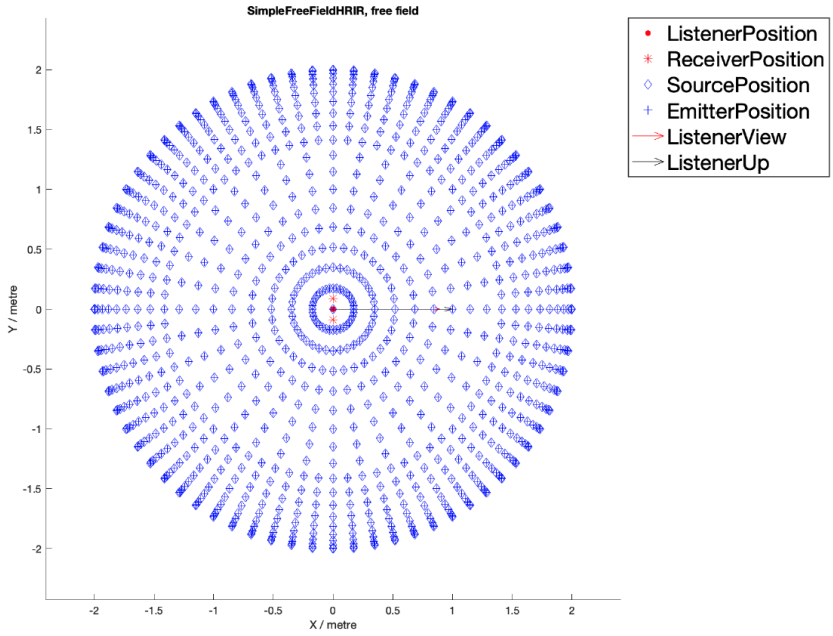

Figure 3.3: CF5 ITA

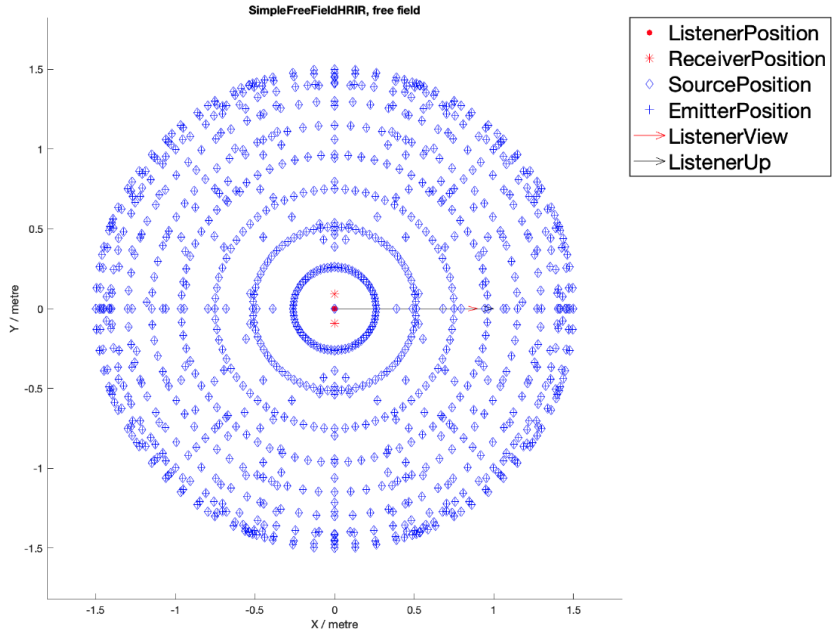

Figure 3.4: SADIE I

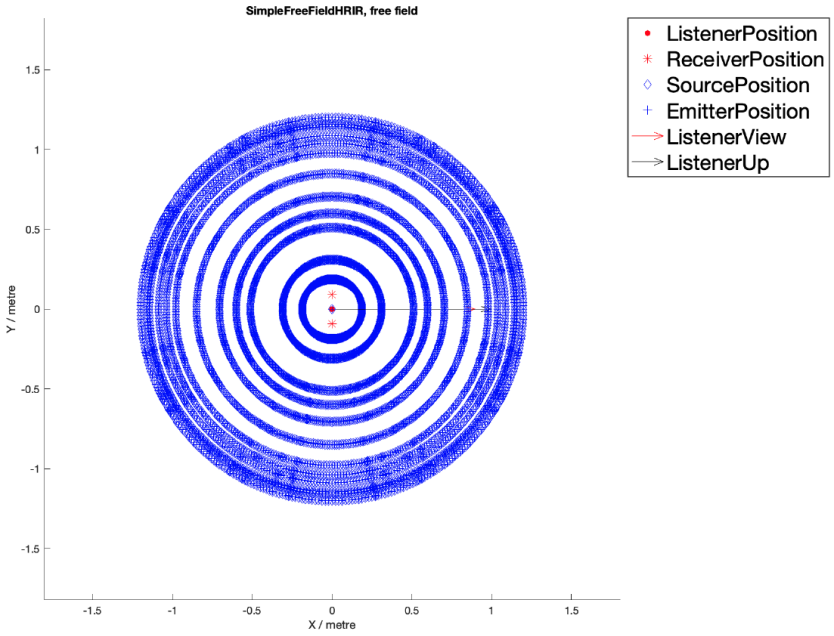

Figure 3.5: SADIE II

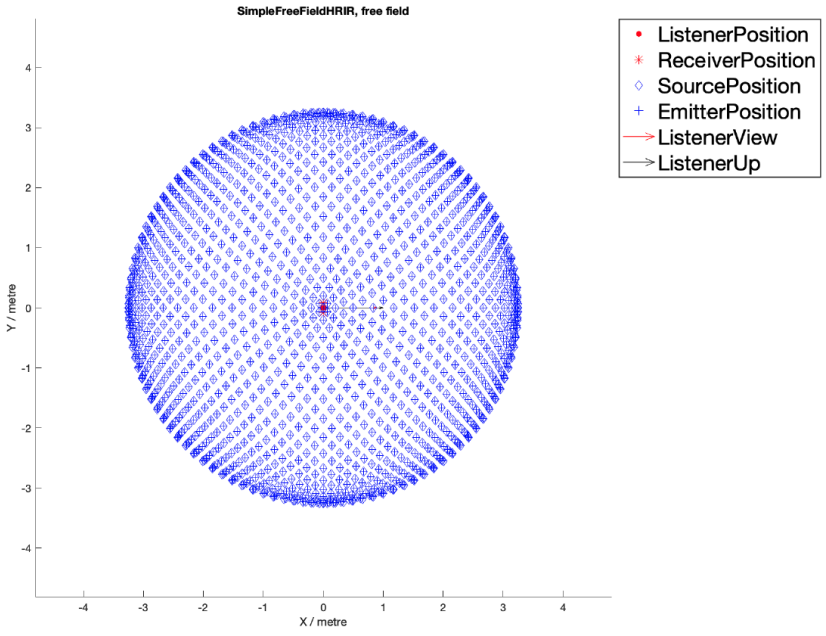

Figure 3.6: THK

performed independently by the author. The analysis relied on standard descriptive and inferential statistical procedures implemented by the author. Hongbo

3.2.3 Test Stimuli

Broadband noise bursts are widely used in localisation accuracy and spatial discrimination experiments due to their flat spectral distribution and clearly defined temporal structure~~Broadband noise bursts are typically used as stimulus for listening tests, in particular for localisation accuracy and spaciousness tests~~ Hongbo [13, 33, 56, 61–65]. However, the primary objective of the present study is not to evaluate fine-grained localisation thresholds, but rather to investigate general perceptual preference and overall listening quality across different generic HRTF databases. In this context, ecological validity and timbral realism are of greater importance than strict spatial discrimination sensitivity [47, 66]. Hongbo

For this reason, natural broadband stimuli were selected instead of synthetic noise bursts. Speech signals have been extensively used in binaural perception research and represent everyday listening conditions [46, 67–69]. ~~This study considers broadband but natural stimuli such as speech signals which have also been used extensively in binaural perception studies [67–69].~~ Hongbo Alongside speech, we have used synthesised piano and a recorded Jazz ensemble which have broader spectral content. The Jazz ensemble is employed as the test stimuli for localisation quality, which is elucidated below:

(1) Piano at both 0 degrees azimuth and elevation, 12 seconds duration.

(2) Male speech at both 0 degrees azimuth and elevation, 12 seconds duration.

(3) Jazz ensemble with all instruments at 0 degrees elevation. Drums are located at 15 degrees azimuth, guitar at 60 degrees azimuth, bass at -30 degrees azimuth and piano at -50 degrees azimuth. See Figure 2 as a schematic diagram

~~The reference audio is the original stereo mix, and the low anchor is a 3.5kHz low pass filtered mono audio.~~[Hongbo]

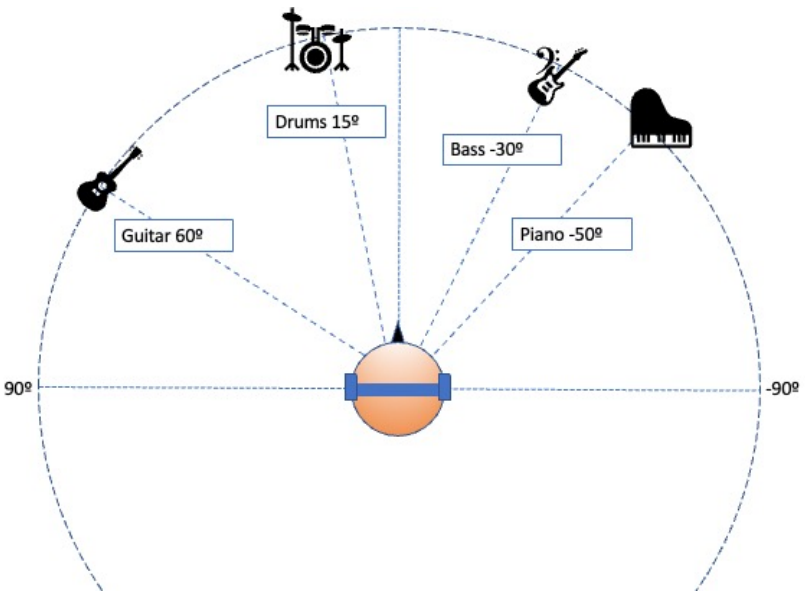

Figure 3.7: Schematic diagram for instruments' position in Jazz ensemble stimuli

3.2.4 Test Paradigm and Rating Scale

Impacted by COVID-19 restrictions, this test was performed completely online using the Web-MUSHRA framework [70], which could be achieved within the web browser without installing any additional software.

The listening test followed the MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) paradigm with methodological adaptations tailored to the objectives of this study. In a conventional MUSHRA test, the reference signal is typically a lossless version of the stimulus and serves as the upper-quality benchmark; test items are rated according to their perceptual proximity to this reference. A low anchor, by contrast, is an intentionally degraded version of the reference signal and defines the lower bound of the rating scale, helping listeners calibrate their scoring behaviour.[Hongbo] In the present study, the original stereo mix was included as a reference condition and the lower anchor is a 3.5 kHz low pass filtered mono audio; however, the reference here was not treated as a traditional upper anchor in the scoring interpretation. This decision was a conscious methodological choice rather than a constraint. The primary objective of the experiment was to evaluate general perceptual preference and timbral naturalness of different HRTF renderings, rather than to assess degradation relative to a technically "ideal" signal. Since the majority of commercially distributed music content is produced and consumed in stereo format, the original stereo mix represents a widely accepted perceptual baseline in everyday listening contexts. [Hongbo]

Consequently, the stereo reference served as a comparative benchmark to contextualise listener judgments, rather than as a strict top-scoring anchor. This approach allowed the evaluation to focus on whether HRTF-based spa-

tialisation maintains perceptual quality and listener preference without introducing undesirable timbral deviations from familiar stereo reproduction. ~~The listening test was based on the MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) listening test paradigm with some customization from the author. Notably, in the normal MUSHRA test protocol, the reference audio is generally a lossless signal, so it is listed as a top scoring option; a higher score of a test stimuli indicates that it is closer to the reference (better sound quality). However, in the presented test, the reference audio in the externalisation, general preference and localisation quality tests was not traditionally used as a reference, and only served as information for comparison.~~ Hongbo

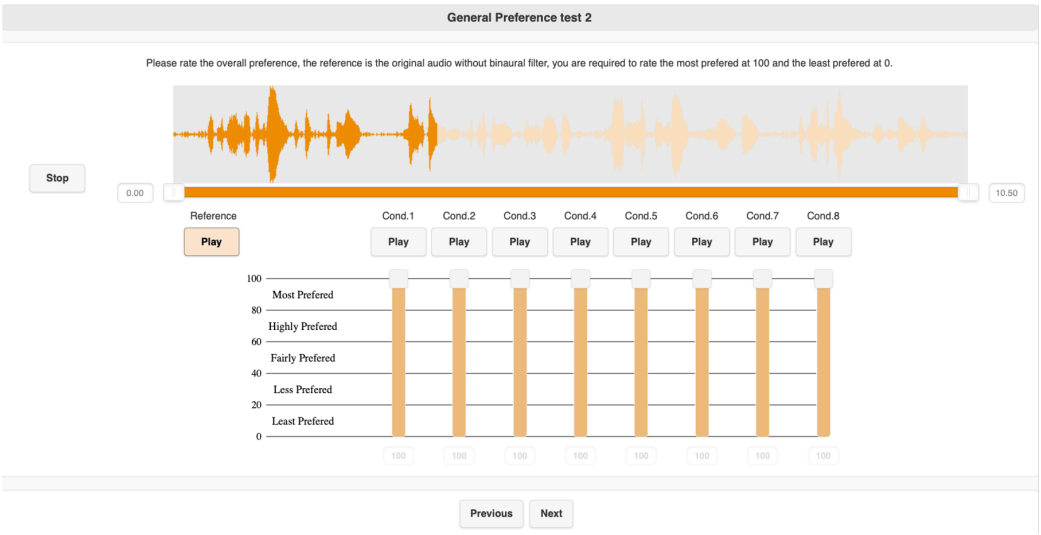

Figure 3.8: A screenshot of the UI from the test website

Thirty participants took part in the study. Participants' ages ranged from 20 to 45 years old(mean age = 26).[Hongbo]

All participants were students or practitioners in audio or music production and had prior experience in listening tests. Participants were recruited through internal university mailing lists and professional networks, and all 30 participants completed the test in full. Each participant received a £10 online shopping voucher as compensation.[Hongbo] The test began with one short training session to familiarise participants with the Web-MUSHRA interface and rating procedure. The main evaluation consisted of three stimulus types: male speech, piano solo, and a jazz ensemble recording. General preference, high-frequency coloration, low-frequency coloration, and externalisation were assessed for all three stimulus types (12 rating tasks in total). In addition, localisation quality was evaluated for the jazz ensemble stimulus (1 additional task), resulting in 13 evaluation tasks per participant.[Hongbo] Stimuli were presented in randomised order to minimise order effects. The sequence of attributes and stimulus types was independently shuffled for each participant. For example, one participant might first evaluate externalisation for the male speech stimulus, whereas another might begin with high-frequency coloration for the piano stimulus. [Hongbo] Participants were asked to complete the test on a personal computer using circumaural headphones. Beyerdynamic DT990 Pro headphones were recommended, although this was not mandatory. The expected completion time was approximately 25–40 minutes.~~All participants were students and practitioners in audio or music production and had prior experience in listening tests. They were asked to complete the test on a personal computer with circumaural headphones. Beyerdynamic DT990 Pro~~

~~headphones were recommended , but this was not mandatory.~~[Hongbo]

Figures 3.9-3.14 show spectrograms of the azimuthal energy at 0 degrees elevation for the 6 HRTFs databases used in this experiment.For clarity and conciseness, only the left-ear responses are presented.[Hongbo]. We see that although they are all measured on a KU100, they exhibit different spectral responses. Accordingly, besides the conventional spatial properties of externalisation, localisation quality and general preference, we also focus on the frequency colouration of the above filters. The rating scales were designed based on established subjective audio evaluation paradigms, particularly the MUSHRA framework [71], which employs continuous 0–100 scales for perceptual quality assessment. The 0–100 scales used for externalisation, general preference, and localisation quality follow this convention to allow fine-grained perceptual discrimination. For spectral colouration assessment, a symmetric 50 to +50 scale was adopted to facilitate direct comparison with the stereo reference (defined as 0), enabling listeners to indicate both positive and negative deviations in perceived bass and treble balance. Similar bipolar scales have been widely used in timbral evaluation studies [47]. They are listed in the Table 3.2~~Rating scales are also listed in table below:~~[Hongbo]

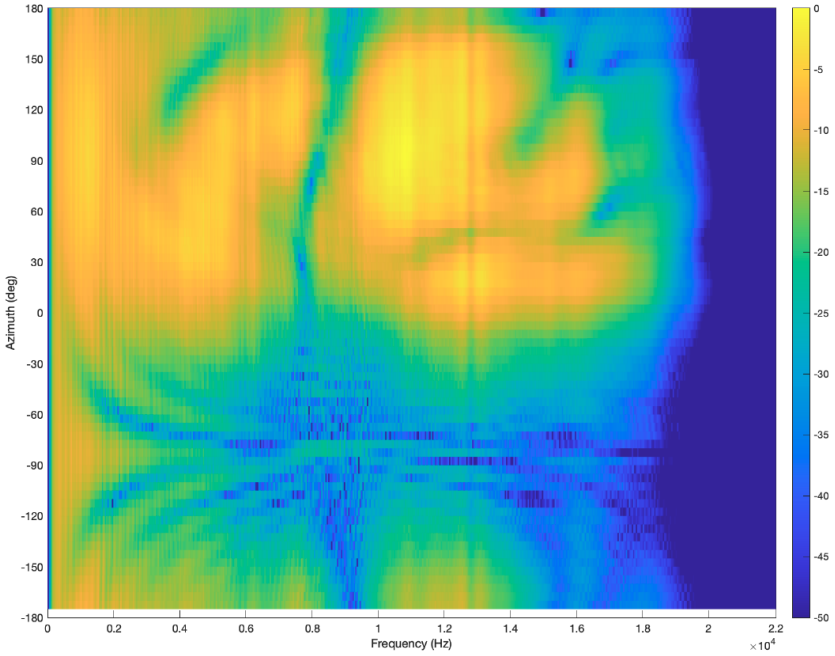

Figure 3.9: CF1 IRCAM

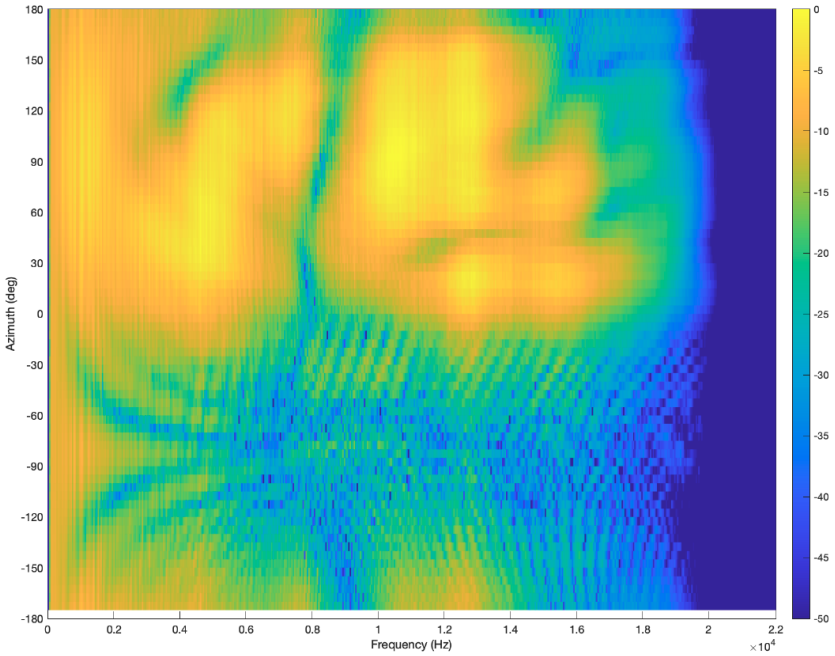

Figure 3.10: CF4 IRCAM Cross Mode

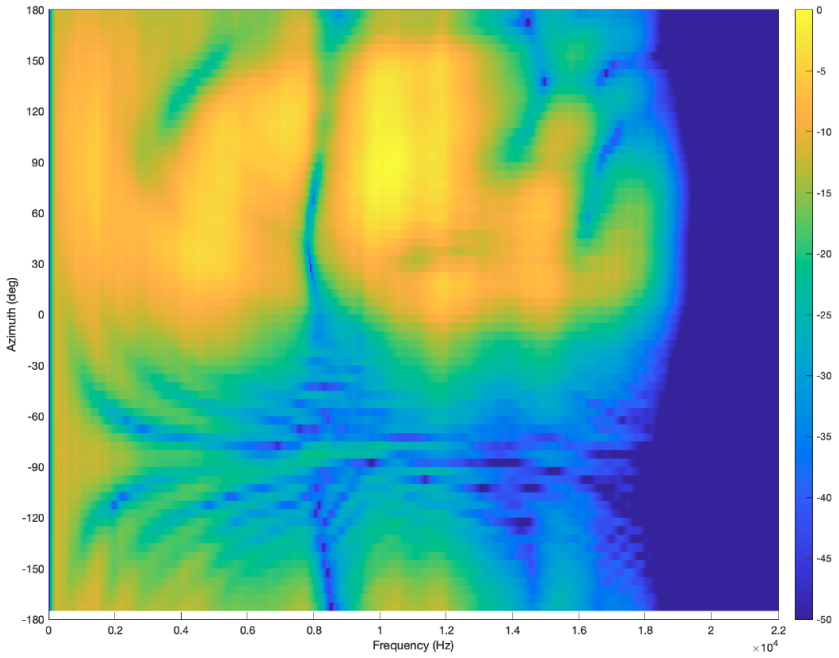

Figure 3.11: CF5 ITA

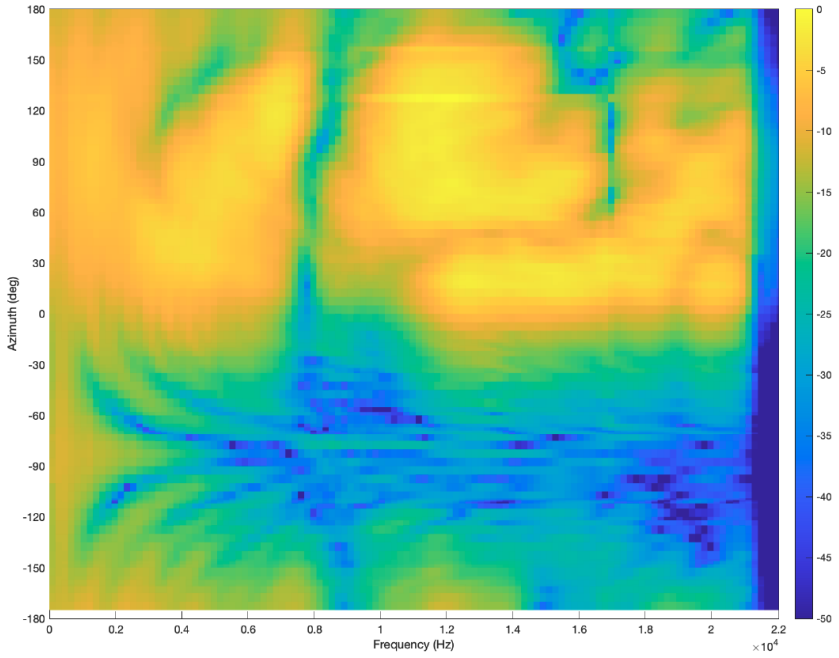

Figure 3.12: SADIE I

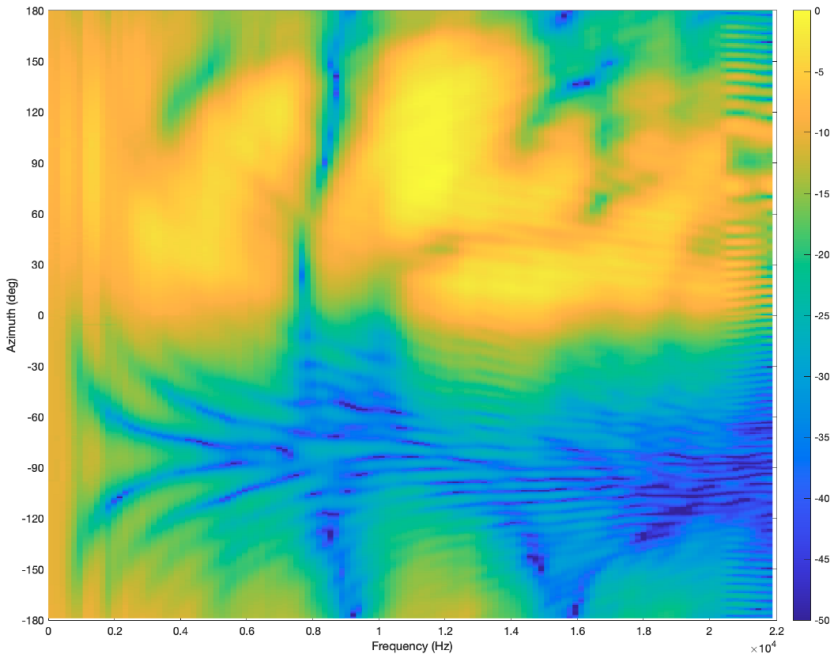

Figure 3.13: SADIE II

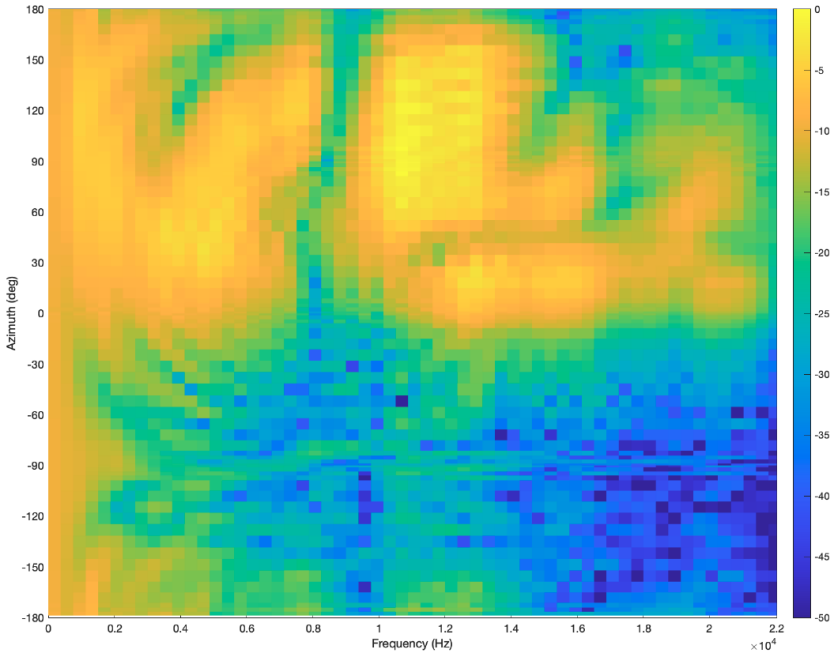

Figure 3.14: THK

Table 3.2: Rating Scales for Stimulus

Attributes	Description	Rating standard
Externalisation	The locatedness of sources to distant points in space	0-100, no actual reference proportional to the level of externalisation
High Frequency Coloration	Level of treble compared with the reference audio	-50 - 50, reference at 0 more treble higher the score
Low Frequency Coloration	Level of bass compared with the reference audio	-50 - 50, reference at 0 more bass higher the score
General Preference	The overall preference of the test audio	0 - 100, no actual reference rate the more preferred audio higher
Localisation Quality	Spatial separation of the sounds and the accuracy of those different sounds locate according to the description	0 - 100, no actual reference rate the more accurate audio higher

3.3 Preliminary observation on experimental results

A preliminary observation is shown in this section.[Hongbo]30 Results were collected with 15 of those using DT-990 headphones. One potential concern is that the use of circumaural headphones, rather than in-ear monitors, may introduce additional acoustic filtering due to interaction with the listener's outer ear. However, the present study was conducted fully online due to COVID-19 restrictions, which limited the possibility of controlling hardware conditions across participants.Circumaural headphones were recommended primarily because most participants were students or professionals in audio and music production, for whom over-ear studio headphones are the standard monitoring tool. The Beyerdynamic DT990 Pro, in particular, is widely used in professional and academic audio contexts and has a well-documented and relatively stable frequency response, making it a reasonable practical reference device. The recommendation was not mandatory, in order to accommodate remote participation and increase accessibility.~~A~~

~~preliminary observation is shown in this section.~~[Hongbo]

3.3.1 Result of frequency coloration test

The three different types of audio signals in this listening test - piano, male speech and jazz ensemble were marked as 1, 2 and 3, respectively. In both frequency coloration tests, we expected the reference audio to be rated as close to zero as possible, with higher scores for the remaining audio indicating more treble and vice versa.

The high-frequency coloration test results revealed that references were all rated approximately to zero which was in line with our expectations. The results also indicated that speech with THK, speech with SADIE I and Jazz ensemble with THK were rated as the three options closest to the reference audio. Furthermore, all audios using ITA filters received the highest scores, indicating that the participants felt that ITA was had more treble, while IRCAM and IRCAM XMod were rated as less treble. To more effectively illustrate the differences between individuals, diverging scales were adopted to present the above results, as presented in Figure 3.15.

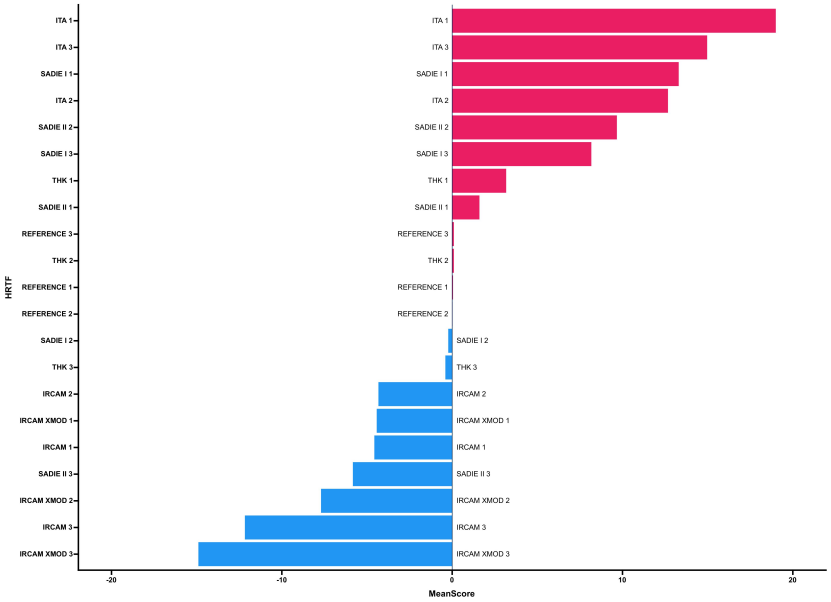

Figure 3.15: Diverging scales of the result from high frequency coloration comparison test

In the low-frequency coloration test, the results indicated that the speech with SADIE II was rated with most bass, following by all audio with SADIE I. On the other side, all audio with the IRCAM filters were rated least bass, as presented in Figure 3.16.6[Hongbo]

Tables3.3 and 3.4~~below~~[Hongbo] present the calculated score for each stimuli, where the total score is the sum of the absolute value of each sub score; the higher the score, the more the coloration. As we could roughly observe from two tables, in the high-frequency coloration test, some HRTF scores were more stable than others. For instance, IRCAM, IRCAM XMod were rated to have less treble, ITA was perceived to have more, and THK tended to be stable and generally close to the reference audio, this result are coincided with the slight attenuation at 4kHz to 8kHz range in the spectrograms for

IRCAM and IRCAM XMod. Interestingly, in SADIE I, the piano solo and jazz ensemble sections scored higher, while the male speech test was close to zero; in SADIE II, the male speech section scored higher, and the jazz ensemble section scored slightly lower. It was therefore revealed that although we used the same HRTF, participants could have perceived significantly different frequency coloration with different audio convolved with this HRTF. It was speculated that the above finding was achieved because the selected stimulus had different frequency bands: the bass part of the jazz ensemble had more bass than the piano solo and male speech, while the hi-hat in the drum kit(also in the Jazz ensemble) had a significantly sharper sound. The above characteristics were potentially enhanced by participants during the test. This would be also true for the low frequency colouration test: in the male speech test the ITA scores were slightly higher, and the SADIE II scores were significantly higher. Results for the~~at~~[Hongbo] low frequency coloration test also coincide with[Hongbo] the fact that low frequency extension existed in the spectrograms for SADIE I and SADIE II. It may be argued that the use of circumaural headphones could introduce additional individual filtering effects due to interaction with each listener's outer ear, potentially influencing perceived spectral balance. While such effects cannot be entirely excluded in a remote listening setup, several factors suggest that this was unlikely to be the primary cause of the observed differences.[Hongbo]

First, the spectral variations discussed above were stimulus-dependent and systematically differed across specific HRTF databases, rather than appearing as random inter-subject deviations. Second, no consistent relationship was observed between headphone model and rating patterns in the

dataset. Finally, all stimuli were evaluated within the same playback condition for each participant, meaning that any individual ear-related filtering would have acted as a constant factor across all comparisons.[Hongbo]

Therefore, although circumaural headphone reproduction may introduce minor individual spectral modifications, the relative differences observed between stimulus types and HRTF databases are more plausibly attributed to interactions between source spectral content and HRTF characteristics rather than outer-ear re-filtering effects alone.[Hongbo]

A statistical analysis is~~was~~[Hongbo] conducted in the next section.

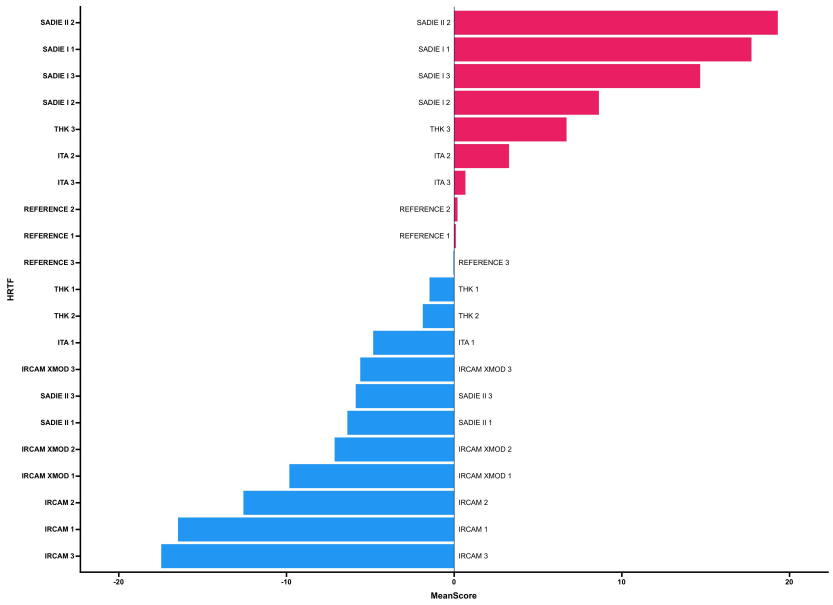

Figure 3.16: Diverging scales of the result from low frequency coloration comparison test

Table 3.3: Calculated score of high frequency coloration comparison test

HRTF	Piano solo	Male speech	Jazz ensemble	Total score
IRCAM	-4.57	-4.33	-12.17	21.97
IRCAM XMod	-4.43	-7.70	-14.90	27.03
ITA	19.00	12.67	14.97	46.64
SADIE I	13.30	-0.23	8.17	21.70
SADIE II	1.60	9.67	-5.83	16.65
THK	3.17	0.10	-0.40	3.67

Table 3.4: Calculated score of low frequency coloration comparison test

HRTF	Piano solo	Male speech	Jazz ensemble	Total score
IRCAM	-16.47	-12.57	-17.47	46.51
IRCAM XMod	-9.83	-7.13	-5.60	22.56
ITA	-4.83	3.27	0.67	8.77
SADIE I	17.73	8.63	14.67	41.03
SADIE II	-6.37	19.30	-5.87	31.54
THK	-1.47	-1.87	6.7	10.04

Where Total Score $= \mid Piano \mid + \mid Speech \mid + \mid Jazz \mid$

3.3.2 Externalisation and localisation quality test

In the externalisation test, THK and IRCAM XMod scored relatively high, whereas all of the HRTFs were rated very close to each other, except for the reference. As expected, since the stereo mix reference is not spatialised, participants could easily perceive a lack of externalisation of the reference compared to the other HRTFs. However, we also tentatively speculated that the reason for the similarity of scores across all HRTFs is that the test was conducted at statically and the stimulus were rendered at 0 degree in the front, subjects could not clearly distinguish between each other. The result is presented in Figure 6 and Table 5.

Table 3.5: Score of Externalisation test

HRTF	Piano solo	Male speech	Jazz ensemble	Total score
IRCAM	58.43	61.17	59.47	179.07
IRCAM XMod	66.90	58.73	59.60	185.23
ITA	60.77	59.03	57.22	177.02
SADIE I	54.37	59.33	56.43	170.13
SADIE II	58.03	52.57	61.07	171.67
THK	63.07	61.20	61.97	186.24
Reference	37.13	40.37	41.40	118.9

Where Total Score $= \mid Piano \mid + \mid Speech \mid + \mid Jazz \mid$ [Hongbo]

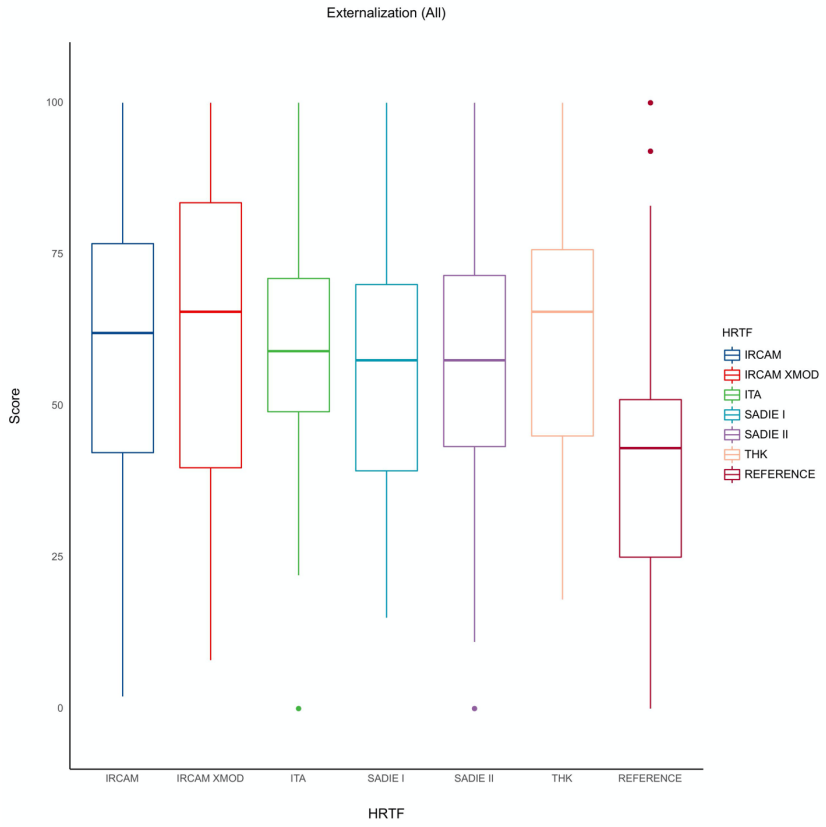

Figure 3.17: Boxplot of the results from the externalisation test

Since the localization test requires multiple sources for comparison, only the jazz ensemble was evaluated for this attribute. As revealed by the results, the Reference and IRCAM XMod scored lowest, whereas other databases were rated close to each other. The relevant results are presented in Figure 3.18~~Figure 7~~[Hongbo] and listed in Table 3.6~~Table 6~~[Hongbo]. Both results for Externalisation and Localisation tests should be further validated through statistical analysis, which is elucidated in the next section.

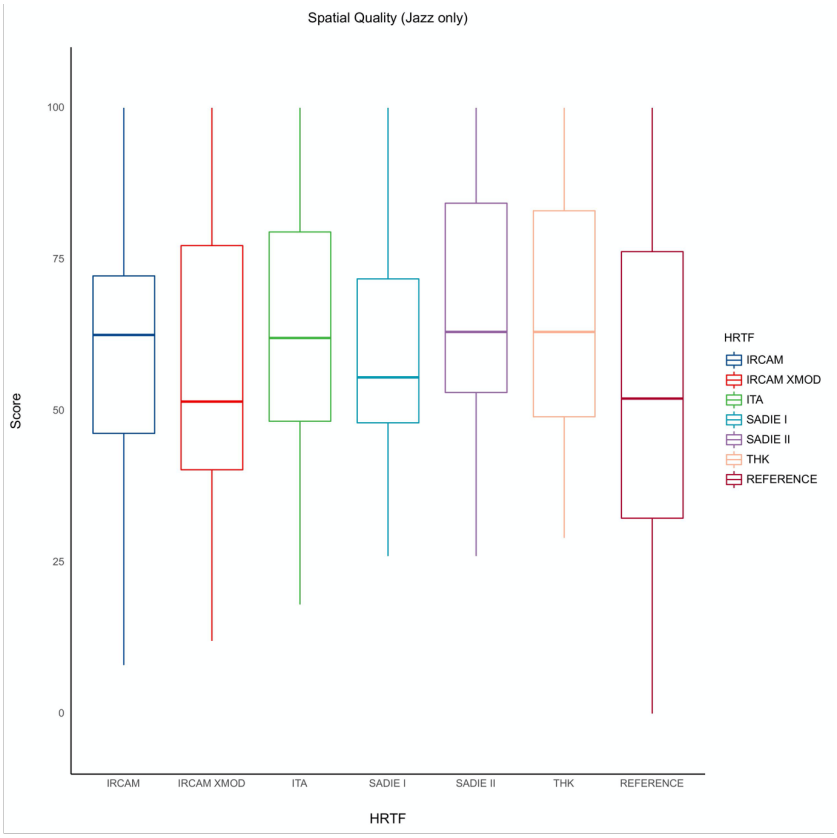

Figure 3.18: Boxplot of the result from localisation quality test

Table 3.6: Score of Localisation quality

HRTF	Jazz ensemble
IRCAM	59.60
IRCAM XMod	56.73
ITA	61.40
SADIE I	59.23
SADIE II	67.17
THK	64.50
Reference	54.20

3.3.3 General preference test

The term general preference indicates how well the audio was perceived and how comfortable the listener was with it. In this test, the majority of participants scored the reference audio (original stereo mix) the highest, and this score was well ahead of the other HRTFs, as presented in Figure 3.19 and Table3.7~~following figures and tables~~[Hongbo] (a further statistical verification is presented in the next section).

Table 3.7: Score of General Preference

HRTF	Piano solo	Male speech	Jazz ensemble	Total
IRCAM	53.17	40.13	41.30	134.60
IRCAM XMod	53.87	40.53	51.10	145.50
ITA	57.37	58.87	56.93	173.17
SADIE I	41.73	60.60	54.37	156.70
SADIE II	57.50	61.70	56.20	175.40
THK	53.0	60.23	61.33	174.56
Reference	73.70	78.40	63.37	215.47

Where Total Score $= \mid Piano \mid + \mid Speech \mid + \mid Jazz \mid$[Hongbo]

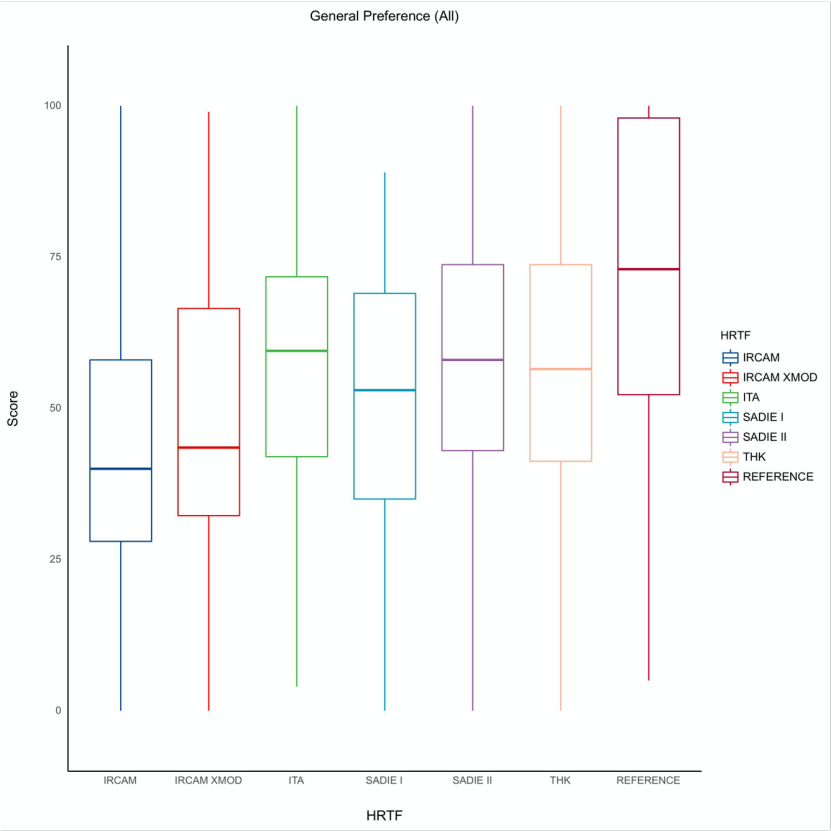

Figure 3.19: Boxplot of the result from general preference test

3.4 Statistical Analysis on the Experimental Results

3.4.1 One-Way ANOVA for Results

In the statistical analysis of the results, we followed the ITU-R BS.1534-3 guidelines and first performed a one-way ANOVA test on the above results, as presented in the following table. [71] For visual clarity, statistically significant results (p 0.05) are highlighted throughout Tables 3.9–3.13. Tables 3.8–3.10 present the ANOVA outcomes, while Tables 3.11–3.13 report the

corresponding Tukey's HSD post-hoc pairwise comparisons.Hongbo

Table 3.8: One-way ANOVA for results

DF=6	GP1	GP2	GP3	EXT1	EXT2	EXT3	LOC	HF1	HF2	HF3	LF1	LF2	LF3
P Value	<0.001	<0.001	0.012	<0.001	0.007	0.003	0.331	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Significant Difference(ANOVA)	Y	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y

P ≤ 0.05 means there is a significant difference DF = degree of freedom Y=Yes N=NO Highlighted cells indicate SDHongbo

In the first ANOVA, except for the localisation quality test where no significant difference was found, there were significant differences in all the other tests. It was~~We~~Hongbo initially speculated that the reason for~~of~~Hongbo the significant differences was~~existed is~~Hongbo because of the the addition of the reference audio. Moreover, since the use of the Beyerdynamic DT-990 headphones might also affect the results, the reference audio was then removed, and the headphone factor was added to conduct a two-way ANOVA.

Table 3.9: Two-way ANOVA for results(reference removed)

DF=5 *	GP1	GP2	GP3	EXT1	EXT2	EXT3	LOC	HF1	HF2	HF3	LF1	LF2	LF3
P Value	0.069	<0.001	0.030	0.246	0.710	0.903	0.445	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Significant Difference(ANOVA)	N	Y	Y	N	N	N	N	Y	Y	Y	Y	Y	Y

P ≤ 0.05 means there is a significant difference DF = degree of freedom Y=Yes N=NO Highlighted cells indicate SDHongbo

Table 3.10: Two-way ANOVA for results (with/without DT 990 headphones)

DF=5 *	GP1	GP2	GP3	EXT1	EXT2	EXT3	LOC	HF1	HF2	HF3	LF1	LF2	LF3
P Value	0.648	0.988	0.853	0.343	0.300	0.794	0.806	0.128	0.478	0.733	0.287	0.100	0.371
Significant Difference(ANOVA)	N	N	N	N	N	N	N	N	N	N	N	N	N

P ≤ 0.05 means there is a significant difference DF = degree of freedom Y=Yes N=NO Highlighted cells indicate SDHongbo

According to the Table 3.9 and 3.10~~table above~~Hongbo, there was no significant difference between the results for~~of~~Hongbo all test audio in the externalisation test, localisation quality test and piano solo for the~~in~~Hongbo general preference tests when the reference audio was removed. The use or not of the~~and the use of~~Hongbo DT-990 headphones ~~or not~~Hongbo had no effect on the results.

3.4.2 Post-hoc Test for Results

In accordance with ITU-R BS.1284 recommendation, for those tests where a significant difference was found, Tukey's Honestly Significant Difference post-hoc test was performed to determine which two conditions had the significant difference.

Table 3.11: Tukey's HSD test results for general preference test

		IRCAM XMOD	ITA	SADIE I	SADIE II	THK
General Preference (speech)	IRCAM	1	0.004	0.001	0	0.001
	IRCAM XMOD		0.005	0.001	0.001	0.002
	ITA			0.999	0.993	1
	SADIE I				1	1
	SADIE II					1
General Preference (Jazz ensemble)	IRCAM	0.595	0.112	0.27	0.147	0.015
	IRCAM XMOD		0.931	0.995	0.96	0.548
	ITA			0.998	1	0.979
	SADIE I				1	0.863
	SADIE II					0.959

Highlighted cells indicate SD[Hongbo]

The values presented in the table above are p-values from Tukey's HSD post hoc test. A value ≤ 0.05 indicates a significant difference (and in the following tables).

It was observed from Table 3.11 that in the general preference test for speech signal, there were significant differences between IRCAM and ITA, SADIE I, SADIE II, THK; there were also significant differences between IRCAM XMod and ITA, SADIE I, SADIE II, THK. This finding revealed that ITA, SADIE I, SADIE II and THK actually scored the same, despite the slight difference in scores. IRCAM and IRCAM CrossMod were approximately scored the same and less preferred by participants.

~~It was observed from Table 11 that in the general preference test for speech signal, there were significant differences between IRCAM and ITA, SADIE I, SADIE II, THK; there were also significant differences between IRCAM~~

~~XMod and ITA, SADIE I, SADIE II, THK. This finding revealed that ITA, SADIE I, SADIE II and THK actually scored the same, despite the slight difference in scores. IRCAM and IRCAM CrossMod were approximately scored the same and less preferred by participants.~~[Hongbo]

Table 3.12: Tukey's HSD test results for high frequency coloration test

		IRCAM XMOD	ITA	SADIE I	SADIE II	THK
High Frequency Coloration(piano solo)	IRCAM	1	0	0.001	0.666	0.449
	IRCAM XMOD		0	0.003	0.796	0.593
	ITA			0.832	0.004	0.012
	SADIE I				0.136	0.268
	SADIE II					0.999
High Frequency Coloration(speech)	IRCAM	0.973	0.002	0.938	0.021	0.915
	IRCAM XMOD		0	0.535	0.002	0.486
	ITA			0.044	0.984	0.006
	SADIE I				0.03	1
	SADIE II					0.002
High Frequency Coloration (Jazz ensemble)	IRCAM	0.985	0.801	0.136	0.022	0.052
	IRCAM XMOD		1	0.033	0.023	0.007
	ITA			0.041	0	0.003
	SADIE I				0.771	0.297
	SADIE II					0.771

Highlighted cells indicate SD[Hongbo]

According to Table 3.12~~Table above~~[Hongbo], there were more significant differences between the HRTFs in the high-frequency coloration test. In the[Hongbo] piano solo section, the original score of THK was leading, and a significant difference was only found between ITA and others; no significant difference was found between ITA and SADIE I. Thus, in piano solo THK, SADIE II, IRCAM and IRCAM XMod performed better, with SADIE I following and ITA in the last place. Likewise, in the[Hongbo]speech section, THK, IRCAM and IRCAM XMod were still leading, followed by SADIE I where ITA and SADIE II were in the last place. In the[Hongbo] Jazz ensemble section, THK, IRCAM, SADIE I and SADIE II ~~were~~[Hongbo] performed better, followed by IRCAM XMod and ITA.

Table 3.13: Tukey's HSD test results for low coloration test

		IRCAM XMOD	ITA	SADIE I	SADIE II	THK
Low Frequency Coloration(piano solo)	IRCAM	0.744	0.161	1	0.299	0.027
	IRCAM XMOD		0.906	0.411	0.98	0.0415
	ITA			0	1	0.982
	SADIE I				0	0.001
	SADIE II					0.913
Low Frequency Coloration(speech)	IRCAM	0.825	0.006	0.259	0	0.158
	IRCAM XMOD		0.183	1	0	0.843
	ITA			0.832	0.005	0.857
	SADIE I				0.161	0.174
	SADIE II					0
Low Frequency Coloration (Jazz ensemble)	IRCAM	0.059	0	0	0.069	0.012
	IRCAM XMOD		0.671	0	1	0.994
	ITA			0.013	0.63	0.935
	SADIE I				0	0
	SADIE II					0.99

Highlighted cells indicate SD[Hongbo]

Table 3.13~~Table above~~[Hongbo] shows that there were wide significant differences in the low-frequency coloration test. As indicated by the results of piano solo section, THK, ITA and SADIE II were leading in the table, followed by IRCAM, IRCAM XMod, and by IRCAM and IRCAM XMod, where SADIE I was in the last place. In the male speech section, THK, IRCAM XMod and ITA had relatively better performance, followed by IRCAM and SADIE I, where SADIE II was rated lowest coloured. In Jazz ensemble section, THK, IRCAM XMod, ITA and SADIE II had the same results, in which IRCAM had less bass, and SADIE I had much stronger low frequency coloration.

3.5 Discussion of Listening Test

In accordance with our results on the general preference evaluation that the original stereo mix was rated far higher than all the other HRTFs, existing studies from Amstrong et al. demonstrated that there was a correlation between the attributes richness and preference [29]. This test verified the results to a certain extent: as the audio convolved with HRTFs led to excessive coloration in some content, while participants preferred the more balanced timbre of the stereo reference. Accordingly, the timbre was found as a vital factor that cannot be neglected when designing binaural filters. It is important to clarify the rationale behind the emphasis on user preference in the preliminary tests. Previous work by Carl et al. (2018) on the SADIE II database demonstrated that listeners often preferred the timbral characteristics of KU100-based HRTFs over their own individualised measurements. [29] This finding suggests that KU100 exhibits strong perceptual robustness and general acceptability across listeners.[Hongbo]

The present study aims to develop a simplified binaural filter derived from existing HRTF measurements. In such a context, the selection of an appropriate base HRTF is crucial. Rather than focusing exclusively on localisation precision, the preliminary evaluation therefore prioritised perceptual acceptability — particularly frequency balance and overall preference — to ensure that the selected database would provide a reliable and well-accepted spectral foundation for subsequent simplification.[Hongbo]

Moreover, according to the results of the present listening test, except for the reference audio, ITA, SADIE I, SADIE II and THK HRTF databases were rated higher in general preference; however, ITA was more coloured at

high frequencies, and SADIE I was more coloured at low frequencies. Thus, SADIE II and THK could be the two most suitable databases for next stage research because of their higher rating in general preference and frequency coloration.

Nevertheless, this study still had several limitations. First, due to the global pandemic, the experiment was conducted entirely online. As a result, several uncontrolled variables were introduced. These include differences in listening environments (e.g., background noise levels, room acoustics), variability in playback equipment (different headphone models and frequency responses), and differences in listening levels across participants. Such factors may have introduced additional variability into the perceptual ratings, particularly in the frequency coloration assessments.[Hongbo]

However, these influences are likely to have contributed primarily to random variation rather than systematic bias, as the presentation order of stimuli was randomized for each participant and no single database was associated with a specific listening setup. Furthermore, a two-way ANOVA analysis (Table 3.10) indicated no statistically significant effect of headphone model (DT990 versus others) on the results, suggesting that equipment variability did not systematically skew the findings.[Hongbo]

Second, since the test was performed under static binaural listening conditions and all stimuli were rendered directly in front of the listener, the perception of externalisation may have been diminished [72–74]. Consequently, the externalisation ratings may not fully reflect performance under dynamic listening conditions. Future work should investigate whether incorporating head tracking alters the perceptual outcomes.~~Nevertheless, this study still~~

had several limitations. The first was that due to the global pandemic, the experiment should be conducted online, so some errors occurred due to the different listening environment of each subject, and the lack of consistency of equipment could also have an effect on the accuracy of the final results. Second, since the test was performed under static binaural listening conditions and all stimulus are rendered in the front of the head, therefore the perception of externalisation is diminished [72–74], therefore it was expected that the externalisation of each HRTF database was not accurately evaluated, and whether the addition of head tracking leads to changes in the results should be explored in the future research.Hongbo

4 Listening Test for Interaural Time Difference Simplification

4.1 Background of the Second Listening Test

Building upon the findings from the initial study, which evaluated several HRTFs from different databases, this subsequent listening test aims to delve deeper into the perceptual differences and preferences identified. In the first test, participants evaluated HRTFs measured on the KU100 dummy head using various stimuli—piano, male speech, and a Jazz ensemble—and assessed attributes such as externalization, frequency coloration, general preference, and spatial quality. While no significant differences were observed in externalization and spatial quality, likely due to the static listening environment, distinct preferences emerged in the frequency coloration tests. Notably, the THK database excelled in both high- and low-frequency coloration tests, SADIE II was favored in the high-frequency coloration, and ITA showed superior performance in the low-frequency coloration. Interestingly, the original stereo mix without binaural filtering received the highest overall preference ratings in the present study. Previous work by Carl et al. (2018) on the SADIE II database similarly demonstrated that KU100-based HRTFs were often preferred over individualised measurements. Together,

these findings suggest that perceptual preference in headphone-based spatial audio may be influenced more strongly by spectral balance and timbral familiarity than by anatomical individualisation alone. [29]~~Interestingly, the original stereo mix without binaural filtering was most preferred, aligning with prior research.~~[Hongbo]

Given the more convenient access to the SADIE II database, further research will be conducted based on this resource to continue exploring these perceptual evaluations.

In order to reduce computational complexity while preserving key binaural cues, the present study adopts a decomposition-based signal model in which each HRTF is represented as the combination of a minimum-phase spectral component and an independently simulated Interaural Time Difference (ITD). The minimum-phase component retains the magnitude characteristics of the original HRTF while removing excess phase, thereby preserving the essential spectral shaping cues responsible for elevation and timbral perception [75]. Under the assumption that the primary interaural temporal information can be approximated as a pure delay, the ITD can be separated from the minimum-phase representation and modelled independently [15]. [Hongbo]This decomposition has been widely adopted in binaural signal processing as a practical simplification strategy, as it enables separate manipulation of temporal and spectral cues while maintaining perceptual plausibility. [9, 21, 75] Accordingly, the present framework forms a perceptually motivated and computationally efficient basis for the proposed simplification approach.[Hongbo]

4.2 Experimental Design

4.2.1 Interaural Time Differences Simulation

The accurate simulation of Interaural Time differences (ITDs) is critical for spatial auditory perception in virtual acoustic environments. Several methodologies have been developed to model ITD characteristics. Analytical models, such as the spherical head approximation, [15, 76] provide computationally efficient solutions by deriving ITD from geometric diffraction principles, though their accuracy diminishes at frequencies above 1.5 kHz due to oversimplified anatomical assumptions. For higher precision, numerical simulations like the Boundary Element Method [77] and Finite-Difference Time-Domain techniques [9]solve wave equations using 3D head scans, albeit at significant computational cost. Modern data-driven approaches leverage HRTF databases [6] to[Hongbo]predicts personalized ITDs from anthropometric features, balancing accuracy and practicality for real-time applications. The choice of method ultimately depends on the trade-off between computational efficiency, frequency bandwidth requirements, and anatomical fidelity. This study aims to explore a more efficient and practical ITD estimation strategy suitable for real-time binaural rendering. The model proposed by Duda et al. [15] was selected because it provides an analytically derived approximation of ITD based on anthropometric parameters, offering a balance between physiological plausibility and computational efficiency. Unlike full wave-based numerical simulations such as the Boundary Element Method (BEM) [77, 78] or Finite-Difference Time-Domain (FDTD) techniques [79], which require detailed 3D geometries and significant computational resources, the Duda

formulation offers a lightweight analytical alternative. This makes it particularly suitable for the proposed simplification framework, where computational cost and scalability are key considerations. ~~This study aims to explore a more efficient and convenient solution; therefore, it will utilize the ITD estimation model proposed by Duda et al. [15]~~[Hongbo]

$$T_{\mathrm{ITD}}(\alpha, \beta, s) = \frac{[0.255x_2(s) + 0.0095x_1(s) + 0.09x_4(s) + C]}{c} \tag{4.1}$$

The equation was expanded to accommodate three dimensions as methodology based on Woodworth's formula, [80] where x_2 represents the maximum head width, x_1 denotes the head height, and x_4 signifies the head depth.

4.2.2 Processing the HRTFs

The minimum phase versions of the HRIRs from the KU100 manikin ~~mannequin~~[Hongbo] in the SADIE II database were initially extracted. Subsequently, equation 4.1 was applied to various directional points of these HRIRs. In the upcoming stages of the experiment, potential participants will be asked to input their corresponding anthropometric data, details of which will be discussed in the subsequent sections.

4.2.3 The Reference HRTFs

Apart from the HRTFs generated by the previously mentioned method, two different HRTFs were used as references for comparison. The first set comprises the KU100 HRTFs from the SADIE II database, and the second set includes personalized HRTFs developed using the rapid method described by Armstrong[Hongbo] [81], where each participant's personalized HRTFs were measured.

4.2.4 Test Paradigm

The second listening test was conducted in a controlled Virtual Reality (VR) environment using Meta Quest II equipment. Fourteen participants, all students or professionals in audio-related fields, took part in this stage of the study. All participants had previously undergone individualized HRTF

measurements as described by Armstrong [81], including detailed anthropometric head measurements as defined by CIPIC Database [6] obtained in an acoustically treated listening laboratory. Head width, height, and depth were measured using precision calipers, and photographic documentation was archived to ensure measurement reliability.~~The listening test was conducted in a Virtual Reality environment using Meta Quest II equipment, developed as reported by [82], with Beyerdynamic DT-990 Pro headphones. The stimulus was broadband noise lasting for 3 seconds.~~[Hongbo]

Stage I: The chosen anthropometric parameters (head width, head height, and head depth) were entered into a MATLAB-based implementation of the Duda ITD estimation model described in Section 4.2.3. The computed interaural delay was then combined with the minimum-phase version of the SADIE II KU100 HRTFs to generate an ITD-modified HRTF for each participant. The resulting HRTF set, together with the original KU100 and the participant's individualized HRTF, was subsequently implemented within a custom Max/MSP patch developed by the author, based on the SPAT5 spatial audio toolkit [83]. This stage lasted approximately 3–5 minutes. ~~Participants were instructed to measure their head width, head height, and head depth (as defined by CIPIC) [6] and input this data into the system. Their ITD-modified HRTF, based on the SADIE II-KU100, would then be generated within 3-5 minutes.~~[Hongbo]

Stage II: Once the ITD-modified HRTF was generated, participants were asked to wear the VR headset and headphones to conduct the sound localization test. Head-tracking data were obtained in real time using the SALTE system developed by Johnston et al. [84]. Audio rendering operated indepen-

dently from the visual display pipeline to ensure synchronization accuracy. The VR visual environment, implemented in Unity and developed according to the framework described in [82], allowed participants to indicate the perceived direction of the broadband noise stimulus (3 seconds duration) within the virtual scene. ~~The stimulus was played back through the headphones, and participants were required to report the location/direction of the virtual sound source using the controller in each trial.~~[Hongbo]

There were 25 spatial directions tested in total. Twenty-four of these were derived from combinations of azimuths (150°, 120°, 0°, 30°, 60°, and 180°) and elevations (30°, 0°, 30°, and 60°), with one additional direction at 0° azimuth and 90° elevation.[Hongbo]

Three HRTF conditions were evaluated: (1) the proposed ITD-modified HRTF (HRTF1), (2) the original KU100 HRTF from the SADIE II database (HRTF2), and (3) the participant's individualized HRTF measured previously (HRTF3).[Hongbo]

For each spatial direction, each HRTF condition was presented twice. Therefore, the total number of trials was:[Hongbo]

25 directions × 3 HRTFs × 2 repetitions = 150 trials.[Hongbo]

All trials were randomized for each participant to minimize order effects. The full test session lasted approximately 50 minutes. Participants received a £20 online shopping voucher as compensation.~~There were 25 different directions, with azimuths of -150, -120, 0, 30, 60, and 180 degrees, and elevations of -30, 0, 30, and 60 degrees, resulting in a total of 150 trials, including repetitions.(one direction 6 times in total including using the reference HRTFs)~~[Hongbo]

4.3 Analyse on the Results

14 Results were collected, A preliminary observation is shown in this section.

Table 4.1: HRTF1 Localization Accuracy: MAE and CR for 25 Target Directions

Target AZI (°)	**Target ELE (°)**	**MAE AZI (°)**	**MAE ELE (°)**	**CR AZI (%)**
-150	-30	16.52	16.97	20.0
-150	0	22.06	0.43	0.0
-150	30	49.30	8.00	40.0
-150	60	41.35	9.74	20.0
-120	-30	9.28	11.73	0.0
-120	0	7.72	13.31	0.0
-120	30	8.93	19.49	0.0
-120	60	21.64	23.68	20.0
0	-30	35.89	13.55	60.0
0	0	0.76	30.95	0.0
0	30	29.56	25.53	40.0
0	60	30.34	59.84	60.0
0	90	36.05	89.90	80.0
30	-30	23.74	12.31	0.0
30	0	23.10	8.78	0.0
30	30	40.49	6.18	20.0
30	60	19.47	23.19	0.0
60	-30	19.34	9.89	0.0
60	0	21.64	0.32	0.0
60	30	30.91	14.71	20.0
60	60	29.21	40.28	40.0
180	-30	36.36	14.12	60.0
180	0	1.54	6.21	0.0

HRTF1

Target AZI (°)	Target ELE (°)	MAE AZI (°)	MAE ELE (°)	CR AZI (%)
180	30	44.05	9.23	40.0
180	60	36.16	45.96	60.0

Note: MAE = Mean Absolute Error; CR = Confusion Rate (error ¿90°).

Table 4.2: HRTF2 Localization Accuracy: MAE and CR for 25 Target Directions

Target AZI (°)	Target ELE (°)	MAE AZI (°)	MAE ELE (°)	CR AZI (%)
-150	-30	13.85	29.11	0.0
-150	0	15.72	0.17	0.0
-150	30	22.43	23.75	20.0
-150	60	49.98	19.74	60.0
-120	-30	12.52	3.55	0.0
-120	0	12.10	0.27	0.0
-120	30	12.61	14.86	0.0
-120	60	17.98	31.87	20.0
0	-30	35.93	26.81	60.0
0	0	35.82	25.94	60.0
0	30	36.04	29.43	60.0
0	60	0.90	25.94	0.0
0	90	0.72	42.46	0.0
30	-30	26.81	9.99	20.0
30	0	28.14	12.88	20.0
30	30	43.23	2.60	40.0
30	60	29.22	31.99	20.0
60	-30	20.79	6.34	0.0
60	0	29.69	23.07	20.0
60	30	31.95	7.34	20.0
60	60	18.95	34.36	20.0

HRTF2

Target AZI (°)	Target ELE (°)	MAE AZI (°)	MAE ELE (°)	CR AZI (%)
180	-30	0.23	3.84	0.0
180	0	0.88	0.89	0.0
180	30	89.85	4.99	80.0
180	60	88.51	20.87	80.0

Note: MAE = Mean Absolute Error; CR = Confusion Rate (error ¿90°).

Table 4.3: HRTF3 Localization Accuracy: MAE and CR for 25 Target Directions

Target AZI (°)	Target ELE (°)	MAE AZI (°)	MAE ELE (°)	CR AZI (%)
-150	-30	32.42	9.18	0.0
-150	0	46.59	0.84	60.0
-150	30	39.46	17.95	40.0
-150	60	34.49	34.99	40.0
-120	-30	21.99	27.60	0.0
-120	0	28.72	0.72	20.0
-120	30	17.90	26.88	0.0
-120	60	30.19	20.92	20.0
0	-30	10.39	8.48	0.0
0	0	40.55	12.38	60.0
0	30	71.61	10.17	80.0
0	60	36.80	23.14	40.0
0	90	23.63	0.58	0.0
30	-30	54.67	29.61	60.0
30	0	43.37	22.87	40.0
30	30	50.24	29.68	60.0
30	60	43.67	45.89	60.0
60	-30	20.84	4.68	0.0
60	0	39.71	2.54	40.0

HRTF3

Target AZI (°)	Target ELE (°)	MAE AZI (°)	MAE ELE (°)	CR AZI (%)
60	30	25.59	29.43	20.0
60	60	34.06	59.66	60.0
180	-30	0.07	4.88	0.0
180	0	0.75	14.32	0.0
180	30	89.97	21.44	80.0
180	60	179.83	28.58	100.0

Note: MAE = Mean Absolute Error; CR = Confusion Rate (error ¿90°).

Table 4.4: Overall Localization Performance Across HRTF Conditions

HRTF Condition	Mean AZI MAE (°)	Mean ELE MAE (°)	Mean CR (%)
HRTF1 (ITD-Simplified)	25.4	24.8	23.2
HRTF2 (KU100)	27.0	20.7	24.0
HRTF3 (Individualised)	40.7	18.9	35.2

Note: MAE = Mean Absolute Error; CR = Confusion Rate (error ¿ 90°).

To provide a clearer cross-model comparison, overall mean localization metrics were calculated across all 25 spatial directions and demonstrated in Table 4.4.[Hongbo]

For azimuth accuracy, HRTF1 (ITD-simplified) achieved the lowest mean absolute error (25.4°), slightly outperforming HRTF2 (KU100, 27.0°), while HRTF3 (individualized) showed substantially higher error (40.7°). This indicates that the simplified ITD-based modification did not degrade horizontal localization performance and, in this dataset, demonstrated improved overall stability compared to individualized HRTFs.[Hongbo]

In terms of elevation accuracy, HRTF3 (18.9°) and HRTF2 (20.7°) showed marginally better performance than HRTF1 (24.8°). However, the differences remained within a comparable range, and all models exhibited elevated errors at extreme elevation (90°), suggesting inherent limitations in vertical spectral cue reconstruction under the present VR conditions.[Hongbo]

Regarding front–back confusion, HRTF1 (23.2%) and HRTF2 (24.0%) showed comparable mean confusion rates, whereas HRTF3 exhibited a notably higher average rate (35.2%), particularly in frontal and elevated regions. This suggests that individualized HRTFs did not consistently reduce perceptual reversals and, in certain spatial quadrants, introduced greater instability.

Hongbo

Overall, the results demonstrate that the ITD-simplified HRTF maintains functional localization performance comparable to the full KU100 reference and, in horizontal localization, may provide improved robustness relative to individualized measurements. These findings support the feasibility of the proposed simplification framework as a computationally efficient alternative without substantial perceptual compromise.Hongbo

Beyond the overall performance metrics, model-specific spatial patterns were also observed. The comparative analysis reveals clear variations in spatial localization capabilities across HRTF models, highlighting model-specific strengths and limitations. While certain implementations demonstrated stronger performance in posterior localization or median-plane precision, all models exhibited elevated errors at extreme elevation (notably 90°) and persistent front–back confusions. Hongbo

Lateral asymmetries and hemispheric inversion errors further suggest inherent limitations in current spectral cue differentiation mechanisms under VR listening conditions. These findings emphasize the importance of improving median-plane spectral modeling and refining anthropometric adaptation strategies to address systematic localization variability across spatial quadrants.~~Where HRTF1,2 and 3 refers to HRTF with ITD-Simplified method, KU100 HRTF and Personlised HRTF respectively.The comparative analysis demonstrates clear variations in spatial localization capabilities across HRTF models, revealing model-specific strengths and limitations. While certain implementations excel in particular spatial regions (posterior localization vs. frontal precision), all models exhibit inherent challenges in extreme~~

~~elevation handling and persistent front-back confusions. Lateral asymmetries and hemispheric inversion errors further highlight fundamental limitations in current spectral cue differentiation. The findings emphasize the necessity for targeted optimization of median plane spectral features, adaptive error correction mechanisms, and personalized anthropometric adaptations to address systematic localization failures across spatial quadrants.~~[Hongbo]

5 Conclusions and future work

5.1 Conclusions

This study examined a simplified approach to binaural filtering through two listening tests under controlled experimental conditions. The first test employed a MUSHRA-based subjective evaluation to compare perceived timbral quality and overall preference across binaural renderings, whereas the second test investigated directional localisation performance in a VR environment with head tracking.[Hongbo]

Across the subjective evaluation, the simplified approach elicited comparable or, in some cases, more favourable perceptual responses than the selected generalised HRTFs under the tested conditions, indicating that cue simplification did not necessarily lead to unacceptable timbral degradation. In the VR localisation test, the simplified approach did not achieve the same level of spatial localisation accuracy as fully personalised HRTFs across all directions; however, its horizontal localisation performance was comparable to the KU100 reference and its overall confusion rate remained within a similar range to the KU100 condition. Taken together, these results suggest that the examined approach is unlikely to replace fully personalised HRTFs in applications requiring high localisation precision, but it may represent a viable

alternative in scenarios where rapid and cost-effective HRTF generation is prioritised, subject to the constraints of the experimental design.[Hongbo]

With respect to the experimental hypotheses proposed in Chapter 1, the findings provide partial empirical support.[Hongbo]

Hypothesis 1, which proposed that generalized HRTFs retain value beyond strict localization accuracy, is supported by the subjective evaluation results. Certain generalized HRTFs demonstrated competitive perceptual performance in timbral quality and listener preference, indicating their continued relevance in practical applications.[Hongbo]

Hypothesis 2, which proposed that modifying generalized HRTFs using individual physiological parameters could achieve satisfactory outcomes, is partially supported. The ITD-simplified approach maintained horizontal localization accuracy and confusion rates comparable to the KU100 reference, suggesting functional viability. However, the simplified model did not consistently outperform fully individualized HRTFs across all spatial dimensions, particularly in extreme elevation conditions. Therefore, while promising, the approach cannot yet be considered a full substitute for comprehensive personalization.[Hongbo]

Overall, the study demonstrates that simplified manipulation of specific binaural cues—particularly ITD components—can preserve listener-relevant aspects of spatial perception and subjective quality in controlled listening tests. Nevertheless, the extent to which these findings generalise beyond the present participant sample, HRTF database, and experimental configurations remains limited and warrants further investigation.~~This study introduced a novel, simplified method for generating Head-Related Transfer Functions~~

~~(HRTFs) through two listening tests. Although the spatial localization accuracy of these HRTFs does not match that of standard Personalized HRTFs, they have demonstrated superior performance compared to the selected Generalized HRTFs. This finding suggests that while the new method may not fully replace the need for more complex personalized HRTFs in applications demanding high precision in sound localization, it offers a viable alternative for scenarios where rapid and cost-effective HRTF generation is prioritized. Furthermore, the enhanced performance over Generalized HRTFs indicates that the simplified approach retains a significant degree of the auditory spatial cues essential for effective sound localization, thus bridging the gap between high customization and practicality in everyday applications.~~[Hongbo]

5.2 Future work

Despite the promising outcomes observed under the tested conditions, this study acknowledges several limitations. In particular, the scope of ITD simplification explored in this work was constrained, and the generalisability of the results is influenced by the characteristics of the selected HRTF database. Future research could explore more detailed personalisation of ILD and spectral adjustments, with the aim of addressing a broader range of auditory perceptual factors. Such developments may contribute to further refinement of spatial audio customisation, including VR-based applications; however, the extent and robustness of these effects would require systematic evaluation across a wider range of listeners, datasets, and experimental conditions.~~Despite promising outcomes, the study acknowledges limitations in the scope of ITD simplification and the generalizability of results due~~

~~to the specificities of the chosen database. Future research will extend to a more detailed personalization of ILD and spectral adjustments, aiming to cover a broader range of auditory perceptions. The potential for these advancements to refine VR audio customization substantially enhances user experience while addressing individual auditory needs.~~[Hongbo]

Appendices

A

Ethics Forms

THE UNIVERSITY of York

Application Form for Physical Sciences Ethics Committee Approval

Advice for applicants on completing the form

Please ensure that the information provided is:

- *Accurate and concise*
- *Clear and simple and easily understood by a lay person*
- *Free of jargon, technical terms and abbreviations*

Further advice and information can be obtained from your departmental representative on the PSEC and at: http://www.york.ac.uk/admin/aso/ethics/cttee.htm

Please return completed (typed) form to your departmental representative via email to:

elec-ethics@york.ac.uk

Title of project:

Elevation Perception under dynamic condition in Binaural Rendering with Simplified HRTF.

SECTION 1 DETAILS OF APPLICANTS

Details of principal investigator (name, appointment and qualifications)

Hongbo Hu. BEng.
PhD Student

Names, appointments and qualifications of additional investigators *(student applicants should include their project supervisor(s) here)*

Gavin Kearney, PhD.
Associate Professor in Audio and Music Technology.

Location(s) of project

University of York, Audio Lab, Genesis 6, Science Park. Experiments will be undertaken in participants' homes.

SECTION 2 FUNDERS

What is the funding source(s) for the project?

This is a project from a self-funding PhD student, no external funding involved.

Please answer the following:

(i) Does the express and direct aim of the research or other activity raise ethical issues?

YES ☐ NO ✓

(ii) Is there any obvious or inevitable adaptation of research findings to ethically questionable aims?

YES ☐ NO ✓

(iii) Is the work being funded by organisations tainted by ethically questionable activities?

YES ☐ NO ✓

(iv) Are there any restrictions on academic freedoms – notably, to adapt and withdraw from ongoing research, and to publish findings?

YES ☐ NO ✓

If you answered **Yes** to any of the above, please give details below:

SECTION 3 DETAILS OF PROJECT OR OTHER ACTIVITY

Aims (100 words max)

This project aims to find out whether participants are able to perceive height information when using a generic HRTF(HRTF for KU100 from selected database) with ITD manipulation rather than using a completed personlised HRTF.

Background (250 words max)

Head Related Transfer Functions (HRTFs) are key to delivering effective and dynamic immersive audio. It is shown throughout the literature that using personalised HRTFs improves spatial audio perception and localisation in virtual auditory displays significantly. However, HRTF measurement is a complex, time consuming process that has to be carried out in an anechoic chamber using calibrated equipment under well-controlled conditions. Alternatively generalized HRTFs can be employed for correct auralisation and perception, but these can suffer from timbral issues and localization errors. This research aims to manipulate generalized HRTFs so that they can be more tailored to the individual.

The previous phase of the experiment revealed that participants would have a greater preference for a specific HRTF set(THK and SADIE II) and based on these results, we will use that HRTF set (THK) for the next phase of the study. It has been suggested that the non-personalised HRTF may confound or defeat the elevation perception in a particular frequency band($f > 1.5kHz$) when subjects perform localisation tests in median plane, however those results need to be explored in more depth in more experiments. Therefore, in this experiment, participants will be asked to use a generic HRTF synthesised by an innovative ITD Optimisation technique to verify whether they can accurately perceive height information or not.

Brief outline of project/activity (250 words max)

Due to Covid-19 restrictions, the listening test will be conducted remotely in participant's homes.

Consent and demographics will be collected via Qualtrics. Participants are then asked to download a MATLAB application (app) which guides them through the experiment.

The listening test will be divided into two parts:

Part 1 (20 mins):
Participants are required to measure their head size and report the value in the MATLAB app, then they will be asked to distinguish whether there is a spatial difference between 2 presented pink noise signal. When the test is finished, participants are required to send over the test result to the investigator, then those results will be used for ITD Optimisation on the KU100'S HRTF from THK database.

Part 2 (60 mins):

This part of the test required participants to wear an Oculus Quest 2 device for the test. Parameters and information about the Oculus Quest 2 can be found in [1]. Participants will be instructed to install the listening test software on their personal computer to connect their Oculus Quest 2 device, after which they would run the software for the listening test. Some reference animated screens will appear on the Oculus Quest 2 device and participants will be asked to respond using the corresponding buttons on the joystick to indicate the location of the test stimuli they hear. The results of the listening tests will eventually be further analysed for the next stage of the study.

[1] https://www.oculus.com/quest-2/?locale=en_GB&utm_source=gg&utm_medium=a_ps&utm_campaign=11138178459&utm_term=oculus+quest+2&utm_content=516430185834&utm_parent=quest2&utm_ad=110857929242&utm_location=9046887&utm_location2&utm_placement=kwd-563298318914&utm_adposition&utm_device=c&utm_matchype=e&utm_feed&gclsrc=aw.ds&&gclid=Cj0KCQjwktKFBhCkARIsAJeDT0gm2gN-jQUeAFPOyKYJ3jGmc0YLjMU1pIJqJb0wSNiv1KRjvV2ZeeEaAsh0EALw_wcB

Study design *(if relevant – e.g. randomised control trial; laboratory-based)*

The test will be carried in VR device (Oculus Quest 2) and based on the SALTE localization test framework. Synthesised HRTF will be generated from part 1's results and will be convolved with several different test signals, namely an instrument solo, a male voice, a jazz ensemble, and a pink noise signal in different frequency bands. Each set of test stimuli will last 10-12 seconds.

If the study involves participants, how many will be recruited?
Approximate 10 - 15 People due to the restriction of hardware.

If applicable, what is the statistical power of the study, i.e. what is the justification for the number of participants needed?

SECTION 4 RECRUITMENT OF PARTICIPANTS

How will the participants be recruited?

People from Audio Lab, and those experienced listener who has the Oculus Quest 2 headset.

What are the inclusion/exclusion criteria?

Exclusion:
People who:

1. Have hearing, vision or movement disabilities.
2. Are not comfortable with audio noise
3. Have motion sickness when wearing VR headset

Will participants be paid reimbursement of expenses?	YES [ ]	NO [√]
Will participants be paid?	YES [√]	NO [ ]
If yes, please obtain signed agreement		
Will any of the participants be students?	YES [√]	NO [ ]

SECTION 5 DATA STORAGE AND TRANSMISSION

If the research will involve storing personal data, including sensitive data, on any of the following please indicate so and provide further details (answers only required if *personal* data is to be stored).

Manual files	√
University computers	√
Home or other personal computers	√
Laptop computers, tablets	√
Website	√

Please explain the measures in place to ensure data confidentiality, including whether encryption or other methods of anonymisation will be used.

Computers will be password protected. Stored participant data will be anonymised. No paper documentation will exist for this study. Data will be stored in encrypted files/folders.

Please detail who will have access to the data generated by the study.

Investigators only

Please detail who will have control of and act as custodian for, data generated by the study.

Investigators only

Please explain where, and by whom, data will be analysed.

By the investigators at the University of York Audio Lab and at the investigator's homes.

Please give details of data storage arrangements, including where data will be stored, how long for, and in what form.

Data will be stored on investigators' personal computers and at the University of York

Audio lab storage in encrypted files. Data will only be stored on the investigators' laptops until the end of the PhD study.

SECTION 6 CONSENT

Is written consent to be obtained? YES √ NO

If yes, please attach a copy of the information for participants

If no, please justify

Will any of the participants be from one of the following vulnerable groups?

Children under 18	YES		NO	√
People with learning difficulties	YES		NO	√
People who are unconscious or severely ill	YES		NO	√
People with mental illness	YES		NO	√
NHS patients	YES		NO	√
Other vulnerable groups (if 'yes', please give details)	YES		NO	√

If so, what special arrangements have been made for getting consent?

Consent Form will be provided to participants.

SECTION 7 DETAILS OF INTERVENTIONS

Indicate whether the study involves procedures which:

Involve taking bodily samples	YES		NO	√
Are physically invasive	YES		NO	√
Are designed to be challenging/disturbing (physically or psychologically)	YES		NO	√

If so, please list those procedures to which participants will be exposed:

List any potential hazards:

There is a small risk of noise exposure, but this will be mitigated by guiding participants through a set up procedure to set their computer to a safe volume.
There is a small risk of fatigue, which will be mitigated by reminding participants to take regular breaks and by keeping the length of the experiment short.
There is a small risk of dizziness when wearing the VR device, which will be mitigated by reminding participants to take regular breaks and by keeping the length of the experiment short.

List any discomfort or distress:

None.

What steps will be taken to safeguard

(i) the confidentiality of information

Data will not be stored with participants names but using a unique ID.

(ii) the specimens themselves?

N/A

What particular ethical problems or considerations are raised by the proposed study?

None.

What do you anticipate will be the output from the study? ***Tick those that apply:***

Peer-reviewed publications	√
Non-peer-reviewed publications	√
Reports for sponsor	
Confidential reports	√
Presentation at meetings	√
Press releases	
Student project	

Is there a secrecy clause to the research? YES [ ] NO [√]
If yes, please give details below

SECTION 8 SIGNATURES

The information in this form is accurate to best of my knowledge and belief and I take full responsibility for it.

I agree to advise of any adverse or unexpected events that may occur during this project, to seek approval for any significant protocol amendments and to provide interim and final reports. I also agree to advise the Ethics Committee if the study is withdrawn or not completed.

Signature of Investigator(s): ……Hang bo H…….

……

Date: ……27/May/2021……

Responsibilities of the Principal Researcher following approval

- If changes to procedures are proposed, please notify the Ethics Committee
- Report promptly any adverse events involving risk to participants

University of York

Online-based listening test for the subjective perceptual evaluation of different Binaural Filters for KU100 under static conditions.

31st Aug 2020 (v1.0)

PROJECT INFORMATION SHEET

Thank you for agreeing to participate in this study. We are working to evaluate the perceived timbral and spatial differences between different binaural databases for headphone based 3D audio. The results of the test will be used for further simplification of binaural filters.

We will be querying your response to a number of different audio stimuli with different filters. The test is a Web-MUSHRA based listening test GUI, released online.

Please use the link http://audiolab.york.ac.uk/listening-tests/hh973/ to conduct the test.

Once you begin the test, follow the instructions on each sub test and please ensure you are using headphones and the L and R channels are on your ears correctly.

During the test, you will be asked to evaluate: Externalisation, High frequency coloration, Low frequency coloration and general preference between different stimuli and the reference.

You are free to listen to each sound sample as many times as you like, and you can go back to previous samples you have responded.

Once you have completed your tests, please fill out the information form .

If you have any questions about the process, or how your data will be used, please don't hesitate to ask. Further, if you wish to withdraw from the study you are free to do so at any point. Please note however that any data collected up to the point of withdrawal may still be used. Contact details are provided opposite for your convenience.

Primary Investigator: Hongbo Hu
Affiliation: Audio Lab, Dept. of Electronic Engineering, University of York
Email: hh973@york.ac.uk

Supervisor: Gavin Kearney
Affiliation: Audio Lab, Dept. of Electronic Engineering, University of York
Email: gavin.kearney@york.ac.uk

Many thanks,
Hongbo Hu and Gavin Kearney

THE UNIVERSITY of York

Application Form for Physical Sciences Ethics Committee Approval

Advice for applicants on completing the form

Please ensure that the information provided is:

- *Accurate and concise*
- *Clear and simple and easily understood by a lay person*
- *Free of jargon, technical terms and abbreviations*

Further advice and information can be obtained from your departmental representative on the PSEC and at: http://www.york.ac.uk/admin/aso/ethics/cttee.htm

Please return completed (typed) form to your departmental representative via email to:

elec-ethics@york.ac.uk

Title of project:

Elevation Perception under dynamic condition in Binaural Rendering with Simplified HRTF.

SECTION 1 DETAILS OF APPLICANTS

Details of principal investigator (name, appointment and qualifications)

Hongbo Hu. BEng.
PhD Student

Names, appointments and qualifications of additional investigators *(student applicants should include their project supervisor(s) here)*

Gavin Kearney, PhD.
Associate Professor in Audio and Music Technology.

7th May 2015

Location(s) of project

University of York, Audio Lab, Genesis 6, Science Park. Experiments will be undertaken in participants' homes.

SECTION 2 FUNDERS

What is the funding source(s) for the project?

This is a project from a self-funding PhD student, no external funding involved.

Please answer the following:

(i) Does the express and direct aim of the research or other activity raise ethical issues?

YES ☐ NO ☑

(ii) Is there any obvious or inevitable adaptation of research findings to ethically questionable aims?

YES ☐ NO ☑

(iii) Is the work being funded by organisations tainted by ethically questionable activities?

YES ☐ NO ☑

(iv) Are there any restrictions on academic freedoms – notably, to adapt and withdraw from ongoing research, and to publish findings?

YES ☐ NO ☑

If you answered **Yes** to any of the above, please give details below:

SECTION 3 DETAILS OF PROJECT OR OTHER ACTIVITY

Aims (100 words max)

This project is related to 3D audio perception over headphones using binaural filtering. It aims to find out whether participants are able to perceive the relative height of sound sources when using a generic binaural filter with and without personalised binaural cue manipulation.

Background (250 words max)

Binaural filters known as Head Related Transfer Functions (HRTFs) describe how a sound is filtered by the head, outer ear and torso as it reaches the ear canal. They are key to delivering effective and dynamic immersive audio over headphones. It is shown throughout the literature that using personalised HRTFs improves spatial audio perception and localisation in virtual auditory displays significantly [1]. However, HRTF measurement is a complex, time consuming process that has to be carried out in an anechoic chamber using calibrated equipment under well-controlled conditions. Alternatively generalized HRTFs can be employed for correct auralisation and perception, but these can suffer from timbral issues and localization errors [2]. This research aims to manipulate generalized HRTFs so that they can be more tailored to the individual.

It has been suggested that the non-personalised HRTF may confound or defeat the elevation perception in a particular frequency band (f > 1.5kHz) when subjects perform localisation tests in the median plane, however such results need to be explored in more depth [3]. Therefore, in this experiment, participants will be asked to use a generic HRTF synthesised by an innovative optimisation technique to verify whether they can accurately perceive height information or not.

Reference

[1] Algazi,V.R.,Duda,R.O.,Thompson,D.M.,and Avendano, C., "The CIPIC HRTF database," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pp. 99—102, 2001

[2] Mendonc a, C., Campos, G., Dias, P., Vieira, J., Ferreira, J. P., & Santos, J. A. (2012). On the improvement of localization accuracy with non- individualized HRTF-based sounds. *In Audio Engineering Society Convention 129*. San Francisco, CA. Audio Engineering Society.

Hebrank J, and Wright D: "Spectral cues used in the localization of sound sources on the median plane," J Acoust Soc Am, 56 (6):1829–1834, 1974.

Brief outline of project/activity (250 words max)

Consent and demographics will be collected via Qualtrics. Participants are then asked to download a MATLAB application (app) which guides them through the experiment.

The listening test will be divided into two parts:

Part 1 (20 mins): Calibration.

Participants are required to measure their head size and report the value in the MATLAB app. Wearing headphones, they will then be asked to distinguish whether there is a perceived spatial shift between 2 presented pink noise signals, where 1 of the signals is controlled via a slider which adjusts binaural cues. When the test is finished, participants are required to send over the test result to the investigator, and these results are used to derive a personalised test dataset for the individual.

Part 2 (60 mins): Localisation.

This part of the test required participants to wear an Oculus Quest 2 device for the test. Parameters and information about the Oculus Quest 2 can be found in [1]. Participants will be instructed to install the listening test software on their personal computer to connect their Oculus Quest 2 device, after which they would run the software for the listening test. Reference animated screens will appear on the Oculus Quest 2 device and participants will be asked to respond using the corresponding buttons on the joystick to indicate the location of the test stimuli they hear.

Test stimuli will be
A solo piano,
A male voice,
A jazz ensemble,
Pink noise signal in different frequency bands.

The stimuli will be presented binaurally using filters generated from Part 1 of the test. Each set of test stimuli will last 10-12 seconds.

[1] https://www.oculus.com/quest-2/

Study design *(if relevant – e.g. randomised control trial; laboratory-based)*

The test will follow a within-subject design, where each subject will experience both custom and generic binaural filtering. Data analysis will be conducted using analysis of variance.

If the study involves participants, how many will be recruited?
Approximate 10 - 15 People due to the restriction of hardware.

If applicable, what is the statistical power of the study, i.e. what is the justification for the number of participants needed?

Based on test requirements outlined in ITU BS-1116[1], on subjective rating of audio quality, a lower number of participants is sufficient if expert listeners are employed. All listeners will come from the proposer's research group, who all have personal access to the hardware required for the experiment.

[1] ITU-R "Recommendation BS 1116-1, Methods for the Subjective Assessment of Small Impairments in Audio Systems including Multichannel Sound Systems", International Telecommunications Union Radiocommunication Assembly, 1997.

SECTION 4 RECRUITMENT OF PARTICIPANTS

How will the participants be recruited?

Researchers from the AudioLab, Department of Electronic Engineering.

What are the inclusion/exclusion criteria?

Exclusion:
People who:

1. Have hearing, vision or movement disabilities.
2. Are not comfortable with audio noise
3. Have motion sickness when wearing a VR headset.

Will participants be paid reimbursement of expenses?	YES [ ]	NO [√]
Will participants be paid?	YES [√]	NO [ ]
If yes, please obtain signed agreement		
Will any of the participants be students?	YES [√]	NO [ ]

SECTION 5 DATA STORAGE AND TRANSMISSION

If the research will involve storing personal data, including sensitive data, on any of the following please indicate so and provide further details (answers only required if *personal* data is to be stored).

Manual files	√
University computers	√
Home or other personal computers	√
Laptop computers, tablets	√
Website	√

Please explain the measures in place to ensure data confidentiality, including whether encryption or other methods of anonymisation will be used.

Computers will be password protected. Stored participant data will be anonymised. No paper documentation will exist for this study. Data will be stored in encrypted files/folders.

Please detail who will have access to the data generated by the study.

Investigators only

Please detail who will have control of and act as custodian for, data generated by the study.

Investigators only

Please explain where, and by whom, data will be analysed.

By the investigators at the University of York Audio Lab and at the investigator's homes.

Please give details of data storage arrangements, including where data will be stored, how long for, and in what form.

Data will be stored on investigators' personal computers and at the University of York

Audio lab storage in encrypted files. Data will only be stored on the investigators' laptops until the end of the PhD study.

SECTION 6 CONSENT

Is written consent to be obtained?

YES	√		NO	

If yes, please attach a copy of the information for participants

If no, please justify

Will any of the participants be from one of the following vulnerable groups?

Children under 18	YES		NO	√
People with learning difficulties	YES		NO	√
People who are unconscious or severely ill	YES		NO	√
People with mental illness	YES		NO	√
NHS patients	YES		NO	√
Other vulnerable groups (if 'yes', please give details)	YES		NO	√

If so, what special arrangements have been made for getting consent?

Consent Form will be provided to participants.

SECTION 7 DETAILS OF INTERVENTIONS

Indicate whether the study involves procedures which:

Involve taking bodily samples	YES		NO	√
Are physically invasive	YES		NO	√
Are designed to be challenging/disturbing (physically or psychologically)	YES		NO	√

If so, please list those procedures to which participants will be exposed:

List any potential hazards:

There is a small risk of noise exposure, but this will be mitigated by guiding participants through a set up procedure to set their computer to a safe volume.
There is a small risk of fatigue, which will be mitigated by reminding participants to take regular breaks and by keeping the length of the experiment short.
There is a small risk of dizziness when wearing the VR device, which will be mitigated by reminding participants to take regular breaks and by keeping the length of the experiment short.
There is a small risk when wearing the VR headset that the user may be less aware of their real world surroundings. Participants will be instructed to ensure the 'play' area is clear of any potential obstacles.

List any discomfort or distress:

None.

What steps will be taken to safeguard

(i) the confidentiality of information

Data will not be stored with participants names but using a unique ID.

(ii) the specimens themselves?

N/A

What particular ethical problems or considerations are raised by the proposed study?

None.

What do you anticipate will be the output from the study? ***Tick those that apply:***

Peer-reviewed publications	√
Non-peer-reviewed publications	√
Reports for sponsor	
Confidential reports	√
Presentation at meetings	√
Press releases	
Student project	

Is there a secrecy clause to the research? YES [ ] NO [√]
If yes, please give details below

SECTION 8 SIGNATURES

The information in this form is accurate to best of my knowledge and belief and I take full responsibility for it.

I agree to advise of any adverse or unexpected events that may occur during this project, to seek approval for any significant protocol amendments and to provide interim and final reports. I also agree to advise the Ethics Committee if the study is withdrawn or not completed.

Signature of Investigator(s): ……Hanjbo H…….

……

Date: ……10/Jul/2021……

Responsibilities of the Principal Researcher following approval

- If changes to procedures are proposed, please notify the Ethics Committee
- Report promptly any adverse events involving risk to participants

University of York

Listening test for evaluating the localisation accuracy of ITD-modification binaural filters

PROJECT INFORMATION SHEET

30th November 2022

It has been suggested that the non-personalised HRTF may confound or defeat the elevation perception in a particular frequency band (f > 1.5kHz) when subjects perform localisation tests in the median plane; however, such results need to be explored in more depth. Therefore, in this listening test, you will be asked to use a generic HRTF synthesised by an innovative optimisation technique (cross-head time delay method)to verify whether you can accurately perceive a correct virtual sound source localisation or not.

The whole listening test will last for 45-60minutes.

The listening test procedure will be as follows:

1. We will take some physical measurements of your head size(width and depth). This process will take approximately 2 minutes.

2. The data we gathered from step one will be input into the MATLAB software on the laptop, and a new HRTF set will then be generated. This process will take approximately 5 minutes.

3. You will be asked to wear the VR headset(Meta Quest 2) and using the hand controller to conduct the listening test, and following is the instruction:

    a. You will be instructed to adjust the volume to a comfortable level.
    b. Pink noise signal will be played back to you, and you are free to move your head inside the green circle(you will see it in the UI in the VR headset), the sound will go off if you move your head out of the green circle.
    c. You will need to point out the direction of the sound source using your controller by clicking the trigger(oral instruction will also be presented to you).
    d. There will be totally 150 trials present to you.

4. The listening test on VR headset will be last for 30-45 minutes.
5. You will be paid £15 in amazon vouchers following the completion of the listening test.
6. If you have any questions about this participant information sheet or concerns, please contact Hongbo Hu(hh973@york.ac.uk) in the first instance.

References

[1] M. Valente, "Strategies for selecting and verifying hearing aid fittings," *(No Title)*, 2002.

[2] H. Møller, "Fundamentals of binaural technology," *Applied acoustics*, vol. 36, no. 3-4, pp. 171–218, 1992.

[3] D. N. Zotkin, R. Duraiswami, E. Grassi, and N. A. Gumerov, "Fast head-related transfer function measurement via reciprocity," *The Journal of the Acoustical Society of America*, vol. 120, no. 4, pp. 2202–2215, 2006.

[4] H. Su, A. Marui, and T. Kamekawa, "The effect of hrtf individualization and head-tracking on localization and source width perception in vr," in *Audio Engineering Society Convention 146*. Audio Engineering Society, 2019.

[5] D. N. Zotkin, R. Duraiswami, L. S. Davis, A. Mohan, and V. Raykar, "Virtual audio system customization using visual matching of ear parameters," in *2002 International Conference on Pattern Recognition*, vol. 3. IEEE, 2002, pp. 1003–1006.

[6] 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 (Cat. No. 01TH8575)*. IEEE, 2001, pp. 99–102.

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

[8] R. S. Woodworth, *Experimental Psychology*. New York: Holt, 1938.

[9] B. Xie, *Head-related transfer function and virtual auditory display*. J. Ross Publishing, 2013.

[10] H. Wallach, "The role of head movements and vestibular and visual cues in sound localization." *Journal of Experimental Psychology*, vol. 27, no. 4, p. 339, 1940.

[11] S. Perrett and W. Noble, "Available response choices affect localization of sound," *Perception & psychophysics*, vol. 57, no. 2, pp. 150–158, 1995.

[12] E. A. Lopez-Poveda and R. Meddis, "A physical model of sound diffraction and reflections in the human concha," *The Journal of the Acoustical Society of America*, vol. 100, no. 5, pp. 3248–3259, 1996.

[13] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[14] W. G. Gardner and K. D. Martin, "Hrtf measurements of a kemar dummy-head microphone," MIT Media Lab, Tech. Rep., 1995.

[15] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *The Journal of the Acoustical Society of America*, vol. 104, no. 5, pp. 3048–3058, 1998.

[16] H. Kuttruff, *Room Acoustics*, 5th ed. London: Spon Press, 2009.

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

[18] A. T. Rosell, "Methods of measuring impulse responses in architectural acoustics," MSc dissertation, Technical University of Denmark, 2009.

[19] G.-B. Stan, J.-J. Embrechts, and D. Archambeau, "Comparison of different impulse response measurement techniques," *Journal of the Audio Engineering Society*, vol. 50, no. 4, pp. 249–262, 2002.

[20] D. Zotkin, J. Hwang, R. Duraiswaini, and L. S. Davis, "Hrtf personalization using anthropometric measurements," in *2003 IEEE workshop on applications of signal processing to audio and acoustics (IEEE Cat. No. 03TH8684)*. Ieee, 2003, pp. 157–160.

[21] H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen, "Head-related transfer functions of human subjects," *Journal of the Audio Engineering Society*, vol. 43, no. 5, pp. 300–321, 1995.

[22] D. Griesinger, "Laboratory reproduction of binaural concert hall measurements through individual headphone equalization at the eardrum," in *Audio Engineering Society Convention 142*. Audio Engineering Society, 2017.

[23] Nura Audio Pty Ltd, "Nura personalized sound technology," https://www.nura.io/how-it-works, 2019, accessed: 2026-02-22.

[24] Dolby Laboratories, "Dolby personalized audio technology overview," https://professional.dolby.com/personalized-audio/, 2018, accessed: 2026-02-22.

[25] A. Monfredini and E. Dalmolin, "Personalized spatial audio through perceptual calibration," *Journal of the Audio Engineering Society*, vol. 69, no. 5, pp. 312–325, 2021.

[26] K. Watanabe, Y. Iwaya, Y. Suzuki, S. Takane, and S. Sato, "Dataset of head-related transfer functions measured with a circular loudspeaker array," *Acoustical science and technology*, vol. 35, no. 3, pp. 159–165, 2014.

[27] 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.

[28] T. T. Sandel, D. C. Teas, W. Feddersen, and L. A. Jeffress, "Localization of sound from single and paired sources," *the Journal of the Acoustical Society of America*, vol. 27, no. 5, pp. 842–852, 1955.

[29] C. Armstrong, L. Thresh, D. Murphy, and G. Kearney, "A perceptual evaluation of individual and non-individual hrtfs: A case study of the sadie ii database," *Applied Sciences*, vol. 8, no. 11, p. 2029, 2018.

[30] A. Brammer, J. Piercy, I. Pyykkö, E. Toppila, and J. Starck, "Method for detecting small changes in vibrotactile perception threshold related to tactile acuity," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 1238–1247, 2007.

[31] F. L. Wightman and D. J. Kistler, "Headphone simulation of free-field listening. ii: Psychophysical validation," *The Journal of the Acoustical Society of America*, vol. 85, no. 2, pp. 868–878, 1989.

[32] W. György, "Hrtfs in human localization: Measurement, spectral evaluation and practical use in virtual audio environment," Ph.D. dissertation, Brandenburg University of Technology, 2002.

[33] W. M. Hartmann and B. Rakerd, "On the minimum audible angle—a decision theory approach," *The Journal of the Acoustical Society of America*, vol. 85, no. 5, pp. 2031–2041, 1989.

[34] J. C. Makous and J. C. Middlebrooks, "Two-dimensional sound localization by human listeners," *The journal of the Acoustical Society of America*, vol. 87, no. 5, pp. 2188–2200, 1990.

[35] A. W. Mills, "On the minimum audible angle," *The Journal of the Acoustical Society of America*, vol. 30, no. 4, pp. 237–246, 1958.

[36] F. Pausch, L. Aspöck, and P. Majdak, "Effect of head movements on binaural localization accuracy in virtual environments," *Frontiers in Neuroscience*, vol. 14, p. 570373, 2020.

[37] J. Roßkopf, M. Pollow, and S. Weinzierl, "Impact of head-tracked binaural auralization on sound localization accuracy," *Acta Acustica*, vol. 8, p. 10, 2024.

[38] J. Meng, Z. Wang, and X. Zhang, "Minimum audible angle in virtual sound reproduction using headphones," *Frontiers in Psychology*, vol. 12, p. 656052, 2021.

[39] M. A. Steadman, C. Kim, and S. Malhotra, "Training improves sound localization with nonindividualized hrtfs in virtual reality," *Scientific Reports*, vol. 9, p. 19039, 2019.

[40] P. Stitt, B. F. G. Katz, O. Avraham *et al.*, "Auditory accommodation to poorly matched non-individual hrtfs," *Scientific Reports*, vol. 9, p. 7682, 2019.

[41] R. P. Tame, D. Barchiese, and A. Klapuri, "Headphone virtualization: Improved localization and externalization of non-individualized hrtfs by cluster analysis," in *Audio Engineering Society Convention 133*. Audio Engineering Society, 2012.

[42] Research Institute of Electrical Communication (RIEC), Tohoku University, "Riec hrtf database," http://www.riec.tohoku.ac.jp/pub/hrtf/, 2008, accessed: 2026-02-22.

[43] K. Sunder and W.-S. Gan, "Individualization of head-related transfer functions in the median plane using frontal projection headphones," *Journal of the Audio Engineering Society*, vol. 64, no. 12, pp. 1026–1041, 2016.

[44] R. Algazi, R. O. Duda, and D. M. Thompson, "The use of head-and-torso models for improved spatial sound synthesis," 2002.

[45] "Everything to know about the hrtf directional audio update in valorant patch 2.06." [Online]. Available: https://www.sportskeeda.com/valorant/everything-know-hrtf-directional-audio-update-valorant-patch-2-06

[46] D. R. Begault, *3-D Sound for Virtual Reality and Multimedia*. Academic Press, 1994.

[47] F. Rumsey, *Spatial Audio*. Focal Press, 2012.

[48] H. Møller, "Fundamentals of binaural technology," *Applied Acoustics*, vol. 36, no. 3-4, p. 171–218, 1992.

[49] n. p. chevalier, p. majdak, e. wilk, and t. görne, "rapid hrtf measurement in a loudspeaker dome," *journal of the audio engineering society*, october 2018.

[50] j.-g. richter, g. behler, and j. fels, "evaluation of a fast hrtf measurement system," *journal of the audio engineering society*, may 2016.

[51] r. ranjan, j. he, and w.-s. gan, "fast continuous acquisition of hrtf for human subjects with unconstrained random head movements in azimuth and elevation," *journal of the audio engineering society*, august 2016.

[52] "Neumann ku 100." [Online]. Available: https://en-de.neumann.com/ku-100

[53] "Kemar." [Online]. Available: https://www.grasacoustics.com/industries/audiology/kemar

[54] A. W. Bronkhorst, "Localization of real and virtual sound sources," *The Journal of the Acoustical Society of America*, vol. 98, no. 5, p. 2542–2553, 1995.

[55] 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*, vol. 94, no. 1, p. 111–123, 1993.

[56] z. ben hur, d. alon, p. w. robinson, and r. mehra, "localization of virtual sounds in dynamic listening using sparse hrtfs," *journal of the audio engineering society*, august 2020.

[57] J. Usher and W. L. Martens, "Perceived naturalness of speech sounds presented using personalized versus non-personalized hrtfs." Georgia Institute of Technology, 2007.

[58] R. Nicol, L. Gros, C. Colomes, M. Noisternig, O. Warusfel, H. Bahu, B. F. Katz, and L. S. Simon, *A roadmap for assessing the quality of experience of 3D audio binaural rendering*, 2014.

[59] A. Andreopoulou, D. R. Begault, and B. F. Katz, "Inter-laboratory round robin hrtf measurement comparison," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 895–906, 2015.

[60] "Listen hrtf database." [Online]. Available: http://recherche.ircam.fr/equipes/salles/listen/

[61] D. Schönstein and B. F. Katz, "Variability in perceptual evaluation of hrtfs," *Journal of the Audio Engineering Society*, vol. 60, no. 10, pp. 783–793, 2012.

[62] J. Vliegen and A. J. Van Opstal, "The influence of duration and level on human sound localization," *The Journal of the Acoustical Society of America*, vol. 115, no. 4, pp. 1705–1713, 2004.

[63] R. L. Martin, K. I. McAnally, and M. A. Senova, "Free-field equivalent localization of virtual audio," *Journal of the Audio Engineering Society*, vol. 49, no. 1/2, pp. 14–22, 2001.

[64] J.-M. Pernaux, M. Emerit, J. Daniel, and R. Nicol, "Perceptual evaluation of static binaural sound synthesis," in *Audio Engineering Society Conference: 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*. Audio Engineering Society, 2002.

[65] h. su, a. marui, and t. kamekawa, "the effect of hrtf individualization and head-tracking on localization and source width perception in vr," *journal of the audio engineering society*, march 2019.

[66] J. A. Moorer, "The perceptual effects of digital signal processing in music reproduction," *Journal of the Audio Engineering Society*, vol. 25, no. 7/8, pp. 434–441, 1977.

[67] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi, "Binaural technique: Do we need individual recordings?" *Journal of the Audio Engineering Society*, vol. 44, no. 6, pp. 451–469, 1996.

[68] 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*, vol. 49, no. 10, pp. 904–916, 2001.

[69] V.-V. Mattila, "Descriptive analysis of speech quality in mobile communications: Descriptive language development and external preference mapping," in *Audio Engineering Society Convention 111*. Audio Engineering Society, 2001.

[70] M. Schoeffler, F.-R. Stöter, B. Edler, and J. Herre, "Towards the next generation of web-based experiments: A case study assessing basic audio quality following the itu-r recommendation bs. 1534 (mushra)," in *1st Web Audio Conference*, 2015, pp. 1–6.

[71] International Telecommunication Union, "Itu-r bs.1534-3: Method for the subjective assessment of intermediate quality level of audio systems (mushra)," 2015, recommendation ITU-R BS.1534-3.

[72] J. M. Loomis, C. Hebert, and J. G. Cicinelli, "Active localization of virtual sounds," *The Journal of the Acoustical Society of America*, vol. 88, no. 4, pp. 1757–1764, 1990.

[73] W. O. Brimijoin, A. W. Boyd, and M. A. Akeroyd, "The contribution of head movement to the externalization and internalization of sounds," *PloS one*, vol. 8, no. 12, p. e83068, 2013.

[74] T. Leclère, M. Lavandier, and F. Perrin, "On the externalization of sound sources with headphones without reference to a real source," *The Journal of the Acoustical Society of America*, vol. 146, no. 4, pp. 2309–2320, 2019.

[75] A. Kulkarni and H. S. Colburn, "Role of spectral detail in sound-source localization," *Nature*, vol. 396, pp. 747–749, 1998.

[76] R. S. Woodworth and H. Schlosberg, *Experimental psychology*. Oxford and IBH Publishing, 1954.

[77] Y. Kahana and P. A. Nelson, "Boundary element simulations of the transfer function of human heads and baffled pinnae using accurate geometric models," *Journal of sound and vibration*, vol. 300, no. 3-5, pp. 552–579, 2007.

[78] B. F. Katz, "Boundary element method for sound field calculations," *Acta Acustica united with Acustica*, vol. 87, pp. 600–609, 2001.

[79] D. Botteldooren, "Finite-difference time-domain simulation of low-frequency room acoustic problems," *The Journal of the Acoustical Society of America*, vol. 98, no. 6, pp. 3302–3308, 1995.

[80] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen, "Creating interactive virtual acoustic environments," *Journal of the Audio Engineering Society*, vol. 47, no. 9, pp. 675–705, 1999.

[81] C. Armstrong, "Improvements in the measurement and optimisation of head related transfer functions for binaural ambisonics," Ph.D. dissertation, University of York, 2019.

[82] 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, 2019.

[83] T. Carpentier and M. Noisternig, "Spat5 spatial audio toolkit for max," 2015, iRCAM.

[84] D. Johnston, B. Tsui, and G. Kearney, "SALTE Pt. 1: A Virtual Reality Tool for Streamlined and Standardized Spatial Audio Listening Tests," *Journal of the Audio Engineering Society*, no. 536, oct 2019.